A Visual Attention Model for High Dynamic Range (HDR) Video Content

by

Yuanyuan Dong

B.Eng. (Communication Engineering), Tianjin University, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2014

Abstract

High dynamic range (HDR) imaging is gaining widespread acceptance in computer graphics, photography and multimedia industry. Representing scenes with values corresponding to real-world light levels, HDR images and videos provide superior picture quality and more life-like visual experience than traditional 8-bit Low Dynamic Range (LDR) content. In this thesis, we present a few attempts to assess and improve the quality of HDR using subjective and objective approaches.

We first conducted in-depth studies regarding HDR compression and HDR quality metrics. We show that High Efficiency Video Coding (HEVC) outperforms the previous version of compression standard on HDR content and could be used as a platform for HDR compression if provided with some necessary extensions. We also find that, compared to other quality metrics, the Visual Information Fidelity (VIF) quality metric has the highest correlation with subjective opinions on HDR videos. These findings contributed to the development of methods that optimize existing video compression standards for HDR applications.

Next, the viewing experience of HDR content is evaluated both subjectively and objectively. The study shows a clear subjective preference for HDR content when individuals are given a choice between HDR and LDR displays. Eye tracking data were collected from individuals viewing HDR content in a free-viewing task. These eye tracking data collected are utilized in the development of a visual attention model for HDR content.

Last but not least, we propose a computational approach to predict visual attention for HDR video content, the only one of its kind as all existing visual attention models are designed for HDR images. This proposed approach simulates the characteristics of the Human Visual System (HVS) and makes predictions by combining the spatial and temporal visual features.

The analysis using eye tracking data affirms the effectiveness of the proposed model. Comparisons employing three well known quantitative metrics show that the proposed model substantially improves predictions of visual attention of HDR.

Preface

All of the work presented in this thesis was conducted in the Digital Multimedia Laboratory at the University of British Columbia, Vancouver campus. All non-original figures and tables have been used with permission from applicable sources mentioned in their descriptions.

A version of Chapter 3 has been published as Y. Dong, P. Nasiopoulos, and M. T. Pourazad, "HDR video compression using high efficiency video coding (HEVC)," UBICOMM 2012, The Sixth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, September 2012. I was the lead investigator responsible for all areas of research, data collection, as well as the majority of manuscript composition. M. T. Pourazad was involved in the early stages of research concept formation and aided with manuscript edits. P. Nasiopoulos was the supervisor on this project and was involved with research concept formation, and manuscript edits.

Part of the results of Chapter 3 are taken from the work published as M. Azimi, A. Banitalebi, Y. Dong., M.T. Pourazad, and P. Nasiopoulos, “A survey on the performance of the existing full reference HDR video quality metrics: A new HDR video dataset for quality evaluation purposes,” International Conference on Multimedia Signal Processing, Venice, Italy, November 2014. I was involved in research concept formation and test data collection. M. Azimi was responsible for data collection and manuscript composition. A. Banitalebi was involved in data collection and test implementation. M.T. Pourazad was involved in the early stages of research concept formation and aided with manuscript edits. P. Nasiopoulos was the supervisor on this project and was involved with research concept formation, and manuscript edits.

A version of Chapter 4 has been published as Y. Dong, E. Nasiopoulos, M.T. Pourazad, and P. Nasiopoulos, "High Dynamic Range Video Eye Tracking Dataset," 2nd International Conference on Electronics, Signal processing and Communications, ESPCO, Greece, November 2014. I was the lead investigator responsible for all areas of research, study implementation, data collection, as well as manuscript composition. E. Nasiopoulos was involved in study implementation and data collection. M.T. Pourazad was involved in the early stages of research concept formation and aided with manuscript edits. P. Nasiopoulos was the supervisor on this project and was involved with research concept formation, and manuscript edits.

Part of the results of Chapter 4 are taken from the work published as E. Nasiopoulos, Y. Dong, and A. Kingstone, "Evaluation of High Dynamic Range Content Viewing Experience Using Eye-Tracking Data," 10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness, QSHINE, Greece, August 2014. I was involved in study implementation and manuscript composition. E. Nasiopoulos was involved in research concept formation, study implementation and manuscript composition. The work was conducted with the guidance and editorial input of A. Kingstone.

A version of Chapter 5 is to be submitted to an IEEE Journal. I was the lead investigator responsible for designing and implementing the proposed algorithms, performing all experiments, analyzing the results, and writing the manuscripts. M.T. Pourazad and P. Nasiopoulos provided guidance and editorial input for this project.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2D	Two Dimensional
3D	Three Dimensional
AUC	Area Under Curve
AVC	Advanced Video Coding
AWGN	Additive White Gaussian Noise
CAM	Color Appearance Model
CC	Correlation Coefficient
CfP	Call for Proposal
CSF	Contrast Sensitivity Function
CVI	Contrast Versus Intensity
DS	Double Stimulus
FDM	Fixation Density Map
fps	frames per second
GOP	Group of Picture
HD	High Definition
HDR	High Dynamic Range
HDR-VDP-2	High Dynamic Range Visible Difference Predictor 2
HEVC	High Efficiency Video Coding
HVS	Human Visual System
IR	Infrared
ITU-T	International Telegraph Union- Telecommunication Standardization Sector
JCT-VC	Joint Collaborative Team on Video Coding
KL	Kullback-Leibler Divergence
LCD	Liquid Crystal Display
LDR	Low Dynamic Range
LED	Light Emitting Diode
MOS	Mean Opinion Score
MPEG	Moving Pictures Experts Group
PCC	Pearson Correlation Coefficient

PSNR	Peak Signal-to-Noise Ratio
PU	Perceptually Uniform
QP	Quantization Parameters
RMSE	Root Mean Square Error
RD	Rate-Distortion
ROC	Receiver Operating Characteristic Analysis
RoI	Region of Interest
SCC	Spearman Rank-Order Correlation Coefficient
SDR	Standard Dynamic Range
SSIM	Structural Similarity
TV	Television
TVI	Threshold Versus Intensity
UHD	Ultra High Definition
VCEG	Video Coding Experts Group
VDP	Visible Difference Predictor
VIF	Visual Information Fidelity

Acknowledgements

I give sincere thanks to my supervisor, Dr. Panos Nasiopoulos, for his guidance and support throughout my M.A.Sc program.

I would also like to thank Jan Lüdert, without whom this thesis would not have been possible.

Finally, I thank my parents for their love and support.

Chapter 1: Introduction

High dynamic range (HDR) technologies are rapidly growing and gaining widespread acceptance in computer graphics, photography and multimedia industry. With the goal to provide life-like visual experience, the visual quality of HDR is vastly higher than conventional low dynamic range (LDR) content. The difference is as big as, if not more, the difference between black-and-white and color television [1].

For LDR, most of the color images are represented with one byte (8 bits) per pixel for each of the red, green, and blue channels, i.e., three bytes of data per pixel. Using such a representation scheme, there are only 256 different shades for each of the red, green and blue components. 256 different shades in each color family are much less than what human eyes can perceive and inadequate to represent many scenes in real life. Different from LDR, HDR captures and storages the luminance and color information of the real world, so both very dark and very bright objects can be truthfully represented in the same image.

With the truthful representation of the real world, more details and information about the scenes, and a life-like visual experience, HDR imaging is affecting not only specialized fields like film and photography, but literally every stage of the digital media pipeline and many other areas. High-end cinematographic cameras are already utilizing HDR techniques to provide content with better quality. Video game developers and graphics card vendors are incorporating HDR into video game engines to deliver more believable virtual worlds even if the final video is displayed using the existing LDR displays. In medical imaging, HDR displays could show better contrast than existing medical displays, thus better facilitate diagnosing. Besides, HDR techniques can find applications in computer vision, surveillance and scientific visualization.

In the last few years, HDR imaging has been gaining widespread acceptance and the related technologies are maturing rapidly. Camera sensors have been developed with the capability to capture HDR images and videos. It is also possible to generate HDR content by fusing multiple exposures using LDR sensors through software solutions. The display industry has also started to take note of the potential of HDR technology. Prototypes of HDR display are built with dynamic ranges of well beyond 50,000:1 according to [2]. Moreover to ensure smooth transition from LDR to HDR service, the backward compatibility with current LDR display systems is under investigation. At the introductory phase of HDR systems, HDR displays (that accept 10-bit or 12-bit signals) and LDR systems (that accept only 8-bit data) will coexist. Thus, the broadcasters should provide both LDR and HDR signals for consumers. To efficiently allow for this overlap, a number of tone-mapping operators have been developed which convert 10-16 bit HDR content to the 8-bit LDR signal.

As all HDR related technologies are still at the infancy level, there are many uninvestigated topics and areas worth research attention in order to fully leverage the potential of HDR. In this thesis, we present our work that focuses on assessing and improving the quality of HDR using subjective and objective approaches. This includes evaluations of existing video compression standards and quality metrics for HDR videos, eye tracking studies utilizing HDR content and display and the design of a visual attention model for HDR content.

Chapter 2 provides the background information on HDR, human visual attention model and eye tracking. Generating and displaying of HDR content are discussed in Section 2.1. Two of the commonly used ways to compress HDR are presented in Section 2.2. An overview of visual attention model is provided in Section 2.3. Finally, the principle and applications of eye tracking are summarized in Section 2.4.

Chapter 3 presents two studies leveraging HDR videos and a prototype HDR display. In the first study, we compare the performance of the HEVC standard with that of H.264/AVC for compressing HDR content. The study confirms that HEVC does not only offer superior compression performance for LDR content as shown from previous studies, but also for HDR videos. The compression improvement in the HDR case is in line with that of LDR. In the second study, the performance of existing full-reference video quality metrics on HDR content is evaluated. A new HDR video dataset with different representative types of distortions is created for the study. By comparing the opinion scores from subjective tests and the quality scores from objective tests, the study shows the ability of each quality metric to predict the quality of HDR videos.

Chapter 4 presents an HDR video eye tracking dataset collected from naïve participants viewing HDR videos in a few viewing tasks. The study also shows a clear subjective reference for HDR displays when individuals are given a choice between HDR and LDR displays.

Chapter 5 presents a new computational approach to predict visual attention for HDR content, and the experiments conducted to evaluate the proposed approach. To begin with, we highlight limitations stemming from available saliency detection models when applied to HDR images and videos. Then we propose a new saliency detection method for HDR images and videos by incorporating the color and intensity perception of the Human Visual System (HVS). The general philosophy behind this method is to utilize the bottom-up structure while bearing the properties of the HVS in mind. Both spatial and temporal cues are taken into account, leading to two saliency maps: the spatial saliency map and the temporal saliency map. To obtain the spatial saliency map, we use the HVS model to decompose feature channels from an HDR input and then follow the procedure of the classical bottom-up method. To compute the temporal saliency

map, an optical flow based method is used to estimate motion. Finally, a dynamic fusion method is proposed to combine both the spatial and temporal saliency maps.

Chapter 2: Background

2.1 High Dynamic Range Imaging

The human visual system (HVS) is able to adapt to a huge range of light conditions, from about 10^{-6} cd/m^2 to 10^8 cd/m^2 (10^{14}:1). In a single view, the human eyes can perceive a dynamic range at the order of 10^5:1 [3]. Contrary to the wide range of light intensity perceived by HVS, the vast majority of existing consumer cameras and display devices are only able to support LDR content with contrast ratio of approximately 100:1 to 1000:1. A new-generation of imaging system promises to overcome this restriction by capturing and displaying high dynamic range (HDR) images and videos which contain information that covers the full visible luminance range and the entire color gamut.

2.1.1 Generation of HDR Content

In order to fully capture and represent the color space and the dynamic range visible to human eyes, quite a few solutions have been proposed in recent years. One solution is to combine multiple LDR images captured at different exposure levels [4]. Figure 2.1 shows a few different exposures of a scene. Each image in the sequence has some pixels properly exposed and other pixels underexposed or overexposed. Since each of the exposure contains certain details, combining all exposures provides an HDR image (see Figure 2.2) covering a wider dynamic range than any single one of the exposures. This technique has been widely used in many applications such as the camera of Apple iPhone.

In real life, objects in the scene don't remain still in all exposures and the camera is not perfectly still either. So it is necessary to have techniques such as image-alignment and lens flare removal to overcome the subtle relative motion between exposures.

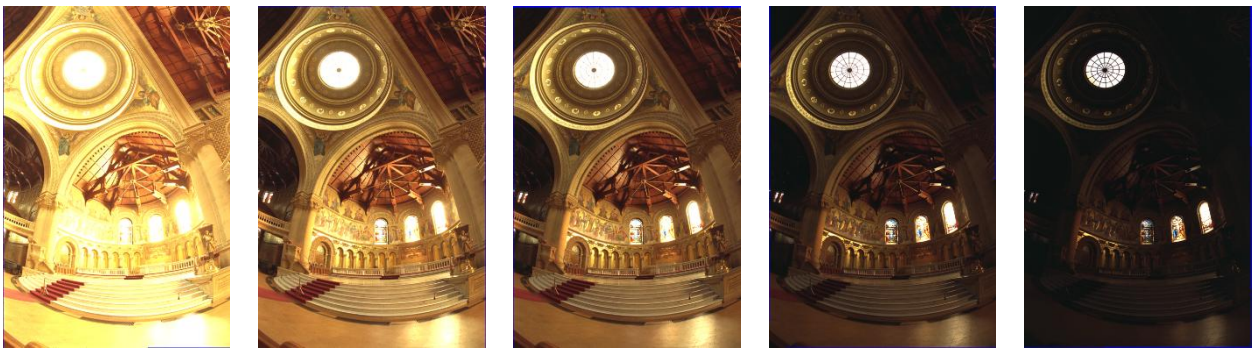


Figure 2.1 A scene captured at different exposures.

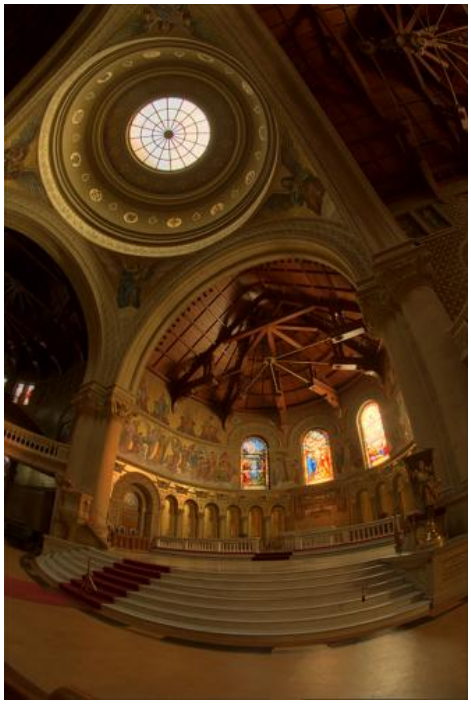


Figure 2.2 An HDR image generated by combining multiple exposures.

Using this idea, camera manufacturer RED invented its HDR solution for video cameras. The RED camera records two exposures within the interval that a standard video camera would record only one, one normal exposure and one high light protection exposure [5]. The primary exposure (the "A frame") uses the standard aperture and shutter settings. The secondary exposure (the "X frame"), highlight protection exposure, uses an adjustable shutter speed that is 2-6 stops faster than the primary exposure. The two steams are combined to create an HDR video with controlling of motion blur. Due to the different shutter speed of the two exposures, the blur of motion in A and X frame is proportionally different. When the two videos are combined,

movement is rendered with a sharp component and a more blurred component on the leading edge. Thus, the motion appears smoother [5]. We use RED SCARLET cameras to capture some of the test sequences in this thesis.

In another solution of HDR video capturing [6], a mirror-rig is used to capture different exposures at the same time. A common glass pane with antireflective coating is used as a beam splitter. The beam splitter has a ratio of around 1: 16 between reflection and transmittance, which means a shift of 4 stops of camera exposure. Two Alexa cameras, a CMOS sensor based motion picture camera made by Arri, are employed in the system as a highlight preserving camera and a lowlight-preserving camera respectively. In post-production, the highlight preserving image is aligned to the lowlight-preserving image to increase spatial fit [6]. The videos captured using this system are posted on the project website [7]. Some of the video sequences are used in this thesis.

Similarly, Technicolor has captured some HDR sequences with a rig of two Sony F3 or F65 cameras [8]. Some of the sequences produced with this system have been submitted to MPEG and JCT-VC. Two of the five sequences submitted are used in this thesis.

2.1.2 HDR Display

Besides capturing of HDR content, displaying of HDR has received a lot of attention from both industry and research community in the last decade. Tone mapping operators have been proposed to reduce the dynamic range of HDR content, so that HDR content can be displayed on typical existing LDR screens. Even though tone mapping operators have made it possible to show HDR content on regular displays in a pleasing way and often better than LDR content, it is a matter of fact that conventional displays are not capable of providing a real-world visual experience. To improve the dynamic range and visual experience of current displays, Seetzen et al. proposed “dual-modulation” displays in [2]. The fundamental idea is to use two

optical modulators (displays) in such a way that the combined contrast is the product of two displays (this is the theoretical value. In real life, the combined contrast could be much lower than the theoretical value).

Based on this principle, the first design used a high intensity video projector as a "backlight" and an LCD panel in front of the projector. In a more advanced implementation (which has its limitations too), the projector is replaced with a low-resolution array of ultra-bright light-emitting diodes (LEDs). The LEDs can be controlled individually or in clusters, so they are dimmed or off behind dark regions in the image. Comparing to the first design, the latter design has the advantage of allowing for thin displays and reduced energy requirements.

There is only one HDR display available in the market at a prototype level and very expensive at the time of this writing. This is the HDR47E S 4K, built by SIM2 Multimedia, which licensed the technology from Dolby, offering a peak brightness of 4000 cd/m^2 and contrast ratio of 20, 000:1 [9]. This is based on the original technology invented by Brightside Technologies, a UBC spin-off that was bought by Dolby [10].

2.2 High Dynamic Range Video Coding

As with all HDR technologies for capturing and displaying, HDR compression is a topic worth research attention, since it is going to enable efficient transmission of HDR content. The transmission of HDR content requires provisions beyond those used in transmission of conventional LDR content as HDR videos involve much more information than their LDR counterparts. Depending on whether or not the LDR data formats are supported, HDR compression schemes can be classified into two categories: backward-compatible HDR video compression and direct HDR video compression using the latest video coding standard, the High Efficiency Video Coding (HEVC).

2.2.1 Backward-Compatible HDR Video Compression

Figure 2.3 summarizes the coding structure of backward-compatible HDR video compression. The input HDR frame is first tone-mapped to an 8-bit LDR frame. This LDR frame is then encoded using video coding standards, e.g., H.264/AVC, to generate a compressed stream. This stream can be decoded and displayed on any LDR players and devices. Besides the LDR stream, a residual stream is generated, which contains the information loss introduced by tone-mapping and video compression. For HDR devices at the receiver end, both streams are combined to recover the HDR video [11].

This approach ensures backward compatibility with the exiting LDR displays, and allows reconstruction of the HDR content for HDR displays, but it comes at a cost of low compression efficiency and suffers from information loss introduced by tone mapping and inverse tone mapping.

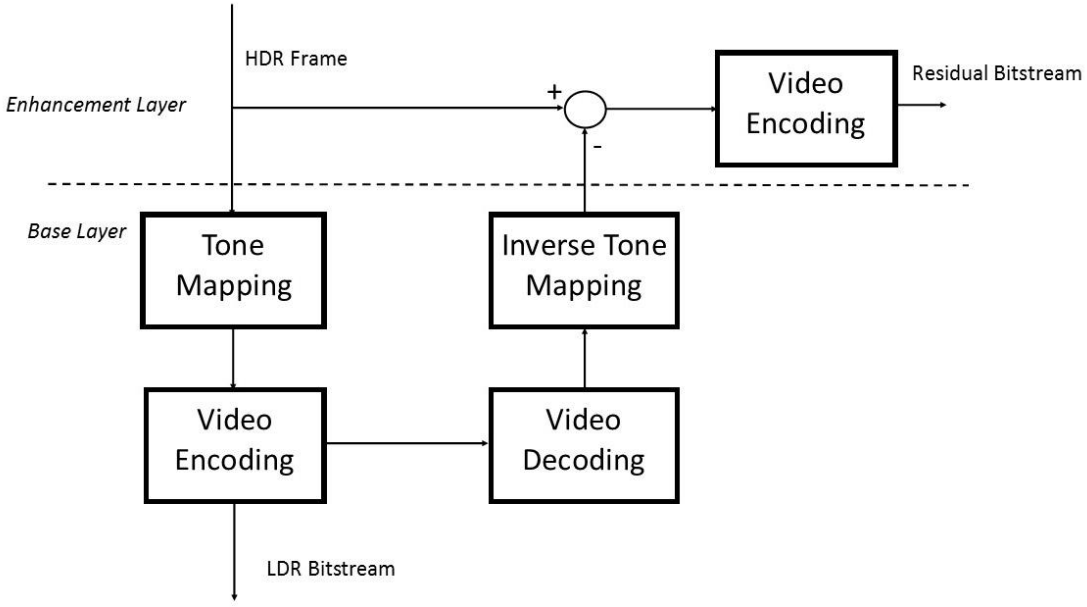


Figure 2.3 Structure of backward-compatible HDR video compression.

2.2.2 HDR Video Compression Using the HEVC Standard

The previous video compression codec H.264/AVC was able to compress videos so that they can be streamed over the network. However, the new types of videos such as Ultra HD (UHD), 4K and HDR content have a lot more details and much larger raw data size. A more efficient video coding standard is critical to transmit and store these digital videos.

The limitations of current video coding standards prompted the ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG) to establish a Joint Collaborative Team on Video Coding (JCT-VC) with the objective to develop a new video coding standard, HEVC. A formal Call for Proposals (CfP) on video compression technology was issued in January 2010. Since then, JCT-VC has put a considerable effort towards the development HEVC standard, with the aim to double the compression efficiency compared to the existing H.264/AVC standard. Eventually, in April 2013 HEVC was ratified as the new video compression standard [12]. At the time of this writing, JCT-VC is working on a number of extensions to support more video services including scalable coding extensions, multi-view extensions and HDR extensions.

Compared to the previous standard, H.264/AVC, HEVC has four key advantages: 1) reduced bitrate requirements by half with comparable image quality; 2) ability to support higher resolution such as 4K/UHD video; 3) improved parallel processing methods to speed up the computation; 4) improved picture quality in terms of noise level, color spaces, and dynamic range. Meanwhile, HEVC has made changes to accommodate the requirement of HDR content. In order to support the increased dynamic range of image brightness in HDR content, there is the need to increase the bit-depth of encoding video, so that more shades of grey, as well as more shades of colors of each color family can be represented. To fulfill this need, HEVC extended the

bit-depth of previous standard to support bit-depth from 8 bits to 16 bits in the July 2014 draft [13].

2.3 Human Visual Attention Model

As one of the most important features of the HVS, visual attention has been studied for over a hundred years. The first trial to describe visual attention dates back to 1891 given by psychologist William James [14]:

Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others…

When the things are apprehended by the senses, the number of them that can be attended to at once is small.

Real life visual environments contain far more information than the HVS is able to process. Human cognition reacts to such over stimuli by selectively attending to some objects while ignoring other regions of the visual environment. This cognitive selection enables the effective allocation of limited visual processing resources; reducing mental efforts in object detection and recognition.

Visual attention research investigates the properties of the HVS and simulates the cognitive selection, geared towards predicting salient regions or objects. So far, research has focused on two main mechanisms directing visual attention [15]. First, the top-down attention, also called overt attention, which is voluntary and task-driven [16]. It is influenced by cognitive factors such as task, experience, emotions, expectations and knowledge of the observer [17]. It has been studied in various natural environments such as web search [18] and multimedia

learning [19]. Since the top-down attention is highly dependent on the task and the observer, most of the existing computational models focus on the bottom-up process, which is involuntary, fast, stimulus-driven, and mainly dependent on the intrinsic features of the visual stimuli itself.

Visual saliency mostly refers to the bottom-up process. Visual saliency detection is defined as the process predicting salient areas through computational models [17]. Since visual saliency detection can predict what an observer focuses on given a visual input, it has a very wide range of engineering applications, including object detection and recognition [20], object tracking [21], robotics [22], image and video compression [23] [24], and dynamic content resizing [25].

In order to develop, evaluate and optimize the computational visual attention models or saliency detection methods, the ground truth of human visual attention is required. There are mainly two ways to record human visual attention in subjective experiments. In some cases, experiments require participants to manually label the relevant or important areas in the presented content [26]. The other way is to obtain the ground truth through eye tracking experiments given the strong link between visual attention and eye movements [27]. Humans typically fixate their gaze on the location that they are currently attending. Thus, gaze pattern serves as a surrogate to what humans are actually attending to in an image [28]. Eye tracking experiments record observers' gaze over time and the resulting gaze patterns can be post-processed into fixation density maps (FDM). The average FDM over all observers is considered as the ground truth of human visual attention [29].

2.4 Eye Tracking

Eye movement research and eye tracking started to flourish in the last decade, with great advances in both eye tracking technology and psychological theory to link eye tracking data to

cognitive processes [30]. With the ability to record gaze movement and analyze attention shift, eye tracking has become a useful tool in many research disciplines such as psychology, cognitive neuroscience, neurology, and marketing. In more recent times, eye tracking has also shown growth as a means of interacting with the computer. In those applications, eye tracker serves as a powerful input device and the system interact with the user on the basis of eye movements.

Currently, the most widely used designs of eye tracking systems are video-based eye trackers. A typical eye tracking system setup includes a video camera to record the movements of the eye(s) and a computer to analyze the gaze data. In addition, infrared (IR) light sources are often used to help improve the accuracy of gaze position tracking [31]. The user sits in front of the monitor on which the content is presented and the eye tracking system records the user's gaze vector as screen co-ordinates. Eye positions and other information are then extracted from the recoded video.

Typically, eye tracker records three kinds of eye movements: saccades, fixations and eye blinks. A saccade is a jump-like, rapid eye movement which allows the eye to move from one location to another rapidly. A fixation happens when the eye gaze pauses at a certain position. Fixations are characterized by consecutive eye data having a velocity below a given threshold [32]. On average, there are about 2-4 eye fixations per second per observer and fixations last around 200ms when reading text and 300ms when viewing of a scene. Human eyes alternate between fixations and saccades; the resulting series of fixations and saccades are called scanpaths.

Chapter 3: HDR Compression and Quality Metrics

3.1 HDR Video Compression Using HEVC

The performance of the HEVC standard on HDR content has not been taken into account at the time of developing this standard and all the tests were conducted using LDR content[1]. Given the difference in properties and characteristics between LDR and HDR content, it is important to consider how HEVC will perform on HDR videos and from these tests try to identify challenges and additions or changes to the new standard. In this work we investigate the compression performance of HEVC on HDR video content, to examine if HEVC has the potential to be used as a platform for a devoted HDR compression scheme. We conduct experimental tests on HDR content and compare the performance of HEVC with that of H.264/AVC standard. Comparable experiment settings of two codecs are introduced which could also be used in other similar tests.

3.1.1 Test Sequences

For our experiment, four test sequences were selected from the database provided by JVT of ISO/IEC MPEG & ITU-T VCEG [33] [34]. These test videos are in YUV 4:2:0 format, with a resolution of 1080p and a frame rate of 50 fps. The bit depth of two of the videos is 10 bits and that of the other two is 12 bits. Figure 3.1 shows snapshots of the four test sequences. The specifications of the test sequences are summarized in Table 3.1.

[1] When this study regarding HDR video compression using HEVC was conducted, HEVC was not finalized yet and no test or study of HEVC had involved HDR content. After HEVC was finalized in 2013, JCT-VC kept working on various extensions of HEVC to support more requirements and services. In the 107th meeting at San Jose, a draft requirement of HDR video coding was approved for the first time [77]. This document was submitted as a request to MPEG to investigate if existing standards adequately satisfy the needs of HDR. This document provided various discussion points to help in reaching the goal to support HDR videos in the framework of HEVC.

Table 3.1 HDR test sequences.

Name	Bit Depth	Resolution	Frame Rate
Capital	10	1920x1080	50 fps
Freeway	10	1920x1080	50 fps
Library	12	1920x1080	50 fps
Sunrise	12	1920x1080	50 fps

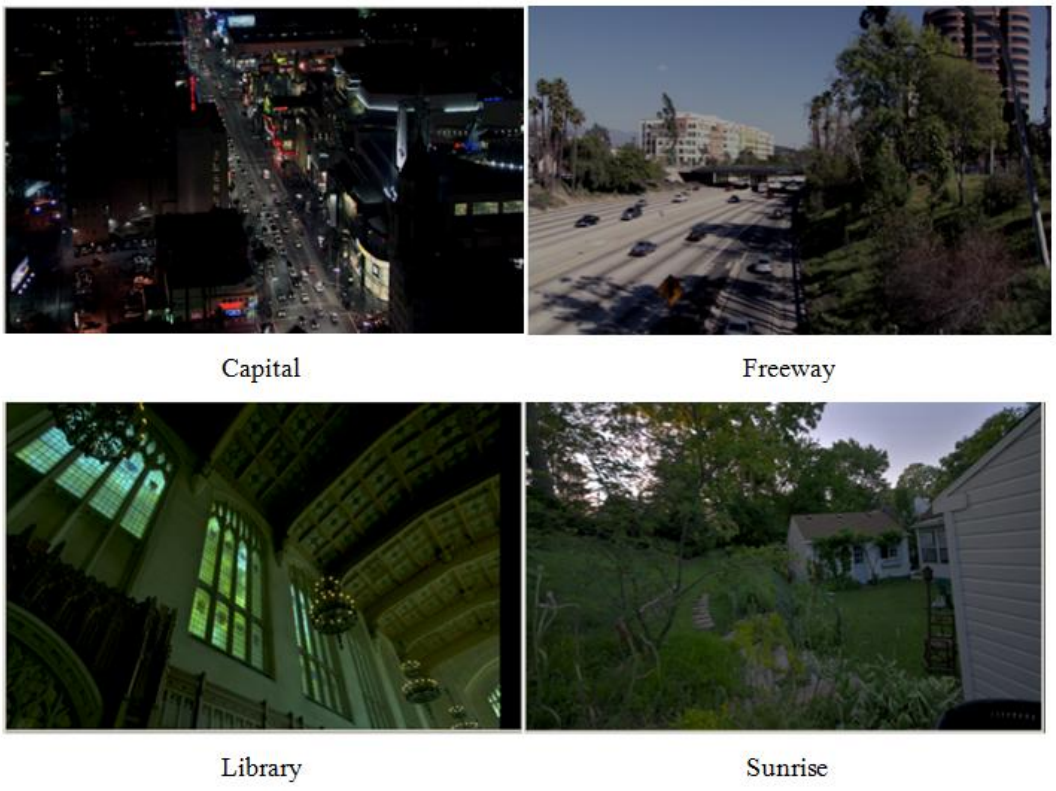


Figure 3.1 Snapshots of the test sequences.

These test sequences have been generated from HDR video content that was originally stored in floating point format and in a linear RGB space. The representation of the sequence was created by first normalizing the RGB values to the set [0, 1]. Then these normalized values were converted to the YCbCr format using the ITU-R BT.709 reference primaries. Chroma planes were subsampled by a factor of two in each dimension using the given separable filter

(refer to [33] for more details). Finally, the resulting 4:2:0 YUV file was quantized linearly with a rounding operation to create the test sequences [34].

3.1.2 HEVC Configuration

To evaluate the performance of HEVC on HDR content, we used the HEVC Test Model 5 (HM 5.0) [35]. Note that HM 5.0 was the latest available HEVC Test Model at the time of conducting this experiment. To enable the highest possible compression performance, the Random Access High Efficiency (RA-HE) configuration was used in our experiment: Hierarchical B pictures, Group of Picture (GOP) length of 8, ALF (Adaptive Loop Filter), SAO (Sample Adaptive Offset) and Rate Distortion Optimized Quantization (RDOQ) were enabled. In order to obtain a reasonable span of Rate-Distortion (RD) curves, the following Quantization Parameters (QPs) were used: 28, 32, 36, and 44. QP is the parameter, which controls the quantization step size, and in turn decides the level of quantization error involved during compression. A higher QP value leads to a larger quantization step size and worse video quality.

3.1.3 H.264/AVC Configuration

In our experiment, the performance of HEVC is compared with the state-of-the-art video compression standard H.264/AVC (JM 16.2). To accommodate HDR content, the configuration of H.264/AVC was set to High 4:4:4 Profile, which accepts up to 14 bits. In our experiment we used hierarchical B pictures, GOP length of 8, CABAC entropy coding and RDOQ enabled. These settings were recommended for comparing H.264/AVC to HEVC by MPEG/VCEG in the joint CfP (for more details check the Alpha anchor in [36]). The same QP settings as those in the HEVC case are used for H.264/AVC.

All the above-mentioned configuration settings were chosen to ensure a fair comparison between HEVC and H.264/AVC. However, these codecs are so different and have different tools

and configuration options. As a result, aside from the necessary changes and above-mentioned settings, the default settings are used for the rest of available options.

3.1.4 Results and Discussions

To evaluate the performance of HEVC versus H.264/AVC for coding HDR content, we conducted our experiment using the infrastructure provided in the previous sections. Figure 3.2 shows the RD curves for all the test sequences and Table 3.2 lists the average PSNR improvement and average PSNR savings achieved by HEVC over the H.264/AVC standard.

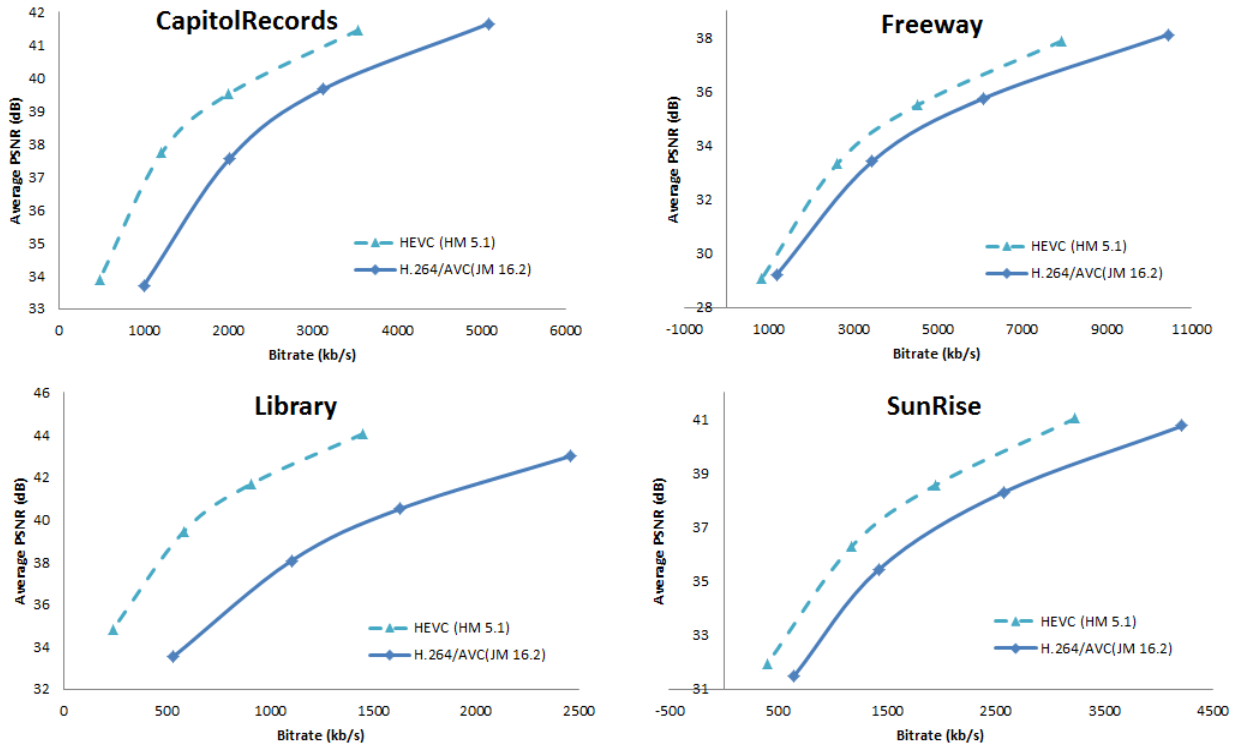


Figure 3.2 Rate-Distortion curves for HDR videos.

Table 3.2 Average compression improvement.

Name	Average PSNR Improvement	Average Bitrate Saving
Capital	1.26 dB	42.12 %
Freeway	1.02 dB	22.74 %
Library	4.88 dB	58.61 %
Sunrise	1.79 dB	32.60 %

As it can be observed, HEVC outperforms H.264/AVC by 22.47% to 58.61% in terms of bitrate (with same PSNR) or 1.02 dB to 4.88 dB in terms of PSNR (with same bitrate). Our results show that the compression efficiency of HEVC when applied to HDR content is dramatically higher than H.264/AVC and seems to follow the performance already witnessed for LDR content.

This study confirms that HEVC does not only offer superior compression performance for LDR content but also HDR videos. The compression improvement in the HDR case is in line with that of LDR. However the overall saving differs among different sequences (content dependent). Service providers could greatly benefit from HEVC due to more efficient use of bandwidth. It is worth mentioning that HEVC's high compression performance comes at the price of increased coding complexity compared to H.264/AVC.

3.2 Evaluation of Existing Full-Reference Quality Metrics on HDR Videos

The main focus of this work is to evaluate the performance of the existing LDR and HDR metrics on HDR video content which in turn will allow for a better understanding of how well each of these metrics work and whether they can be applied in capturing, compressing, and transmitting process of HDR data. To this end a series of subjective tests are performed using DML-HDR video database [37]. This HDR video dataset contains original HDR videos and also distorted HDR videos with several different types of artifacts. Then, the correlation between the results from the existing quality metrics and those from subjective tests is measured to determine the effectiveness of exiting quality metrics for HDR.

3.2.1 DML-HDR Dataset

Since there was not yet an HDR video dataset when this work was conducted, we started the work with creating a comprehensive HDR video database called "DML-HDR" [37]. The

video dataset started with five videos and at the time of this writing, there are ten videos in the dataset. They are available on the webpage of DML-HDR [37].

All videos in the dataset were captured using RED Scarlet-X cameras, which can capture dynamic range up to 18 stops [5]. Video sequences are about 10 seconds long with frame rate of 30 frames per second (fps) and resolution of 2048x1080. Videos in this dataset are available in two formats: RGBE and YUV 12-bit. RGBE is a lossless HDR format and it assigns four bytes to represent each pixel: one byte for the mantissa of each of the R, G, B channels and the remaining one byte is used as a shared exponent [38]. The YUV 12-bit format consists of three channels, Y (Luma), U and V (Chroma). Each channel is represented by integer values between 0 and 4095 (12 bits).

To evaluate the performance of quality metrics, five types of distortions were applied to HDR videos in DML-HDR dataset. These five types of distortions are:

- Additive White Gaussian Noise (AWGN): white Gaussian noise with mean of zero and standard deviation of 0.002 was added to all frames of each video. Based on our knowledge from LDR videos, this value of standard deviation may seem to be too small. However, observations from watching distorted HDR videos on the HDR display showed that AWGN with the standard deviation value of 0.002 is visible. This may be due to their larger dynamic range compared to LDR videos. Note that, before adding the AWGN noise to the HDR videos, all pixel values were normalized between 0 and 1. After adding the AWGN noise, pixel values were converted back to the original scale.
- Intensity shift: the luminance of each HDR video was globally increased over time by 10% of the maximum scene luminance.


Chritmas Tree


Hallway


ICICS


Mainmall


Playground


Stranger


Table


UBC


Waking Girl


Walking on Snow

Figure 3.3 Snapshots of video sequences in DML-HDR dataset. The number of videos in the dataset is still growing to meet the needs of different projects. These are the ten videos in the dataset at the time of this writing.

- Salt and pepper noise: Salt and pepper noise was added to 2% of the pixels in each frame. The distribution of the affected pixels by salt and pepper noise was randomized from frame to frame.
- Low Pass Filter: An 8×8 Gaussian low pass filter with standard deviation of 8 was applied to each frame of all the sequences. Subsequently, rapid changes in intensity in each frame were averaged out.
- Compression artifacts: All the videos were encoded using the HEVC encoder (HM software version 12.1[2]) with random access main10 profile configuration. The HEVC

[2] This was the lasted version of HEVC when this study was conducted.

encoder settings were as follows: hierarchical B pictures, group of pictures (GOP) size of 8, internal bit-depth of 12, input video format of YUV 4:2:0 progressive, and enabled rate-distortion optimized quantization (RDOQ). The quantization parameter (QP) was set to 22, 27, 32, and 37 in order to simulate impaired videos with a wide range of compression distortions.

The compressed videos are available in 12-bit YUV format in the "DML-HDR" video dataset, whereas all other distortion types are available in HDR format (.hdr). This is because the YUV format is the default format used by the HEVC reference software.

3.2.2 Experiment Procedure

In order to evaluate the performance of quality metrics, we conducted both subjective tests and objective tests. In subjective tests, the Mean Opinion Scores (MOS) on distorted videos are collected; in objective tests, quality metrics provide quality scores for each distorted video; finally, results from subjective and objective tests are compared.

3.2.2.1 Subjective Tests

The videos were displayed on a Dolby HDR TV prototype built based on the concept explained in [2]. This system consists of two main parts: 1) a 40 inch full HD LCD panel in the front, and 2) a projector at the back to provide the backside luminance. This HDR display is capable of emitting light at a maximum luminance level of 2700 cd/m^2.

To present HDR content on this display, the original HDR video signal is split into two streams, for the projector and the LCD respectively [2]. The input signal to the projector contains only luminance information of the HDR content and the input signal to the LCD includes both intensity and color information of the HDR video. The process for preparing the input signal to this system is as follows:

- Load an HDR image in RGB space.
- Tone-map the HDR image using Reinhard tone mapping technique [39] to generate RGB_LCD.
- Extract luminance channel Y from HDR image: (Y = 0.2126 R + 0.7152 G + 0.0722 B), and normalize it to [0, 1].
- Evenly split pixel values between projector and LCD, Y_projector = Y^0.5.
- Simulate point spread function of projector (Y_lightfield), low pass filter Y_projector by a Gaussian filter (window size: 12×12, sigma=2).
- LCD signal: RGB_LCD/ Y_lightfield.

The subjective evaluations were conducted in a room complying with the ITU-R BT.500-13 Recommendation [40]. Prior to the actual experiment, a training session was shown to the observers to familiarize them with the rating procedure. The stimuli were designed based on the Double-Stimulus (DS) method [40]. In particular, after each 10-second long reference video, a 3-second gray interval was shown, and then followed by the 10-second long distorted video. Another 4-second gray interval was allocated after the test video, allowing the viewers to rate the quality of the test video with respect to that of the reference one. The scoring is based on discrete scheme where a numerical value from 0 (worst quality) to 10 (identical quality) is assigned to each test video representing its quality with respect to the reference video. Note that in order to stabilize the subjects' opinion, a few dummy video pairs were presented at the beginning of the test but the collected scores for these videos were discarded from the final results. Eighteen adult subjects including ten males and eight females participated in our experiment. The subjects' age range is from 19 to 35 years. Prior to the tests, all the subjects were screened for color blindness using the Ishihara chart and visual acuity using the Snellen charts.

3.2.2.2 Objective Tests

In the objective tests, both HDR quality metrics and LDR quality metrics with extensions to HDR domain are included.

In order to meaningfully use LDR metrics to evaluate the quality of HDR content, two methods of extending LDR metrics to HDR content are used in the test, Perceptually Uniform (PU) encoding [41] and multi-exposure method [42]. PU encoding transforms luminance values in the range of 10^{-5} cd/m^2 to 10^{8} cd/m^2 into approximately perceptually uniform LDR values, so that popular LDR quality metrics, like PSNR and SSIM, are capable of handling all luminance levels visible to the human eyes [41]. In multi-exposure method, the HDR image is tone-mapped with different exposures, uniformly distributed over the dynamic range of the image. The quality of each exposure is computed by a LDR metric and the mean of all quality scores from different exposures is the quality score of the original HDR image by this LDR metric [42].

The LDR metrics used in our experiment are PSNR, SSIM [43], and VIF [44]. Among the existing HDR metrics, HDR-VDP-2 is used in our experiment, as it is the state-of-the-art full-reference metric that works for all luminance conditions (both LDR and HDR) [45].

3.2.3 Results and Discussions

After subjective tests, the outlier subjects were detected based on the ITU-R BT.500-13 recommendation [40]. No outlier was detected in this test. The MOS for each impaired video was calculated by averaging the scores over all the subjects with 95% confidence interval. This score represents the overall rating of quality for each impaired video.

Table 3.3 summarizes the results of the correlation between the objective quality scores and the subjective MOSs. In order to evaluate each metric's accuracy, the Pearson Correlation Coefficient (PCC) and Root Mean Square Error (RMSE) are calculated between MOS values

and the obtained objective quality scores. The Spearman Rank-Order Correlation Coefficient (SCC) is also computed to estimate the monotonicity of results. The PCC and SCC in each column are calculated over the entire video data set. The results are reported based on three impairments categories: a) AWGN, intensity shifting, salt & pepper noise, and low pass filtering, b) compression artifacts, and c) all the impairments used in the study.

Table 3.3 Correlation of subjective responses with objective quality metrics.

Metric/Method	Impairments: AWGN, intensity shifting, salt & pepper noise, and low pass filtering			Impairment: compression, QP: 22, 27, 32, 37			Impairments: AWGN, Intensity shifting, salt & pepper noise, low pass filtering, and compression		
	Pearson Correlation	***Spearman Correlation***	***RMSE***	***Pearson Correlation***	***Spearman Correlation***	***RMSE***	***Pearson Correlation***	***Spearman Correlation***	***RMSE***
HDR-VDP-2	0.390	0.365	9.568	0.872	**0.958**	0.449	0.120	0.182	6.682
PSNR (PU encoding)	0.689	0.381	10.356	0.572	0.736	0.627	0.735	0.738	2.470
SSIM (PU encoding)	0.572	0.486	8.590	0.828	0.899	0.563	0.351	0.123	6.865
VIF (PU encoding)	**0.973**	**0.8492**	8.557	0.833	0.890	0.764	**0.916**	**0.948**	**0.711**
PSNR (Multi-Exposure)	0.889	0.499	1.292	0.685	0.735	1.309	0.885	0.861	1.370
SSIM (Multi-Exposure)	0.723	0.474	10.277	0.741	0.760	1.012	0.673	0.629	3.851
VIF (Multi-Exposure)	0.931	0.696	**0.550**	**0.899**	0.926	**0.398**	0.731	0.846	1.702

A few observations can be made from Table 3.3: 1) In the presence of the AWGN, intensity shifting, salt & pepper noise, and low pass filtering distortions, VIF with PU encoding yields the best performance compared to other metrics. 2) In the presence of the compression artifacts, HDR-VDP-2 and Multi-Exposure VIF outperform all other tested metrics. 3) Overall, in the presence of the tested distortions, VIF with PU encoding shows the best performance in predicting the quality of HDR videos compared to other tested metrics. 4） A possible explanation of VIF performing better is that it measures the mutual information between the original and distorted signal, thus it is less affected by dynamic range differences.

Figure 3.4 shows the objective quality metric results versus subjective test results in the presence of all the impairments used in this study (i.e., AWGN, intensity shifting, salt & pepper noise, low pass filtering, and compression).

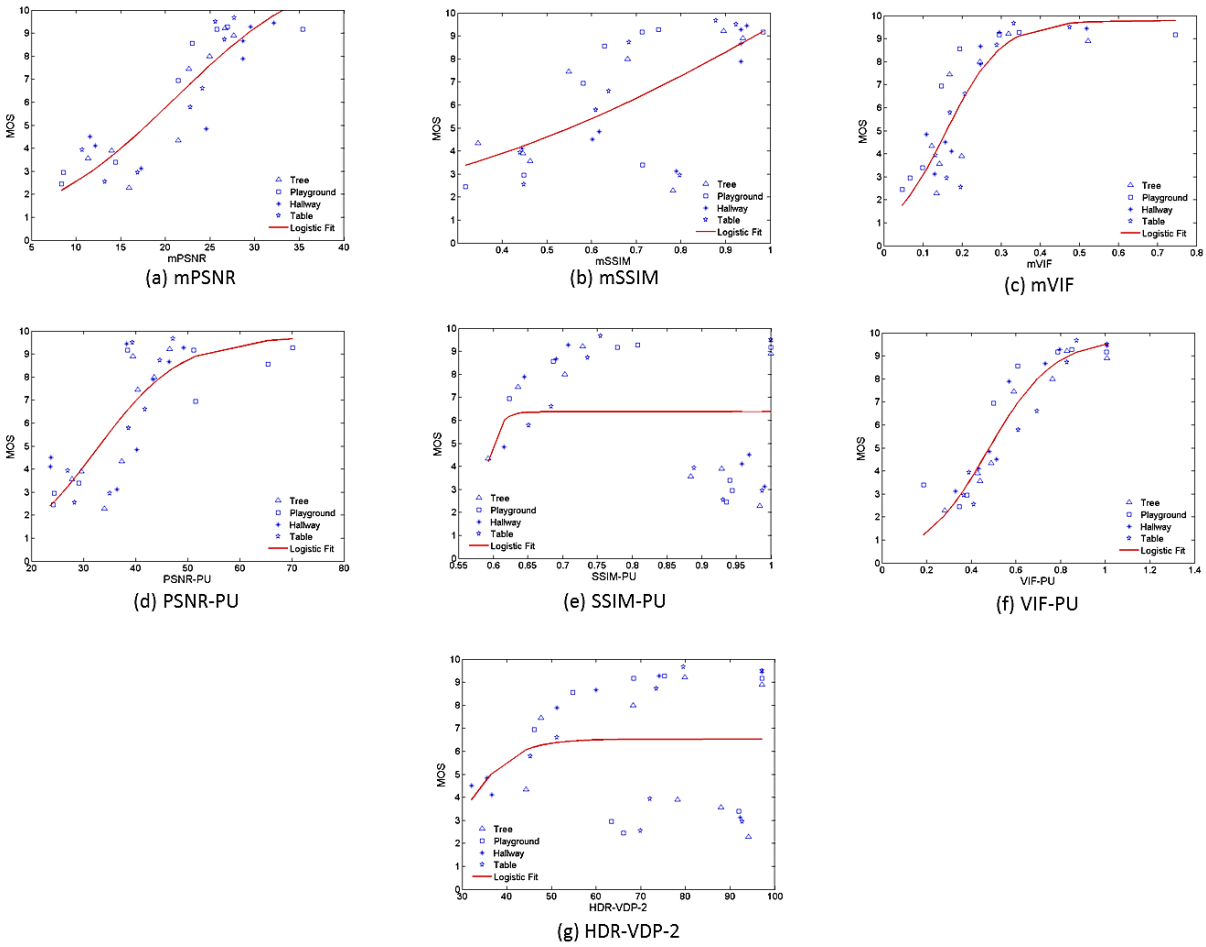


Figure 3.4 Subjective results versus objective opinions. All figures show the results in the presence of AWGN, intensity shifting, salt & pepper noise, low pass filtering, and compression. (a) PSNR (Multi-Exposure), (b) SSIM (Multi-Exposure), (c) VIF (Multi-Exposure), (d) PSNR (PU encoding), (e) SSIM (PU encoding), (f) VIF (PU encoding), and (g) HDR-VDP-2.

3.3 Conclusion

In Section 3.1, we evaluated the performance of the HEVC test model with the state of the art compression standard, H.264/AVC, for compressing HDR content. Configuration settings used for this study were chosen carefully to represent similar scenarios and ensure a fair comparison. Our experiment results show that HEVC outperforms H.264/AVC by 22.47% to

58.61% in terms of bitrate (with same PSNR) or 1.02 dB to 4.88 dB in terms of PSNR (for the same bitrate). However, it is worth noting that the current available version HEVC (at the timing of writing), July 2014 draft [13], is not yet optimized for HDR content and there are still many critical points to be carefully explored for a more effective and efficient HDR coding chain. For example, subjective tests in [46] have been performed on HDR sequences before and after encoding using HEVC. It was concluded that the downsampling/upsampling between chroma (4:4:4 and 4:2:0) created significant artifacts on HDR videos. Also, the best transform scheme to convert floating point HDR formats to 12 bits or higher bit depth integer representation accepted by the core of HEVC is not yet known. This needs to be investigated and standardized in the codec to support HDR content.

In Section 3.2, we report the performance of exiting quality metrics in evaluating the quality of HDR content. In the study, we tested not only existing HDR quality metrics, but also the proposed methods to extend LDR quality metrics to predict the quality HDR videos. Experiments results show that in the presence of compression distortions, HDR-VDP-2 out performs all other metrics. Overall VIF using PU encoding yields the best performance.

The findings of these studies could be utilized in optimization of HEVC standard for HDR coding and compression.

Chapter 4: Eye Tracking Study of HDR Content

HDR technologies have demonstrated that they can play an influential role in the design of cameras and consumer display products. Understanding the human visual experience of viewing HDR content is a crucial aspect of such systems. Although the visual experience of LDR technologies has been well explored, there are limited comparable studies for HDR content.

In this chapter, we present an eye tracking experiment using HDR content. The eye-tracking data were collected while individuals viewed HDR content in a free viewing task. Our study shows a clear subjective preference for HDR display when individuals are given a choice between HDR and LDR displays. Besides, the eye fixation data collected in this study are now available as a dataset named *DML-iTrack-HDR* [47]. This dataset can be used to develop, evaluate, and optimize visual attention models for HDR content.

4.1 Experiment

Eye tracking is a very well-established method of capturing an individual's looking behavior and visual attention. Data collected from an eye tracking experiment are necessary for the development, assessment and optimization of visual attention models. However, to the best of our knowledge, there are no publicly available HDR eye-tracking datasets yet[3]. In order to fill this need in the research community, we created such an HDR video eye tracking dataset called *DML-iTrack-HDR* [2]. Table 4.1 summarizes important parameters and details of our experiments. The summary of fixations for different HDR test videos is provided in Table 4.2.

[3] There were no HDR eye tracking datasets when this project was conducted. At the time of this writing, *DML-iTrack-HDR* is the only available HDR eye tracking dataset to the best of my knowledge.

Table 4.1 Summary of eye tracking experiments.

	Details	**Specification**
Participants	Number(M/F)	18(8/10)
	Age Range	21 - 30
	Test	Snellen chart, Ishihara chart
HDR display	Model	Dolby prototype
	Resolution	1024 x 768
	Peak luminance	2700 cd/m^2
	Screen size	66.5 x 49 cm
Eye tracker	Manufacturer	SMI
	Model	iView X RED
	Sampling frequency	250Hz
	Resolution accuracy	$0.4 \pm 0.03°$
	Setup mode	Stands alone; mounted on tripod
Video Presentation	Repetition	1
	Gray frame with dot	2s
	Stimuli duration	115s
	Dummy stimuli	Two HDR images in the very beginning
Image Presentation	Repetition	1
	Gray frame with dot	2s
	Presentation time	5s each image
Viewing condition	Task	Free viewing
	Ambient light	dim
	Environment	laboratory
Setup	Floor to eyes	110-130cm
	Floor to screen	94cm
	Floor to eye tracker	98cm
	Eye tracker to screen	67cm

Table 4.2 Summary of fixations.

Sequence	**No. of fixations**	**Fixation duration (ms)**	**No. of fixations per second per subject**
bistro01	252	339	2.78
fishing	512	387	2.30
park	713	325	2.71
mainmall	383	326	2.65
market	410	304	2.85
bistro03	239	405	2.34
balloon	252	312	2.80
carousel	651	274	3.20
playground	376	313	2.82
bistro02	938	332	2.61
average	**472.6**	**331.7**	**2.71**

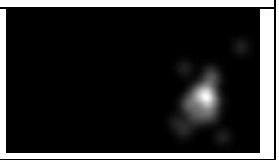
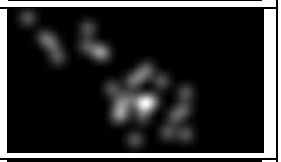
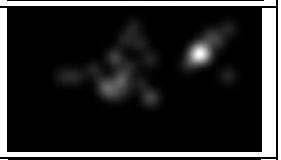
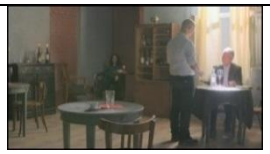
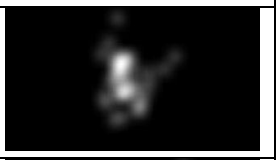
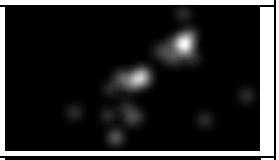
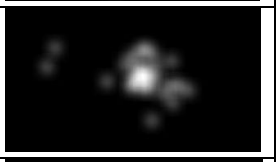
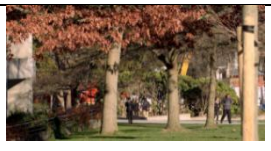
The number of fixations for each clip is the number of fixations from all observers; fixation duration of each clip is the average duration across all fixations.

4.1.1 Stimuli

The stimuli of the eye tracking experiment include twenty-three HDR images and ten HDR video clips. Twenty-three HDR images with various contents were selected for the study. They are in non-compressed Radiance RGBE format. Ten HDR video sequences were selected from three different datasets in this study: a) Technicolor [8], b) Froehlich et al. [6], and c) DML-HDR [37]. Table 4.3 provides details about the videos used in the experiments. The Technicolor HDR videos were captured by a rig of two Sony F3 or F56 cameras with simultaneous low and high exposure settings [8]. The HDR dataset provided by Froehlich et al. [6], were captured by Alexa cameras, a CMOS sensor based motion picture camera made by Arri. The DML-HDR dataset was captured at the University of British Columbia with RED Scarlet-X cameras, which can capture dynamic range up to 18 stops [5].

Content were chosen in order to span a wide variety of scenes that included day or night lighting conditions, different ranges of motion (i.e., minimal motion or fast moving objects) and a wide range of color spectrums. We limited the social context of the scenes, and those scenes that did include social components (i.e., people or human interaction) were kept neutral in nature, in order to effectively isolate different picture elements.

Table 4.3 Video sequences used in eye tracking experiments.

HDR Video Clip	Video Specification	Source Database	Description	Clip Snapshot	Fixation Density Map Snapshot
Balloon	200 frames 30 fps 1920 * 1080	Technicolor [8]	Exterior Medium color spectrum Slow global and local motion		
Market	400 frames 50 fps 1920 * 1080	Technicolor [8]	Exterior High illumination High color spectrum Static scene with slow motion		
Bistro 01	151 frames 30 fps 1920 * 1080	Froehlich et al. [6]	Interior High contract with local bright sunlight at the window Single moving object and slow motion		
Bistro 02	300 frames 25 fps 1920 * 1080	Froehlich et al. [6]	Interior High contract scenery with local bright sunlight at the window		
Bistro 03	170 frames 30 fps 1920 * 1080	Froehlich et al. [6]	Interior Medium illumination		
Carousel	339 frames 30 fps 1920 * 1080	Froehlich et al. [6]	Exterior scene at night Fast moving colorful objects Light sources with changing color		
Park	439 frames 30 fps 1920 * 1080	Froehlich et al. [6]	Exterior scene at night Wide color spectrum Fast motion		
Fishing	371 frames 30 fps 1920 * 1080	Froehlich et al. [6]	Sunlight scene Sunlight reflection on water surface		
Playground	222 frames 30 fps 2048*1080	DML-HDR [37]	Sunlight exterior scene High illumination High color spectrum Fast motion		
Mainmall	241 frames 30 fps 2048*1080	DML-HDR [37]	Medium illumination Slow local motion		

4.1.2 HDR Prototype Display

Experiments were performed using a Dolby prototype HDR TV display that consists of a projector with the resolution of 1024×768 at the back and a 40-inch full HD LCD placed in the front (see Figure 4.1). To deliver HDR viewing experience, the HDR video is processed to generate two calibrated streams, which are sent to the projector and the LDR respectively, according to the procedure described in [6]. The LCD screen is fed by a color stream, while a calibrated luminance stream is sent to the projector (A description of how to generate these two signals can be found in Section 3.2). The maximum brightness level achieved by this HDR display system prototype is 2700 cd/m^2.

Figure 4.1 Prototype HDR display system.

4.1.3 Reference LDR Display

In order to let participants compare HDR display to LDR display, an LDR display is used to present tone-mapped sequences. The LDR display used in this study is a Hyundai 42-inch LCD display. The display was calibrated to white point 6500K and peak brightness 120 cd/m^2 before the study, to represent a typical LDR screen available on the market.

For each HDR video sequence, a corresponding LDR video was generated using the photographic tone reproduction method proposed by Reinhard et al. [39] to show on reference LDR display. Since the image based tone-mapping method processes each frame independently,

flickering appears in some sequences as a result of abrupt changes in the mapping between consecutive frames. As such, an additional temporal coherency algorithm for video tone mapping proposed by Boitard et al. [48] was used, in combination with photographic tone reproduction to generate LDR videos with temporal coherency.

4.1.4 Eye Tracking System

Eye movements of participants were tracked using the SensoMotoric Instruments (SMI) iView X RED system. The eye tracker was mounted on a tripod and participants were seated in a chair that allowed their eye height to be adjusted to meet the set up requirements of the SMI system. This setup is shown in Figure 4.3 and measurements of setup is listed in Table 4.4. The sampling frequency of the SMI is 250 Hz and the resolution accuracy is $0.4 \pm 0.03°$.

Figure 4.2 Eye tracker from SensoMotoric Instruments (SMI).[49].

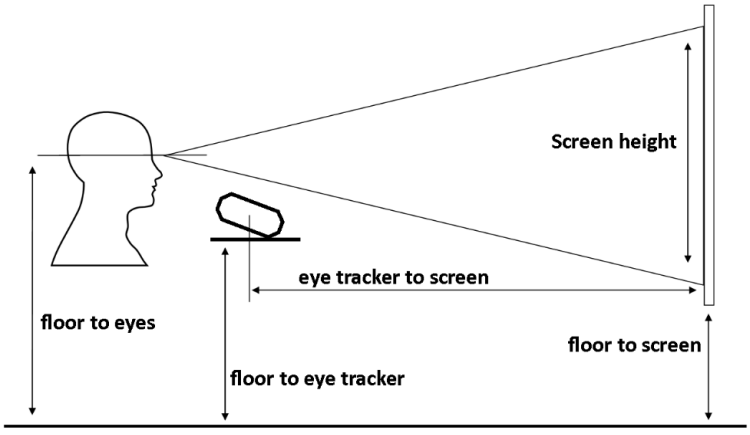


Figure 4.3 Set up of eye tracker.[50].

Table 4.4 Set up of eye tracking system.

Floor to eyes	110 cm
Screen height	49 cm
Screen width	66.5 cm
Floor to screen	94 cm
Floor to eye tracker	98 cm
Eye tracker to screen	67 cm
Eye tracker angle	10.2°

4.1.5 Participants

18 individuals (8 males and 10 females) participated in the study. All participants had normal or corrected to normal vision, and were screened for normal color vision. All subjects were naïve to the purpose of the experiment. Before each task, the participant's eye height and position was adjusted so that their eyes could be tracked accurately.

4.1.6 Procedure

Before each participant viewed the stimuli, a calibration was run which ensures accuracy of the eye tracking data. The calibration stage was repeated if the quality of the calibration was not satisfactory. Each participant was asked to 'free-view' all images and videos in the stimuli. Each video was presented at its native frame rate and each image was presented for 5 seconds. Before each image or video, participants were asked to fixate on a dot presented at one of the four corners of a neutral gray background. Note that by requiring participants to start each trial at one of the corners of the screen, we ensured that participants were free to choose where to first begin looking at the material presented on the displays, thereby avoiding any artificial center bias for viewing images and videos [17]. Center-bias means that a majority of fixations happen to be near the center of scenes, whatever the salience of the content. It has a number of reasons [17] and there is no effective way proved to eliminate the effect. In our study, by adding the corner

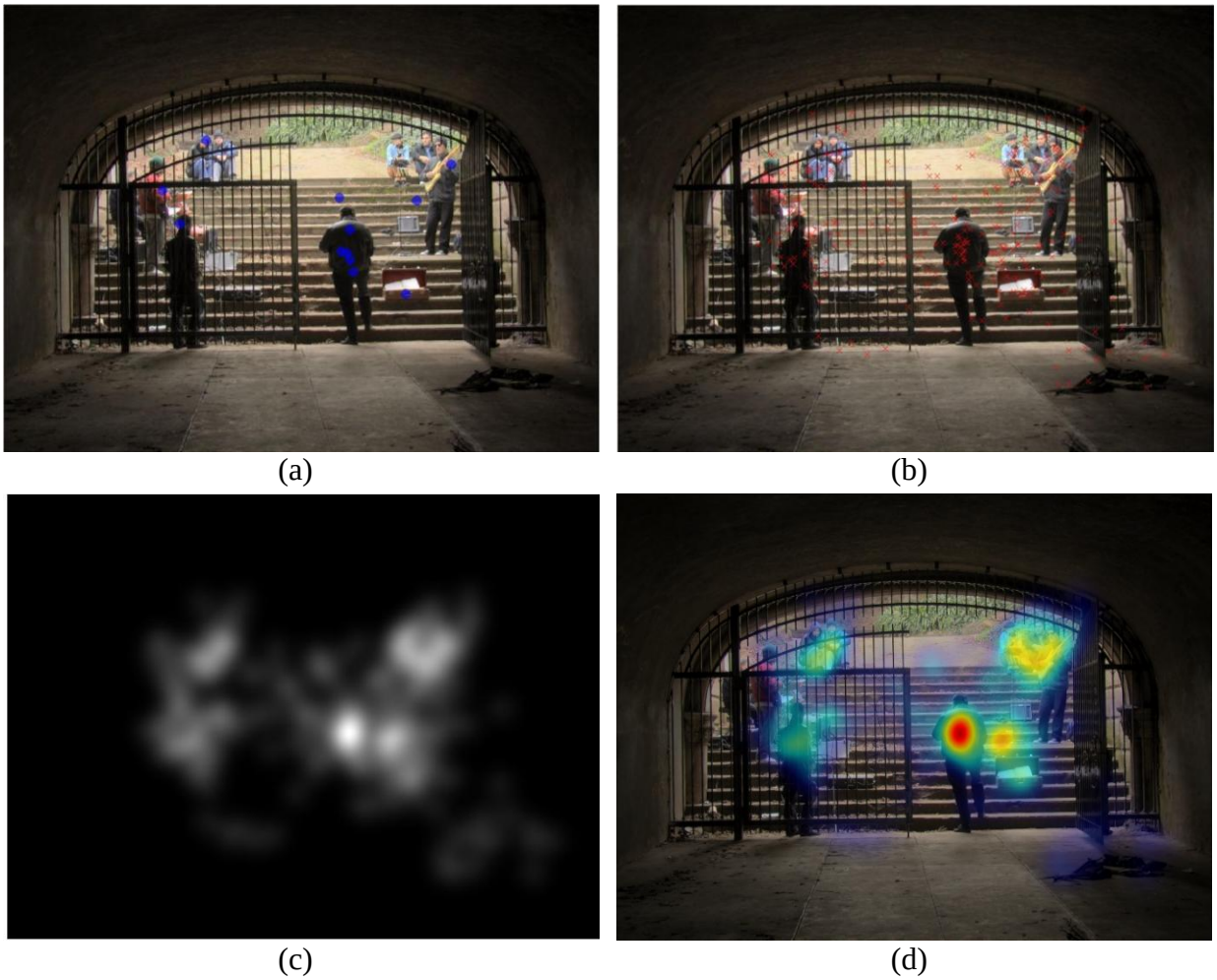
fixation, we think a change may be induced to the results, especially for the fixations just after the clip onset. The corner fixation dot was presented for 2s after each video and the location of the dot was randomized.

Each subject was asked to view the same content on both displays. To avoid a bias of naming of displays, the displays were simple labeled as display A and display B.

4.2 Fixation Density Map

Fixations hits from the eye tracking experiment can be post-processed into FDMs and the average FDM over all observers are considered to be the reliable ground truths of human visual attention. However, there is not yet a standardised methodology for the calculation of FDM. Thus, we follow the commonly used best-practice in research community in generating FDM from our eye tracking experiments.

The process of generating FDM consists the following steps. First, fixation positions are filtered out form all the gaze points recorded by eye tracker. Then, a fixation map per observer is computed from the fixation positions. Fixation maps are then averaged over all observes for each picture. Finally, the fixation map is filtered with a 2D Gaussian kernel to account for eye tracker inaccuracies as well as the decrease in visual accuracy with increasing eccentricity from the fovea [51]. The standard deviation of Gaussian filter is determined by the number pixels correspond to 1 degree of visual angle [32]. Figure 4.4 shows the FDM of one of the HDR images in our eye tracking experiments.



(a) (b) (c) (d)

Figure 4.4 Fixations and fixation density map (FDM). (a) An illustration of the fixations for one participant collected during a 5s free-viewing task. (b) All fixations for 18 different participants who viewed this image in the eye tracking study. (c) Fixation Density Map (FDM) obtained by convolving the fixation map with a 2D Gaussian filter. This maps represents the areas of interest. (d) Fixation Density Map (FDM) on top of the original image. The warmer the color is, the more fixations detected in that area.

4.3 Conclusions

The presented study in this Chapter examines the impact of LDR and HDR displays on subjective and objective measures. Objective eye movement data indicated that naive observers were unaffected by the HDR versus LDR displays. Both displays had a similar effect on frequency and temporal eye movement data, specifically, the number of fixations to areas of interest (AOIs), the time spent fixating those AOIs, and the frequency of revisits to those AOIs. Furthermore, when the two displays were shown to subjects directly against each other, there was a consistent and unanimous preference for the HDR displays. Participants reported

preferring the HDR monitor over the LDR monitor for its superior 'clarity and life-like detail' or because objects were 'looking more real'.

Collectively these data indicate a reliable subjective preference HDR displays emerges when a direct contrast to an LDR display is available, and does not arise from or lead to an objective performance difference in visual attention as measured by eye movement behavior.

Besides, the eye fixation data collected in this study serve as ground truth for evaluation of saliency detection algorithms, which is explained in Chapter 5.

Chapter 5: HVS Based Visual Attention Model for HDR Content

As described in Chapter 2, computational models of visual attention can predict the most relevant and important areas of images or videos presented to human eyes. Understanding how humans perceive HDR, in other words coming up with a visual attention model that best represents such content, is essential to many different stages of the HDR imaging pipeline such as HDR image and video capturing, compression, content resizing, and displaying.

In the last two decades, many models have been proposed to simulate the bottom-up process. Itti, Koch & Niebur proposed a visual attention model using three feature channels: color, intensity, and orientation for LDR images [52]. This model has become the benchmark for comparing alternative models. Itti, Dhavale & Pighin [53] extended this model to LDR video content by adding two temporal feature channels: flicker and motion.

Apart from the spatio-temporal saliency models proposed in [52] and [53], there are also other visual attention models proposed solely for images or for both image and video content. According to [54], there are at least 65 computational visual attention models proposed over the last two decades. However, to the best of our knowledge, the only model has been proposed for HDR image content is the Contrast Features (CF) model in [55].

Although the LDR state-of-the-art models have shown promising results on different types of LDR content, they neglect the HVS properties related to wide luminance ranges and rich color gamut associated with HDR content. The CF model addresses the fact that human eyes are sensitive to contrast, but the color perception of HDR content is not considered; besides, the CF model is not applicable to video content. In this chapter, we address these shortcomings by proposing a new saliency detection method that detects the most visually important areas of HDR images and videos. The proposed model takes into consideration the characteristics of HVS

under a wide luminance range and makes predictions of visual attention by combining spatial and temporal visual features. An analysis of eye movement data affirms the utility of the proposed model. Comparisons employing three well known quantitative metrics show that the model proposed here substantially improves predictions of visual attention of HDR.

5.1 Benchmark Models and Limitations

This section provides a brief background of the Itti et al.'s model [53], the most widely used LDR approach, and the CF model [55], an extension of the Itti et al.'s model for HDR images. These two state-of-the-art models are selected as the benchmarks which the proposed model is compared against.

5.1.1 Itti et al.'s Model and its Limitations

Itti et al.'s model described in [53] is an implementation of the bottom-up framework proposed by Koch and Ullman [56]. Figure 5.1 depicts the structure of Itti et al.'s model for video content. First of all, the visual input is decomposed into several parallel channels (color, intensity, orientation, flicker and motion) and a Gaussian pyramid is formed in each of the channels. The Gaussian pyramid has nine spatial scales, $\sigma = [0, \ldots, 8]$, where $\sigma = 0$ represent the original resolution of frame and level $\sigma = 8$ has a resolution of $1/2^{\sigma} = 1/256$ of the input frame.

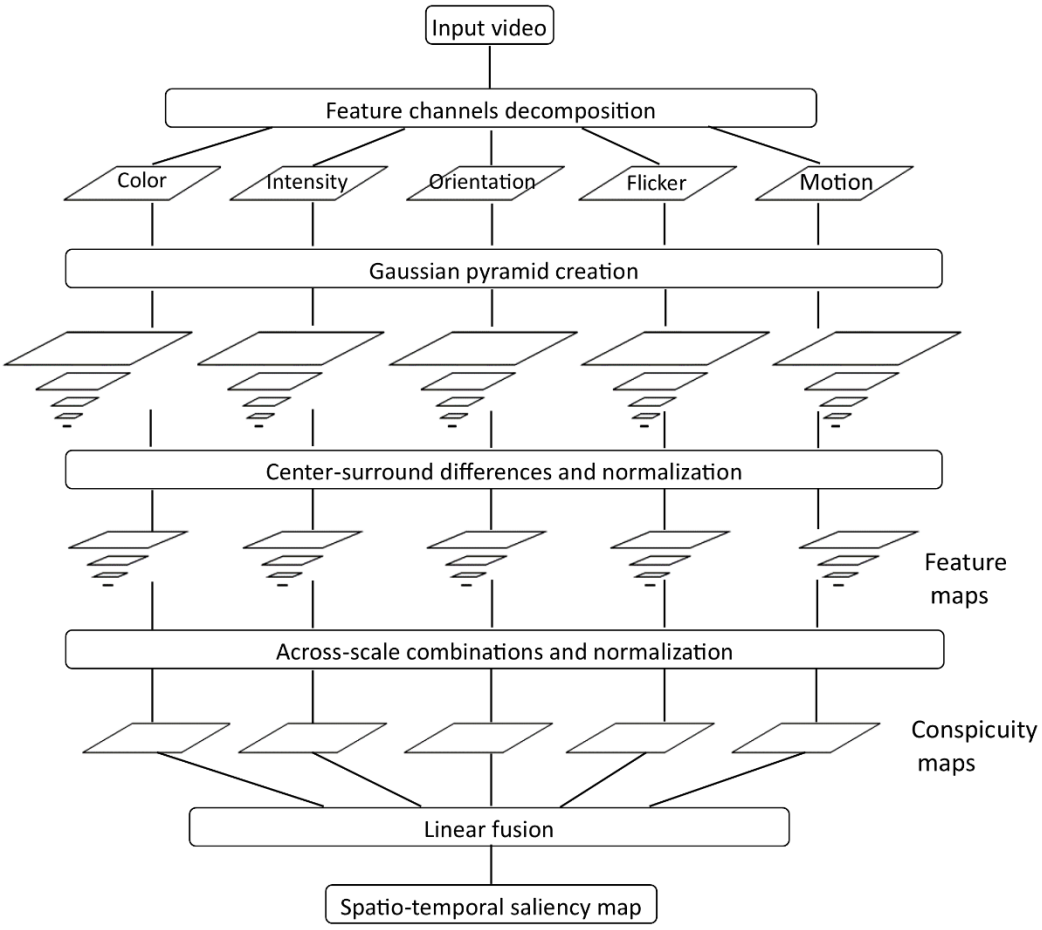


Figure 5.1 Architecture of Itti et al.'s visual attention model.

With r_n, g_n and b_n representing the red, green, and blue channels of input frame n, the intensity map is computed as:

$$I_n = \frac{r_n + g_n + b_n}{3} \tag{5.1}$$

I_n is used to create the Gaussian pyramid of intensity channel, $I_n(\sigma)$.

Two opponent color pyramids, red-green (RG) and blue-yellow (BY), are created by:

$$R_n = r_n - \frac{g_n + b_n}{2} \tag{5.2}$$

$$G_n = g_n - \frac{r_n + b_n}{2} \tag{5.3}$$

$$B_n = b_n - \frac{r_n + g_n}{2} \tag{5.4}$$

$$Y_n = r_n + g_n - 2(|r_n - g_n| + b_n) \tag{5.5}$$

$$RG_n = R_n - G_n \tag{5.6}$$

$$BY_n = B_n - Y_n \ . \tag{5.7}$$

r_n, g_n and b_n used in above formulas are normalized by I_n to decouple hue from intensity [52].

The orientation pyramid $O_n\ (\sigma, \theta)$ is obtained by convolving the intensity pyramid with Gabor filters, where $\theta = \{0^o\ , 45^o\ , 90^o\ , 135^o\ \}$. Gabor filter is the product of a cosine grating and a 2D Gaussian filter, which models the sensitivity profile of orientation-selective neurons in primary visual cortex [57].

The flicker pyramid is computed by subtracting the intensity I_n of current frame and intensity I_{n-1} of previous frame.

The motion pyramid is obtained from spatially-shifted differences between orientation pyramid from current and previous frame.

$$R_n(\sigma, \theta) = |O_n\ (\sigma, \theta) \cdot S_{n-1}(\sigma, \theta) - O_{n-1}\ (\sigma, \theta) \cdot S_n(\sigma, \theta)| \tag{5.8}$$

$S_n(\sigma, \theta)$ is the shifted pyramid and is obtained using the same Gabor filters as in the orientation channel. A fast implementation uses the intensity pyramid rather than the orientation pyramid in equation (5.8). This fast implementation was used in the original publication of this method [53]. Therefore, in the comparison of performance, the same implementation is utilized for the motion pyramid.

These are the five feature channels used in the visual attention model for video content based on Itti et al.'s model. Among these five channels, intensity, color and orientation are spatial channels, and flicker and motion are temporal channels. When the input is an image, only the intensity, color and orientation channels are used.

In the next step, early visual features are extracted to form a set of topographic feature maps in each channel by center-surround operations modeling the receptive fields (RF) of the human eye. Center-surround operation is calculated as the across-scale difference $\ominus$ between a center fine scale c and a surrounding coarser scale s in the Gaussian pyramid. Here $c = \{2, 3, 4\}$, $s = c + \delta$, and $\delta = \{3,4\}$.

$$Intensity: I(c,s) = |I(c) \ominus I(s)| \tag{5.9}$$

$$RG: RG(c,s) = |RG(c) \ominus RG(s)| \tag{5.10}$$

$$BY: BY(c,s) = |BY(c) \ominus BY(s)| \tag{5.11}$$

$$Orientation: O(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)| \tag{5.12}$$

$$Flicker: F(c,s) = |F(c) \ominus F(s)| \tag{5.13}$$

$$Motion: R(c,s,\theta) = |R(c,\theta) \ominus R(s,\theta)| \tag{5.14}$$

After center-surround operations, a normalization operator $N(\cdot)$ is utilized on each scale. This normalization operator suppresses maps containing more noise and promotes maps with strong variations, thus simulating the local competition between neighboring salient locations.

After normalization, feature maps of each feature are summed over using across-scale addition $\oplus$, to obtain five separate "conspicuity maps".

$$Intensity: \bar{I} = \oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} N(I(c,s)) \tag{5.15}$$

$$Color: \bar{C} = \oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} [N(RG(c,s)) + N(BY(c,s))] \tag{5.16}$$

$$Orientation: \bar{O} = \sum_{\theta} N(\oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} N(O(c,s,\theta))) \tag{5.17}$$

$$Flicker: \bar{F} = \oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} N(F(c,s)) \tag{5.18}$$

$$Motion: \bar{R} = \sum_{\theta} N(\oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} N(R(c,s,\theta))) \tag{5.19}$$

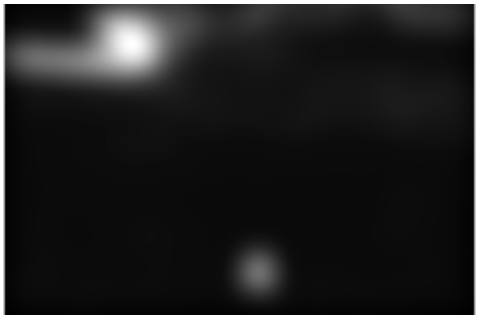
All conspicuity maps are normalized again and combined into one master saliency map:

$$S = \frac{1}{5}\left(N\left(\bar{I}\right) + N\left(\bar{C}\right) + N\left(\bar{O}\right) + N\left(\bar{\mathrm{F}}\right) + N\left(\bar{R}\right)\right) \tag{5.20}$$

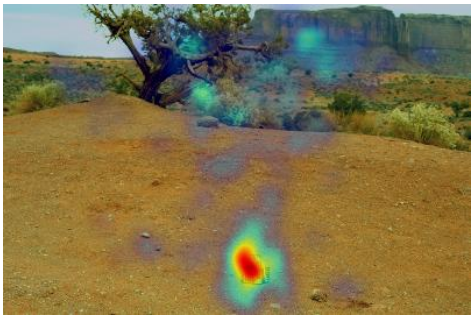
Although this model has shown good results with various LDR images and videos, it has the following drawbacks when applied to HDR content:

- If there is a very bright area in an HDR image, Itti et al.'s model detects that area as the most saliency area, but fails to detect other salient regions. This is evident in Figure 5.2, where the saliency map using Itti et al.'s method does not match the human fixation map obtained from the eye tracking experiment.
- For HDR videos, the presence of very bright areas also heavily degrades the performance of temporal saliency map. As shown in Figure 5.3, both flicker channel and motion channel detect the highlight areas as the salient regions but not the areas such as the hand and the paper that are actually moving.
- The input signal is assumed to be perceptually uniform and linear RGB signal. However, HDR images and videos don't have perceptual uniformity. For this reason, the color, intensity and orientation feature maps are generally not able to capture all the corresponding salient regions (See Figure 5.4).
- Characteristics of the HVS is not taken into consideration in the modeling even though neurobiology of attention was carefully and thoroughly modeled.
- The fusion step simply involves averaging across the spatial and temporal channels. However, studies have revealed that motion is the strongest attractor of attention [58] [59].
- Parameters such as the viewing distance and the pixel density of display have been overlooked.

(a)



(b)



(c)

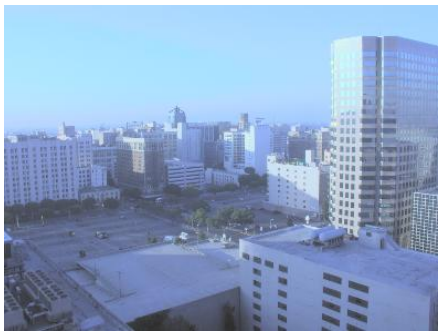
Figure 5.2 Limitations of Itti et al.'s model on HDR images. Saliency map using Itti et al.'s model. Itti et al.'s model detects the brightest area in an HDR image as the most salient area. (a) HDR image, (b) saliency map using Itti et al.'s method, (c) human fixation map from eye tracking experiment on top of original image (red areas represent the most attended areas).

(a)

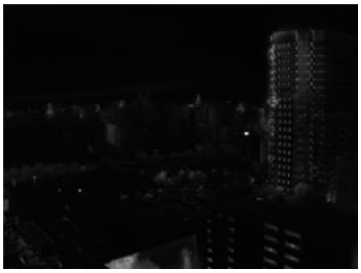
(b)

(c)

Figure 5.3 Limitations of Itti et al.'s model on HDR videos. (a) One fame of HDR clip, bistro01; (b) conspicuity map of flicker channel; (c) conspicuity map of motion channel.



(a)



(b)

(c)

(d)

Figure 5.4 Conspicuity maps of an HDR image using Itti et al.'s model. (a) Tone mapped image of the original HDR image, (b) (c) (d) conspicuity map of color, intensity, orientation using Itti et al.'s method.

5.1.2 Contrast Feature Model and its Limitations

In [55], Bremond, Petit & Tarel proposed an algorithm, denoted CF, for computing saliency maps of HDR images by defining new visual features within the framework of Itti et al. [52]. The main difference between this model and the framework of Itti et al. is using intensity normalized values for the intensity and orientation channels. The rationale is that the HVS is sensitive to contrasts rather than absolute intensity levels.

In Itti et al.'s model, the intensity feature map between center scale c and surround scale s is defined as: $I(c,s) = |I(c) \ominus I(s)|$, where "$\ominus$" stands for the across-scale difference between two maps. In CF, intensity contrast is used instead of the absolute intensity:

$$I'(c,s) = \frac{|I(c) \ominus I(s)|}{I(s)} \tag{5.21}$$

Itti et al.'s model defines orientation for angle θ as the difference between Gabor filters at scales c and s: $O(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)|$. In CF, a new definition is used to detect the

"boarders of boarders" so this feature is homogeneous to contrast. The new orientation feature map between center scale c and surround scale s is calculated as:

$$O'(c,s,\theta) = \frac{O(c,\theta)}{I(s)} \quad (5.22)$$

where $O(c,\theta)$ is the orientation obtained by convolving the intensity with the Gabor filter, and $\theta = \{0^o, 45^o, 90^o, 135^o\}$.

In the CF model, the color channel remains the same as in the original model, because the color feature is already normalized by intensity at every pixel in the Itti et al.'s model.

The predictions obtained by CF were compared with eye tracking experiment results and a better fit was shown compared to the predictions obtained by Itti et al.'s model. However, the following limitations exist in the CF model:

- The eye tracking experiment was conducted with a physical scene of objects and light sources presented in front of subjects, instead of using an HDR display.
- Only one scene/image was studied, so the robustness of the model is not fully validated.
- Some mechanisms of visual adaptations were taken into consideration, but there are other critical properties like intensity and color perception under wide luminance range, which are not modeled in the CF model.

5.2 HVS Based Visual Attention Model for HDR Content

Different from LDR content, HDR can describe a full color gamut and the entire range of luminance visible to a human observer. HDR images and videos store a truthful representation of the depicted scene.

To extend the bottom-up framework invented by Itti et al. to HDR content, we propose a new spatio-temporal saliency detection method as depicted in Figure 5.5. Compared to the

original bottom-up model, three new steps/modules are introduced in our proposed model to address the limitations of the original bottom-up method. First, we add a module denoted HVS model in Figure 5.5 before feature channels are extracted to model the human perception on HDR pixels. This HVS model mainly addresses two issues: the color perception and luminance perception under the wider HDR luminance range. Second, we use a method based on optical flow to obtain the temporal saliency map so the motion information is correctly extracted under various illumination conditions. Third, the original average fusion scheme is replaced by a dynamic fusion scheme which takes into account some known characteristics of the HVS regarding how human perceives the temporal and spatial cues. The following subsections describe these modules in detail.

5.2.1 HVS Model

The module of HVS model in the proposed method addresses the human perception of HDR pixel values. Color perception and luminance perception are tackled in parallel. The perception of colors is predicted by the Color Appearance Model (CAM), which describes how colors are perceived under given lighting conditions. In terms of luminance perception, two stages, Amplitude Nonlinearity and the Contrast Sensitivity Function (CSF), compensate for sensitivity variations of the HVS at different HDR light levels and spatial frequencies.

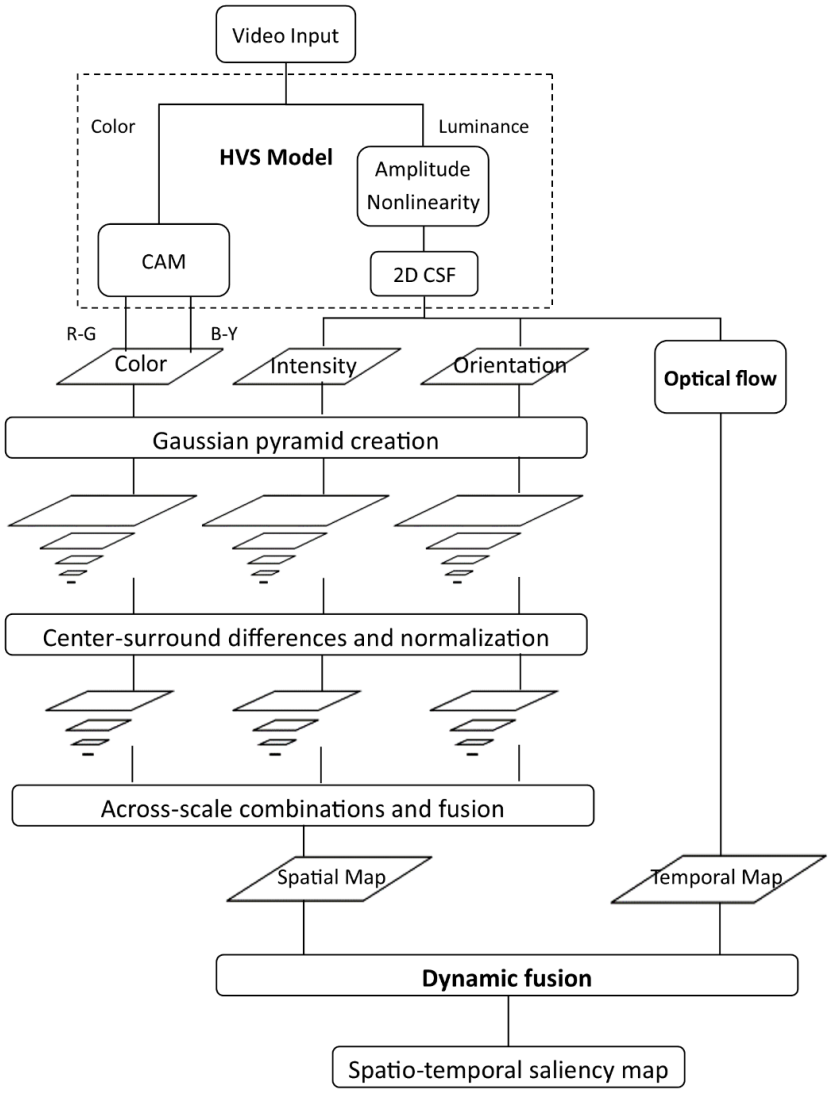


Figure 5.5 Architecture of the proposed visual attention model for HDR content.

5.2.1.1 Color Appearance Model

In Itti et al.'s approach, the two color opponent signals are obtained as linear combinations of R, G, and B. However, studies have shown that the HVS perceives colors differently under different lighting conditions. For example, colorfulness increases with higher luminance levels, which is known as the Hunt effect [60]. To model how the HVS perceives colors under different lighting conditions, CAMs are developed based on psychophysical studies. In our implementation, we use the CAM for extended luminance range, proposed by Kim et al.

[61], to make sure that color perception under a wide luminance range is correctly modeled (see Figure 5.5).

The reason we use the CAM proposed in [61] rather than other CAMs is that high luminance levels were included in the psychophysical experiments used to design this color model. For example, although the LUTCHI data set [62] has been widely used to optimize CAMs (e.g., CIECAM97 and CIECAM02, experiments were carried out with luminance levels mainly below 690 cd/m^2. In contrast, in [61], luminance levels up to 16,860cd/m^2 were included.

We follow the steps in [61] to obtain the two opponent signals (see Figure 5.6). First, chromatic adaptation transform is performed on incoming pixel triples in RGB using CIECAT02 to get white-adapted XYZ values. In the human eye cone cells are responsible for color perception. There are three types of cones each one sensitive to three different spectra: L-cones, M-cones, and S-cones, which are sensitive to long, medium, and short wavelengths. To emulate the color perception, tristimulus XYZ values are transformed into LMS cone space using the Hunt-Pointer-Estevez (HPE) transform described in [63]. Then, cones' absolute responses are modeled by:

$$\mathrm{L}' = \frac{L^{n_c}}{L^{n_c} + {L_a}^{n_c}} \,, \mathrm{M}' = \frac{M^{n_c}}{M^{n_c} + {L_a}^{n_c}} \,, S' = \frac{S^{n_c}}{S^{n_c} + {L_a}^{n_c}} \,, \tag{5.23}$$

where L_a is the absolute level of adaptation luminance measured in cd/m^2, and the exponent n_c is equal to 0.57, which is derived from the experiment. After modeling cone responses, opponent signals are derived as:

$$\begin{aligned} a &= \frac{1}{11}(11L' - 12M' + S') \;\; red - green \; channel \\ b &= \frac{1}{9}(L' + M' - 2S') \;\; yellow - blue \; channel \end{aligned} \tag{5.24}$$

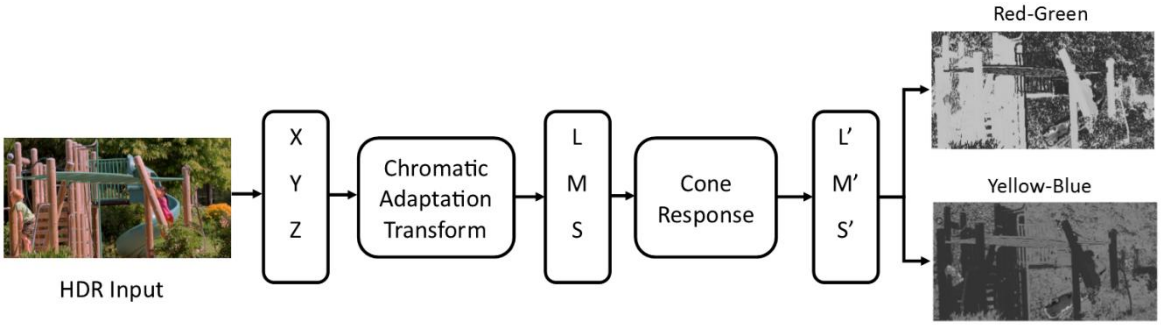


Figure 5.6 Obtain opponent color signals using color appearance model.

5.2.1.2 Amplitude Nonlinearity

Inside the HVS model of Figure 5.5, amplitude nonlinearity accounts for the non-linear response of the HVS to luminance.

The threshold contrast, i.e., minimum contrast to be noticed at a given luminance, decreases along with the increase of adaptation luminance, until adaptation luminance reaches around 100 cd/m^2 (Figure 5.7). When the adaptation luminance is higher than a certain level, roughly 100 cd/m^2, the threshold contrast remains constant. This is known as Weber's law, meaning the perception is proportional to the logarithm of the background intensity.

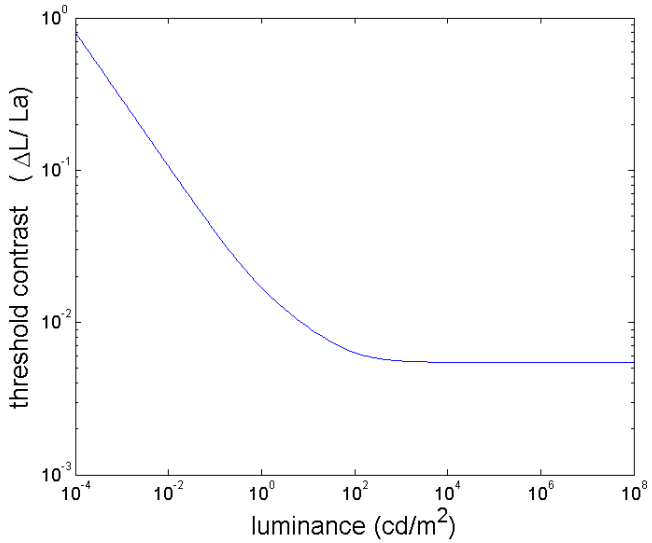


Figure 5.7 Contrast versus intensity (CVI) function. This figure shows the threshold contrast (minimum change of contrast can be detected by human eyes) for a given luminance value.

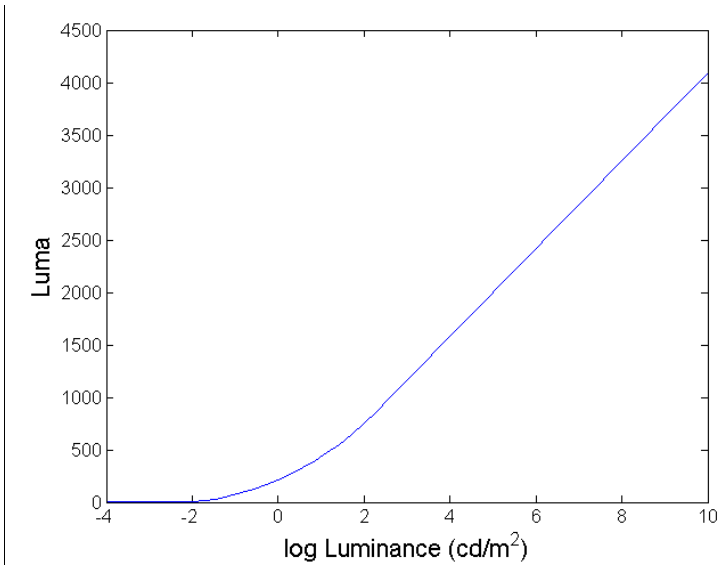


Figure 5.8 Luminance to luma mapping (TVI function). Formulas of this mapping are derived in [64].

The luminance range of HDR content covers the full visible range visible to the HVS. The response of the human eye in this range is not always linear nor always logarithmic, so it is necessary to model the variation of perception at different luminance levels. In the proposed model, we transform the input signal into units of Just Noticeable Difference (JND) in this stage, named amplitude nonlinearity, using the mapping approach presented in [64]. This mapping transforms luminance to the JND scaled space, noted luma, in which adding or subtracting a value of 1 means a just noticeable change to the human eye (Figure 5.8). This mapping contains three different functions depending on the intensity of luminance. At very low luminance levels, below L_1, linear mapping is used. At high luminance levels, above L_2, a logarithmic mapping is used, which corresponds to Weber's law. A power function segment exists between the two luminance thresholds L_1 and L_2. The luma value for adaptation luminance L_a is defined by:

$$luma(L_a) = \begin{cases} 769.18 \cdot L_a, & L_a < L_1 \\ 449.12 \cdot {L_a}^{0.16999} - 232.25, & L_1 \le L_a < L_2 \\ 181.7 \cdot ln(L_a) - 90.16, & L_a \ge L_2 \end{cases} \tag{5.25}$$

$$L_1 = 0.061843\ cd/m^2$$

$$L_2 = 164.1\ cd/m^2$$

The coefficients in above formula and the two luminance thresholds L_1 and L_2 are derived in [64].

5.2.1.3 Contrast Sensitivity Function

Amplitude nonlinearity compensates for the nonlinearity of luminance perception over the entire luminance range. Yet, there is another important property of luminance perception, the sensitivity change at different spatial frequencies, needs to be modeled but overlooked in the existing models.

The CSFs shown in Figure 5.9 and Figure 5.10 depict the visual sensitivity as a function of spatial frequency. The CSF used in the proposed model is developed by Daly [65]:

$$CSF(\rho, \theta, L_a, i^2, d, c) = P \cdot \min\left[S_1\left(\frac{\rho}{r_a \cdot r_c \cdot r_\theta}\right), S_1(\rho)\right], \tag{5.26}$$

where

$$r_a = 0.856 \cdot d^{0.14}$$

$$r_c = \frac{1}{1 + 0.24c}$$

$$r_\theta = 0.11 \cdot \cos(4\theta) + 0.89$$

$$S_1(\rho) = [((3.23\rho^2 i^2)^{-0.3})^5 + 1]^{-0.2} \cdot A_l \epsilon \rho e^{-B_l \epsilon \rho} \sqrt{1 + 0.06 e^{B_l \epsilon \rho}} \tag{5.27}$$

$$A_l = 0.801(1 + 0.7L_a^{-1})^{-0.2}$$

$$B_l = 0.3(1 + 100L_a^{-1})^{-0.15}$$

The parameters are as follows: ρ , spatial frequency in cycles per visual degree; θ, orientation; L_a, adaptation luminance in cd/m^2; i^2 stimulus size in deg^2, $i^2 = 1$; d, distance in meters; c, eccentricity, c = 0; ϵ, constant, $\epsilon = 0.9$; and P, absolute peak sensitivity, P = 250. The formulas for A_l and B_l contain the corrections as found in [66] after the correspondence with the author of the original publication.

Normalized CSF (Figure 5.10) for a particular adaptation luminance as depicted is obtained by normalizing the CSF with the peak sensitivity of this luminance.

$$nCSF\ (\rho, L_a) = \frac{CSF(\rho, L_a)}{max_\rho(\rho, L_a)} \tag{5.28}$$

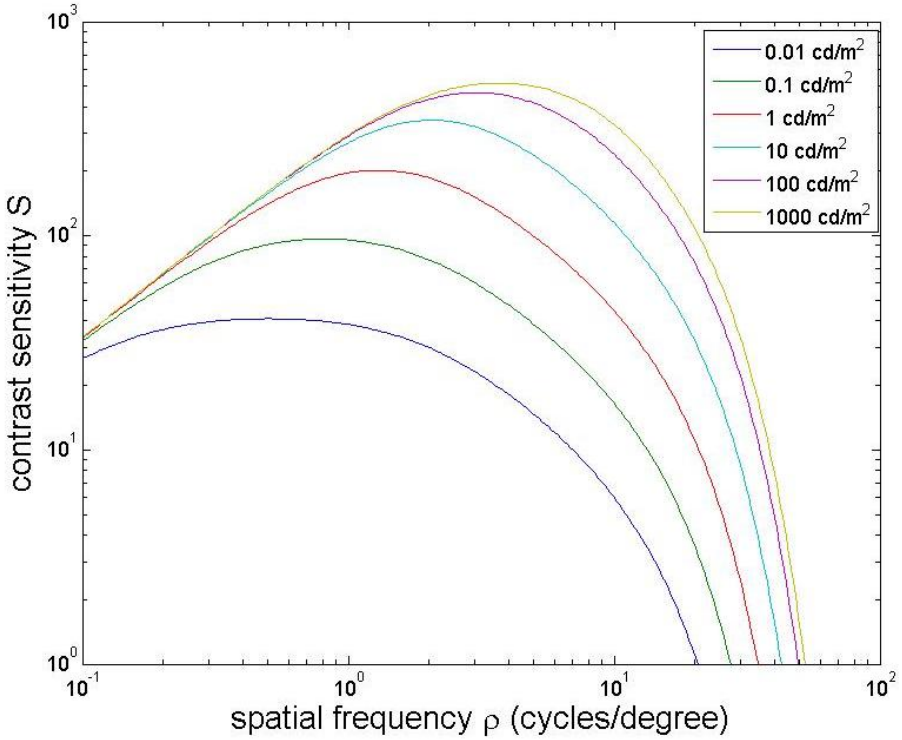


Figure 5.9 Contrast Sensitivity Function (CSF) depends on luminance. This figure shows CSF at six different adaptation luminance levels.

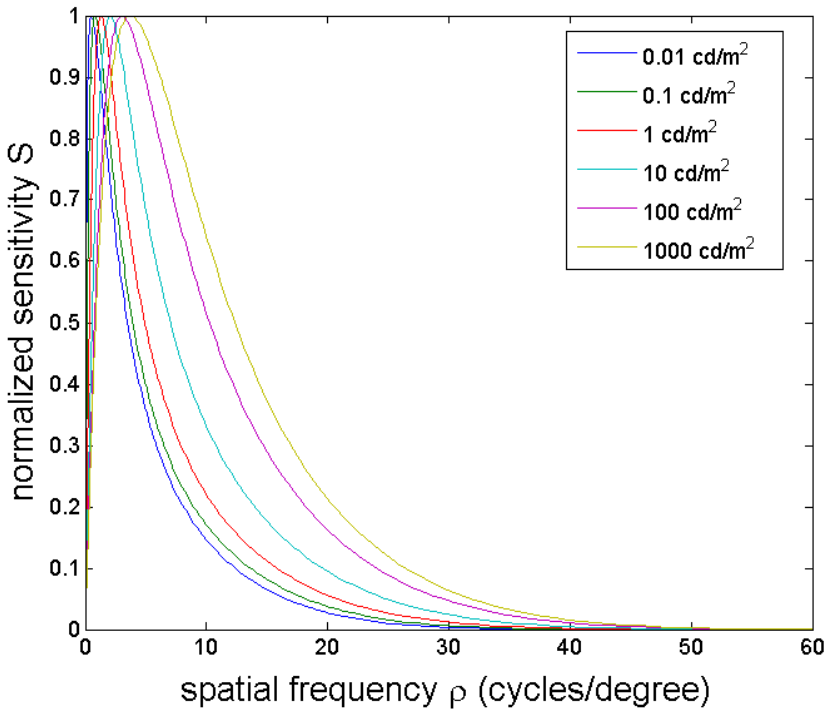


Figure 5.10 Normalized Contrast Sensitivity Function (CSF). This figure shows CSF at six different adaptation luminance levels. Normalized CSF far a particular adaption luminance is obtained by normalizing the CSF with the peak sensitivity of this luminance.

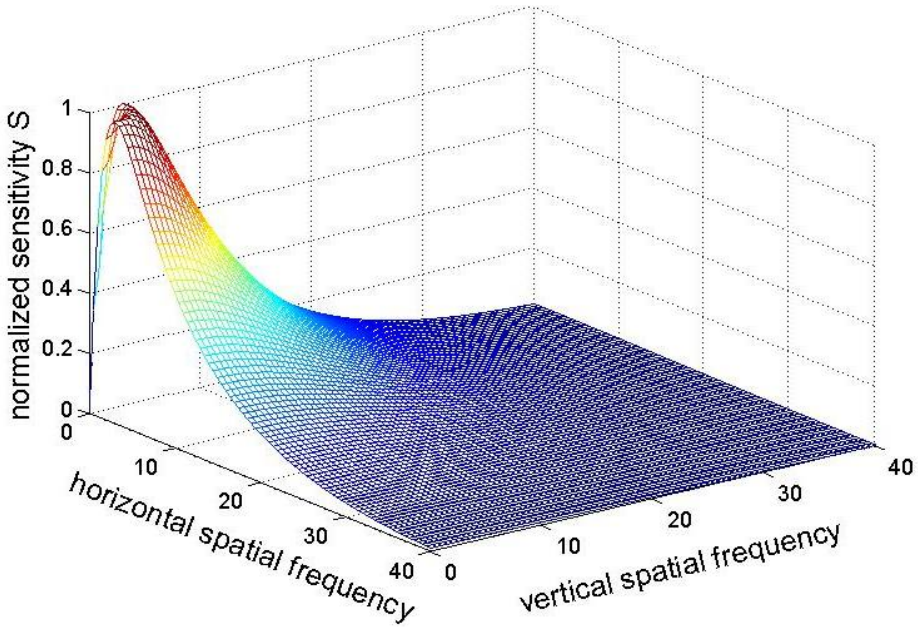


Figure 5.11 2D normalized Contrast Sensitivity Function (CSF). (La=100 cd/m^2)

There are a few things to notice in Figure 5.9 and Figure 5.10. First, the contrast sensitivity drops at very high and very low spatial frequencies, like a bandpass filter. Moreover, as the level of luminance increases, the peak of CSF shifts to higher spatial frequencies, meaning that certain frequencies become more visible at higher luminance levels.

Based on the above observations, in our proposed model, we used the 2D CSF proposed in [65] to filter the luma image, so that the spatial frequencies more senstivie to the HVS are enhanced. Meanwhile, the caluculation of spatial frequenies involves viewing distance, screen size and screen pixel density (pixel per inch), parameters overlooked in the existing models. Last but not least, the multi-CSF method presented in [67] is utilized to address the fact that a certain spatial frequency could have different degree of importance to the HVS at different luminance levels.

Ideally, in order to apply the CSF filter to an HDR image with a wide luminance range, we should use a separate CSF filter for every pixel, since each pixel has a different luminance level, and every luminance level corresponds to a different CSF with the peak at a certain frequency (as shown in Figure 5.9). Given that an HD image has more than two million pixels, this is computationally not feasible. For this reason, we adopt a more effective approach, the multi-CSF method described in [67]. As shown in Figure 5.12, after Amplitude Nonlinearity, the image/frame is filtered in the Fourier domain multiple times, each time using the CSF at a different adaptation luminance level, so that the more visible frequencies are enhanced. Since the shape of CSF remains constant for adaptation luminance larger than 1000 cd/m^2, CSF filters for $L_a = \{0.0001, 0.01, 0.1, 1, 10, 100, 1000\}\ cd/m^2$ are used (see Figure 5.12). Then, all filtered images are converted back to spatial domain. The final filtered image is obtained by interpolation between the two pre-filtered images closest to the adaptation luminance of each given pixel. For

example, if the original luminance of a pixel is $70 cd/m^2$, its value in the final filtered image is linear-interpolated from the filtered image with La = 10 cd/m^2 and the filtered image with La = 100 cd/m^2.

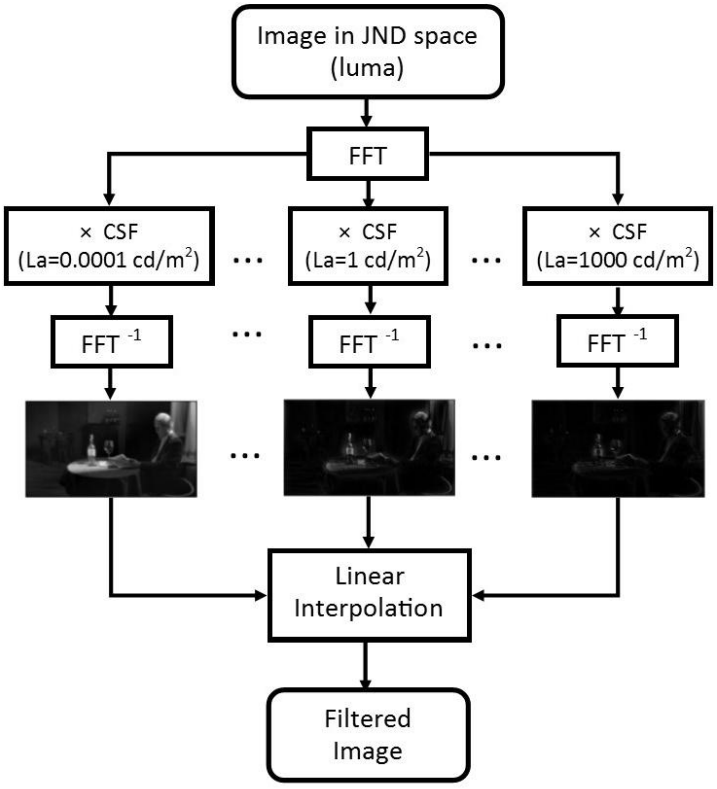


Figure 5.12 To apply CSF on image using the multi-CSF method. [67]. $L_a = \{0.0001, 0.01, 0.1, 1, 10, 100, 1000\}$ cd/m^2.

5.2.2 Optical Flow

In Itti et al.'s model, the temporal saliency map is derived from two features, flicker and motion. Flicker is the difference between the intensity of frames and motion is the spatially-shifted difference the intensity of frames. Our experiments with HDR videos show that neither of these two features can accurately represent the motion information in videos. To improve the accuracy of the temporal saliency map, we use an optical flow based approach to compute a

dense motion vector map between consecutive frames. The magnitude of the motion vector serves as the temporal saliency map.

Optical flow is the distribution of the apparent velocities of objects in an image. By estimating the optical flow between video frames, the velocities of objects in the video can be measured. Horn and Schunck first proposed dense optical flow in [68]. Their method is based on the assumption of intensity consistency and the smoothness of motion over the entire image.

Since optical flow relies on the assumption of intensity consistency between frames, videos with varying lighting in the sequence, such as the sky, tend to cause more artifacts. To make sure that the difference of exposure over time and the variation of lighting in sequences don't cause artifacts in the motion vector, we adopt the residual based optical flow presented in [69] [70]. In this approach, residuals are used for optical flow computation rather than the original frames; the residual in this case is the difference between an intensity image and its smoothed version, i.e., the high frequencies of the image.

In the proposed model, the Horn-Schunck method is used and the residuals are generated using a Gaussian filter. For each frame, a Gaussian filter is convolved with the intensity frame generated from the HVS model (see Figure 5.5). The residual is calculated by taking the difference between the original image/frame and Gaussian filtered image/frame. Figure 5.13 shows a frame from sequence *bistro02* and the temporal saliency map. The moving object in this frame is the walking man but Itti et al.'s method fails to detect this moving object and emphasizes on the bright window.

(a) (b) (c)

Figure 5.13 The temporal saliency map. (a) Example frame from sequence *bistro02*, (b) the temporal saliency map using the proposed approach, (c) the temporal saliency map from Itti et al.'s method.

5.2.3 Dynamic Fusion

After the temporal and the spatial maps are obtained, they have to be combined to form a unique master map. The best way to fuse these maps, which are generated from different visual cues, is still an open issue. Since the biological and neurological mechanisms for feature fusion in the HVS are not fully understood yet, we propose a dynamic fusion scheme for combining temporal and spatial saliency maps based on some known characteristics of the HVS. Figure 5.14 shows an example of a fused spatio-temporal saliency map using the dynamic fusion scheme.

- Motion is one of the most sensitive cues for the HVS. It is strongest attractor of attention among the commonly studied features [31].
- Nothdurft measured the relative saliency of targets defined by various single or double features, concluding that combined feature targets are more salient than single-feature targets [71].
- When watching videos, the human eyes are more sensitive to motion if motion is strong; while the motion is subtle, human attention is attracted more by spatial cues like color, contrast and orientation [72].
- In the HVS, temporal interactions exist among stimuli that do not overlap in time, known as temporal masking. Because of the temporal masking effect, the eye fixation

position is dependent not only on the current content but also on what was displayed prior to the change.

In the proposed fusion method, the product of the spatial and temporal maps is added to address the finding that objects salient both spatially and temporally tend to stand out more than objects that are important only spatially or temporally. Since the relative importance of spatial cues and temporal cues varies based on the intensity of motion in the video, the proposed fusion method contains weighted spatial map and temporal map, rather than the average of them as in Itti et al.'s model. The formula of spatio-temporal saliency map is given by:

$$\begin{aligned} S &= a \cdot \mathrm{S}(s) + b \cdot \mathrm{S}(t) + \mathrm{c} \cdot \ \mathrm{S}(s) \cdot \mathrm{S}(t) \\ a &= \max\left\{0, 0.5\left(1 - \frac{ave\ DOH}{\varepsilon}\right)\right\} \\ b &= \min\left\{1, 0.5\left(1 + \frac{ave\ DOH}{\varepsilon}\right)\right\} \\ c &= 1 \end{aligned} \tag{5.29}$$

In the equation above, $S(s)$ and $S(t)$ are the normalized spatial and temporal saliency maps of a frame. a, b, and c are the parameters to control the relative strength of the spatial saliency map, the temporal saliency maps and the product of two maps, respectively. The average Difference of Histograms (*aveDOH*), serves as an indicator for intensity of motion and ε is an empirical threshold for intensity of motion. When the intensity of movement in video is high, i.e., *aveDOH*>ε, the temporal saliency map is dominating and the spatial map has little contribution to the master saliency map.

The DOH between frame n and frame n-1 in a sequence is defined by:

$$DOH = \left(\sum_{1}^{q} |hist(L_n) - hist(L_{n-1})|\right), \tag{5.30}$$

where hist(.) means histogram, L_n and L_{n-1} are the luminance map of frame n and frame n -1 respectively, and q is the number of levels used in histograms (q = 30 in this paper). It was reported that luminance histogram is a very efficient indicator of image content and motion intensity. Since we are using the luminance map, this metric can also detect light intensity change even there is no actual moving objects in the video. Due to temporal masking effects and the fact that eye fixations last more than 300msec on average, a group of frames (group size = 10 frames in this paper) are take into consideration rather than the single current frame. The average DOH in the group of frames divided by the total number of pixels is used as an index for the level motion in the video, and determines the parameter for dynamic fusion of temporal and spatial saliency map. Figure 5.15 shows the average DOH for sequence *bistro 03*. The peaks before frame 40 match the action of glass falling and the low average DOG in the second part of the video indicates there is less motion, both observations correlating well with the actual intensity of motion.

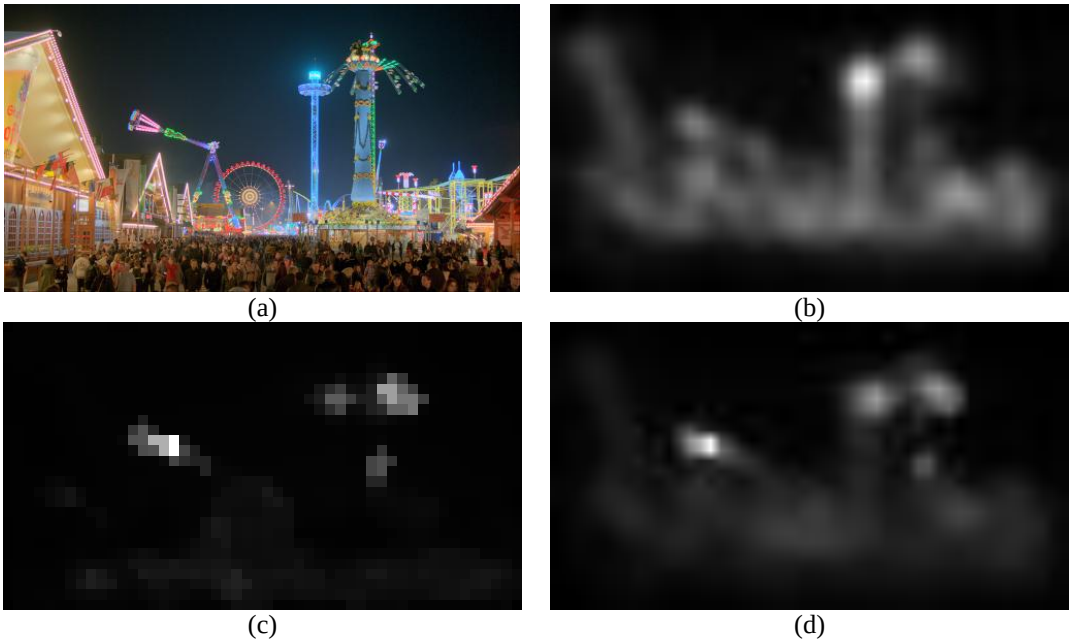


Figure 5.14 Example of spatio-temporal saliency map. (a) frame from sequence *Park*, (b) spatial saliency map, (c) temporal saliency map, (d) fused spatio-temporal saliency map.

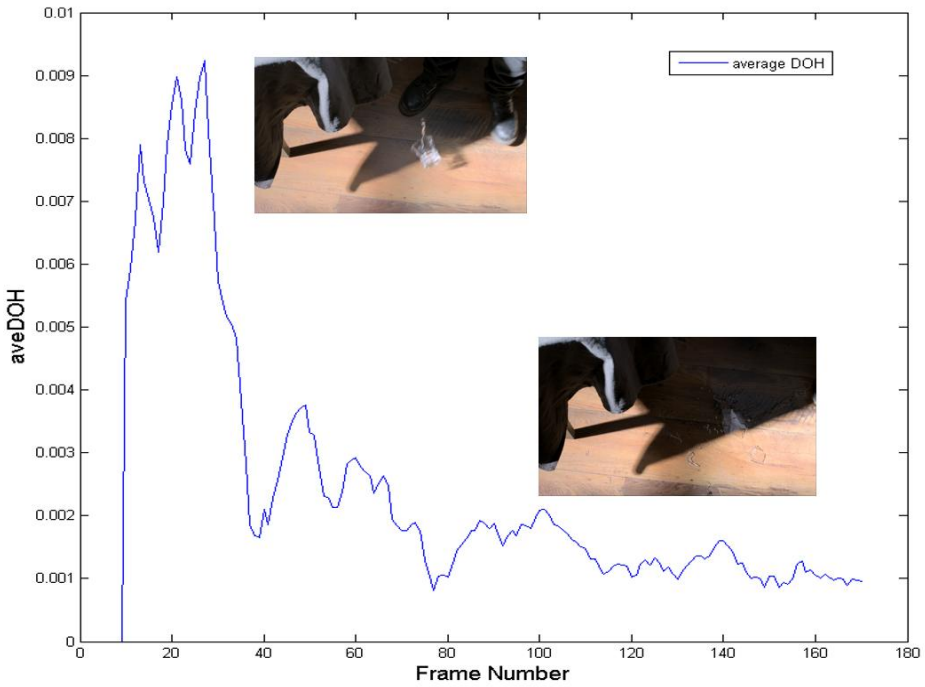
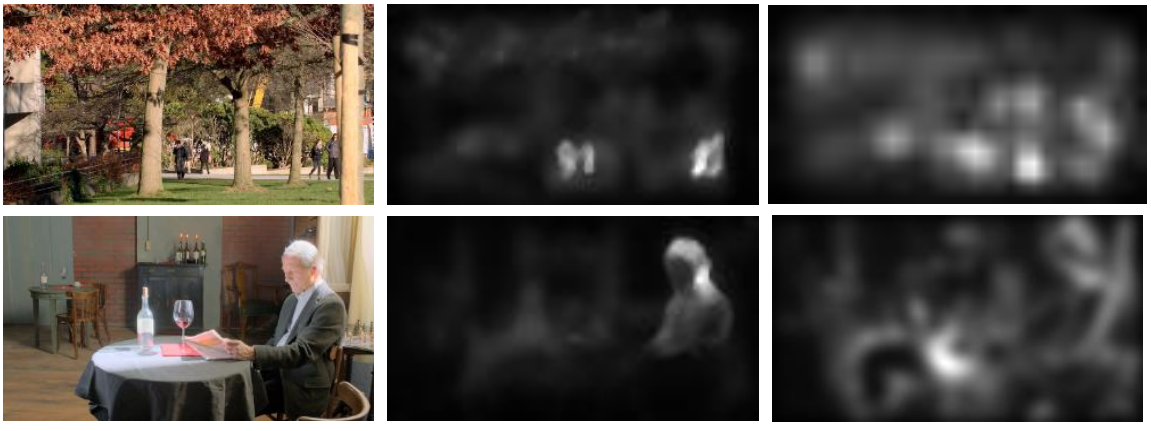


Figure 5.15 Average DOH of sequence *bistro 03*.

5.3 Performance Evaluation

A few saliency maps of videos using the proposed model and Itti et al.'s model are shown in Figure 5.16. In order to quantitatively evaluate our model's performance against the existing models, we chose to use three different statistical metrics. A detail description of the ground truth of visual attention from eye tracking and these metrics is given in the flowing subsections.



(a) (b) (c)

Figure 5.16 Saliency maps using Itti et al.'s model and the proposed model. (a) Example frames of HDR videos; (b) the spatio-temporal saliency map from the proposed model; (c) the spatio-temporal saliency map from Itti et al.'s model.

5.3.1 Human Fixation Density Maps

FDMs were obtained from free view eye tracking experiment we conducted (See Chapter 4). They represent subjects' region of interest (RoI) and serve as ground truth for assessing the performance of the proposed model and benchmark models. For a given image, all fixation data from different subjects are combined together to provide a spatial distribution of human fixation (see Figure 5.17). All fixation hits are filtered by a 2D Gaussian filter whose standard deviation is set to 1° of visual angle. Figure 5.19 depicts a few HDR images used in the experiment and the FDMs obtained from eye tracking study.

For video clips in the experiment, spatial distribution of human fixation for every frame is computed for every subject. If duration of a given fixation is longer than the length of one frame, this fixation hit is appeared in more than one frame. Fixations from all subjects are combined together and filtered by a 2D Gaussian filter. The same Gaussian filter is used for both image stimuli and video stimuli. Figure 5.18 depicts a few frames from one of the video clips, *playground*.

Figure 5.17 Fixations from all subjects of an HDR image. Every color stands for one subject.

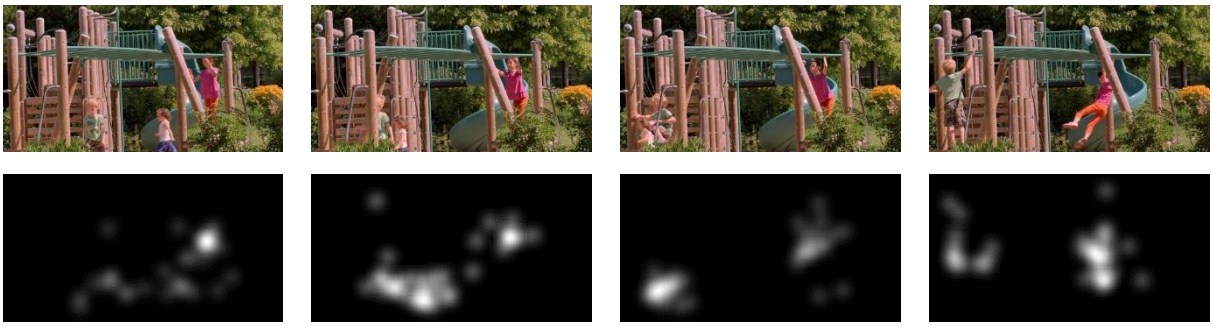
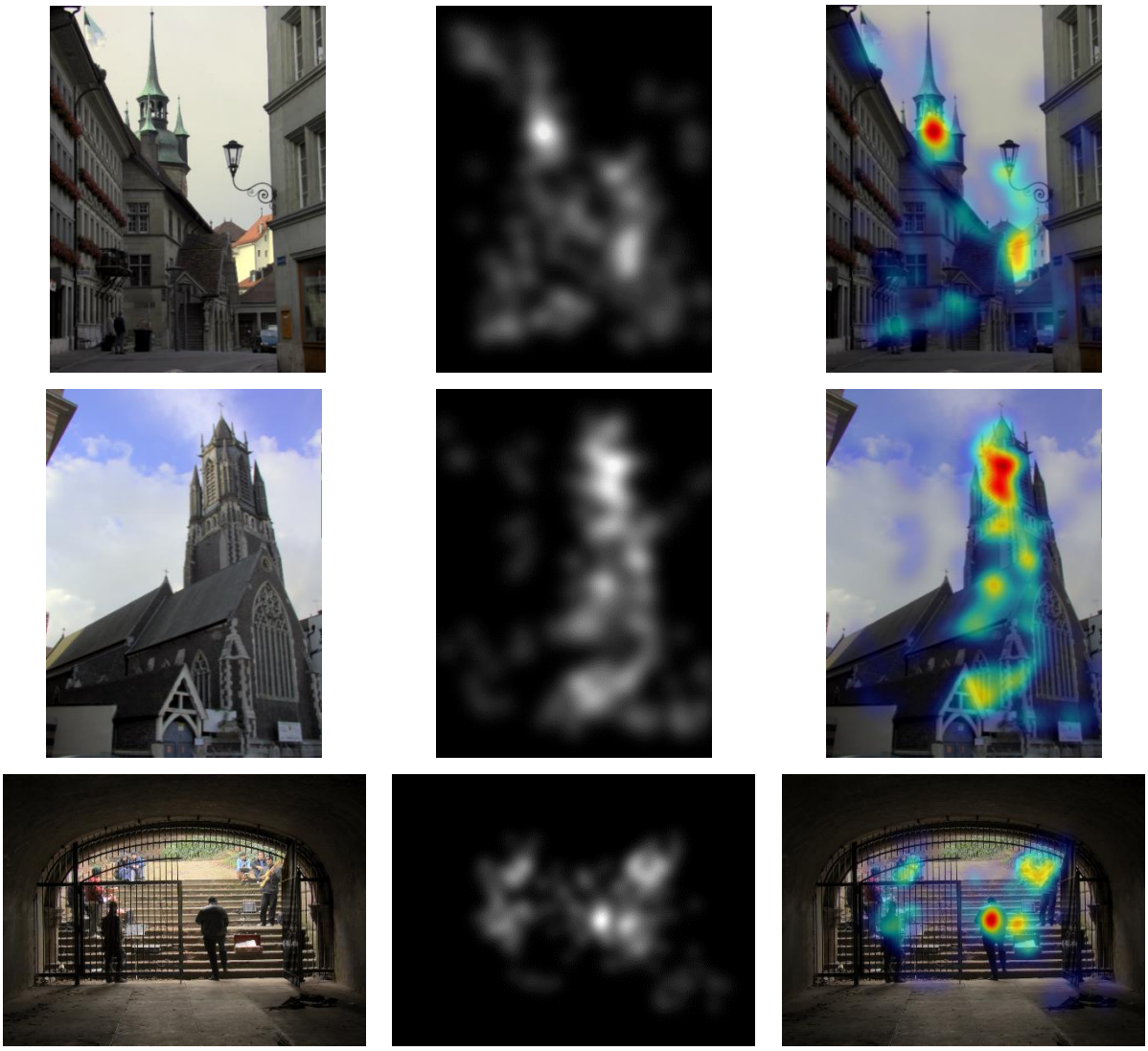


Figure 5.18 Human fixation density maps (FDMs) from sequence *playground*. Top row: a few example frames from sequence Playground; bottom row: fixation density maps from eye tracking study.



(a) (b) (c)

Figure 5.19 Fixation density maps (FDMs) of HDR images. (a) HDR image, (b) fixation density map from eye tracking experiment, (c) fixation density map (heat map) lay on top of original image (red regions represent the most attended area).

5.3.2 Evaluation Metrics

Three statistical metrics have been chosen for comparison of models. The reason to use multiple metrics in the qualitative assessment is to make sure conclusions are robust and independent of metrics. All three metrics are very often used in research community to evaluate

the accuracy of visual attention models. In the following sections, g denotes the FDM derived from eye tracking experiment. The prediction map from computational model is represented by p.

5.3.2.1 Linear Correlation Coefficient (CC):

The first metric, linear correlation coefficient, measures the strength of linear relationship between two data sets:

$$CC\ (p,g) = \frac{cov(p,g)}{\sigma_p \sigma_g}, \tag{5.31}$$

where $cov(p,g)$ is the covariance value between p and g, and σ_p and σ_g are the standard deviations of p and g. The range of CC is between -1 and 1. Values close to 0 mean a poor correlation between the two data sets. When CC is +1 or -1, there is a perfect linear relationship between the two variables.

5.3.2.2 Kullback-Leibler Divergence (KL):

Kullback-Leibler divergence (KL) measures the overall dissimilarity between two probability density functions:

$$KL(p|g) = \sum_x p(x) Log\left(\frac{p(x)}{g(x)}\right). \tag{5.32}$$

$p(x)$ and $g(x)$ are probability density functions deduced from prediction map and ground truth map. This can be done by dividing each location of map by the sum of all pixel values. The KL value varies from zero to infinity. When the KL-divergence value is zero, it means two probability density functions are exactly the same. However, KL-divergence is not a distance metric and not symmetric, which means $KL(p|g) \neq KL(g|p)$. In some published results, $KL(p|g)$ using ground truth as reference is used. In our comparison, average of $KL(p|g)$ and $KL(g|p)$ is reported.

5.3.2.3 Receiver Operating Characteristic Analysis (ROC)

The Receiver Operating Characteristic Analysis (ROC) is probably the most widely used qualitative metric used for assessing of saliency models. In this metric, the input includes a set of fixation points from eye tracking study and the prediction map, which is referred to as "hybrid method" since it mixes two types of information, a continues map and a set of discrete points [32].

The continuous prediction map is thresholded with different values within the data range of prediction map to generate a binary map. Only certain percentage of pixels of the map are kept as white and the rest of pixels are black. For example, top 5%, 10%, or 20% pixels in the map are kept as shown in Figure 5.20. For each threshold, observers' fixations are laid on top of the binary map (see Figure 5.21). Four numbers featuring the quality of classification are computed. They are true positive (TP), true negative (TN), false positive (FP) and false negative (FN) (see Figure 5.22). A TP means a fixation that fall on the white areas and a FN means a fixation that fall on black areas. By varying the threshold, a ROC curve (see Figure 5.23) can be plotted showing True Positive Rate (FPR) as a function of False Positive Rate (FPR). TPR, also called hit rate, is defined as TP/ (TP +FN); while FPR is defined as FP/ (FP +TN).

The area underneath the ROC curve, denoted as AUC, is often used as a numerical score to measure how well the prediction is aligned with ground truth. The area under ROC curve could be obtained by a trapezoid approximation. An area of 1 means the prediction is perfect.

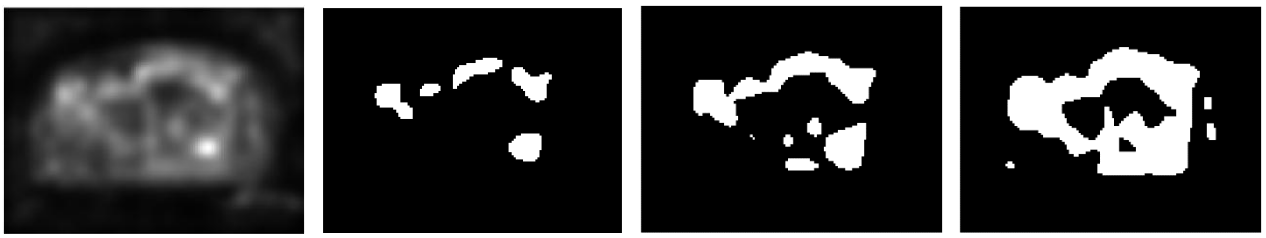


Figure 5.20 Threshold saliency map to binary map. From left to right: top 5%, 10%, and 20% of pixels are kept.

Figure 5.21 Receiver Operating Characteristic (ROC) analysis. Binary image is the thresholded saliency map with 10% of pixels kept. The red dots are the fixation points from one of the subjects in eye tracking experiment.

True Positive	False Positive
False Negative	True Negative

Figure 5.22 Confusion matrix.

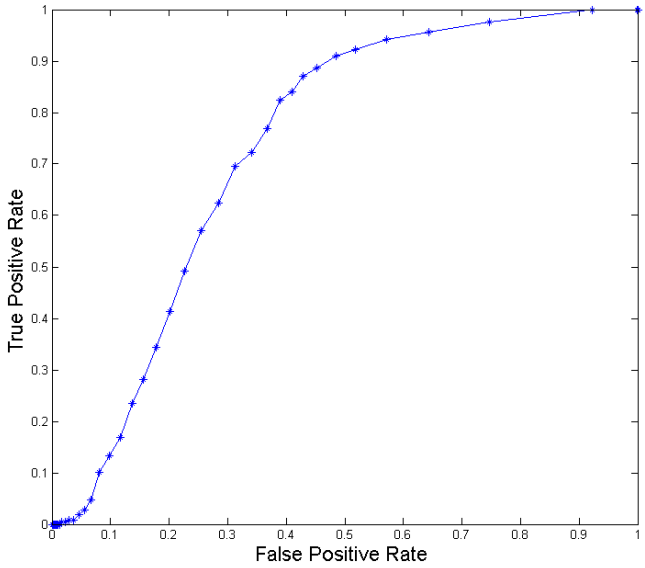


Figure 5.23 ROC curve of one HDR image. Area under curve is 0.734.

Given the inter-subject variability in eye-tracking data, the natural dispersion of fixations among different subjects looking at the same image [73], no saliency algorithm can perform better (on average) than the ROC dictated by inter-subject variability. We calculate an ideal ROC which is a theoretical upper bound to fully evaluate the predictive ability of computational models. The ideal ROC is obtained by a split data technique, measuring how well the fixations of half of the subjects can be predicted by the other half of subjects. All subjects are repeatedly split into two sub-groups with same number of subjects in a random manner. The reported theoretical upper bound value is the values averaged over 100 random samples. Similar techniques are used in previous publications such as [28] and [74].

5.3.3 Quantitative Comparisons and Discussions

As mentioned before, to evaluate the performance of our proposed method, we compared it against two benchmark models, the Itti et al.'s model [53], the best LDR approach, and the Contrast Features (CF) model [55] which is designed for HDR content. Two data sets were used, an HDR image dataset with 23 images, and an HDR video dataset with 10 videos[4]. For fair comparison, we extended the Matlab code of Itti et al.'s model available online [75] to work on HDR images in RGBE format.

5.3.3.1 Experiment 1 – using the HDR image dataset

Table 5.1 provides the average results of CC, ROC, and KL on 23 HDR images. Even though CF method improves the prediction accuracy compared to Itti et al.'s method, the proposed method outperforms both CF and Itti et al.'s methods according to all three metrics. The proposed model performs at approximately 91% of the theoretical limit (Figure 5.24).

[4] These two datasets are the same as in Chapter 4.

Moreover, the lower standard deviations across images achieved by proposed method imply the proposed method is more robust than two benchmark models in predicting attention selection.

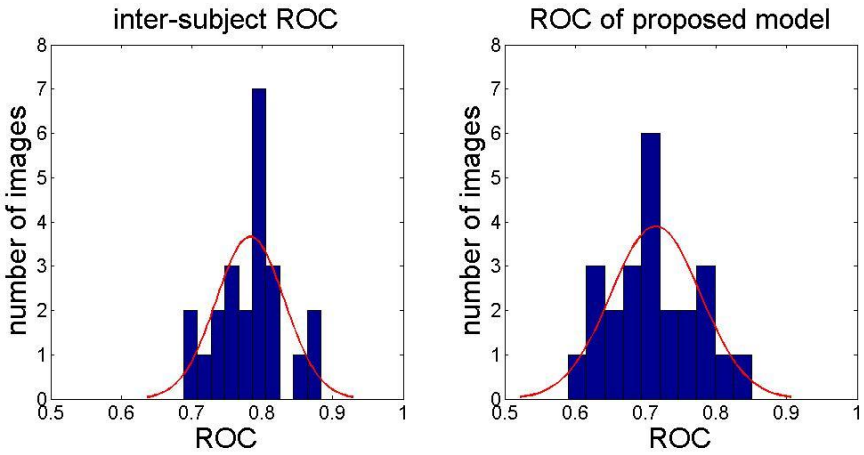


Figure 5.24 Ideal ROC and ROC of proposed model for the HDR image dataset. (a) Distribution of ROC using the split-data technique, providing the theoretical upper bound of ROC. Ideal ROC = 0.783. (b) Distribution of proposed model performance. Mean ROC = 0.711.

Table 5.1 Average results of visual attention models on HDR images.

Model	ROC		CC		KL	
	Ave.	ST dev.	Ave.	ST dev.	Ave.	ST dev.
Proposed	0.71	0.07	0.46	0.12	1.68	0.56
CF	0.68	0.07	0.41	0.16	1.85	0.74
Itti-CIO	0.64	0.08	0.37	0.18	1.86	0.72

5.3.3.2 Experiment 2 – using the HDR video dataset

Table 5.2 shows the average CC, ROC and KL per clip. Since CF method did not propose a solution for video input, it is excluded from the comparison using the video dataset. The average gain of proposed spatio-temporal method is about 0.05, 0.12 and 0.53 for ROC, CC and KL respectively. Based on the analysis of ideal ROC, the predictive efficiency of proposed model is 95% of the ideal ROC (Figure 5.25). Standard deviations using proposed model are reduced according to all three metrics, suggesting a more robust model than Itti et al.'s.

A few more observations can be made from Table 5.2: (1) The lowest scores of both methods belong to sequence carousel, a very busy and cluttered scene with multiple moving objects and blinking bulbs. The possible reason might be when the visual stimuli are complicated and busy, there are more top-down factors driving the attention and those factors are not yet taken into consideration in both models. (2) The largest gain of performance are given by bistro 01 and carousel according to ROC, and fishing and bistro 03 according to CC and KL. What these sequences have in common is very dynamic lighting and salient objects are not always in brightest regions. This suggests that the improvement is yielded from the HVS model which accounts visual sensitivity under broad luminance range. (3) It is worth noting that there are some discrepancies among three metrics, for example, the highest score of all sequences using Itti et al.'s model is given by balloon according to CC and ROC, but KL shows that Itti et al.'s model performs best with park. This suggests that these metrics are not always unanimous in deciding the degree of similarity between predictions and eye tracking data. Therefore, performance should be assessed by a combination of metrics to ensure a fair comparison.

Table 5.3 shows the performance of spatial saliency map and temporal saliency map for both models. The best performance stems from incorporating both spatial and temporal information for both models. Noting also that the proposed spatial and temporal saliency maps outperform those of Itti et al. respectively according to all three metrics.

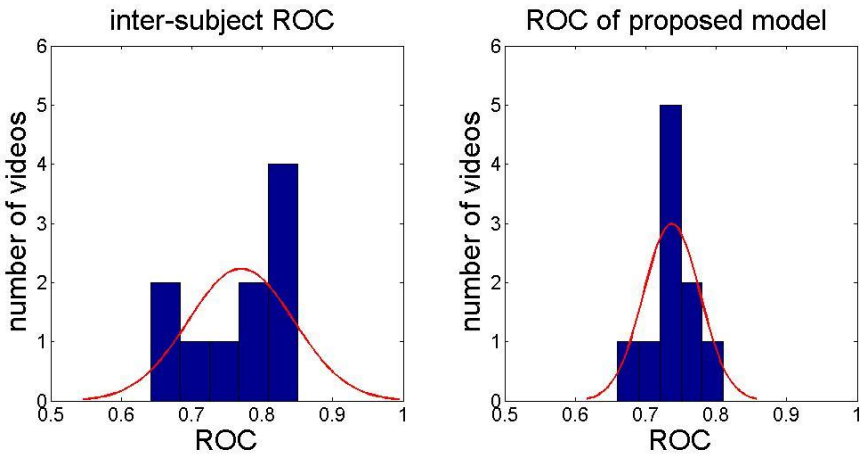


Figure 5.25 Ideal ROC and ROC of proposed model for the HDR video dataset. (a) Distribution of ROC using the split-data technique, providing the theoretical upper bound of ROC. Ideal ROC = 0.770. (b) Distribution of proposed model performance. Mean ROC = 0.737.

Table 5.2 Results of visual attention models on each HDR video.

	ROC			CC			KL		
	Itti	Proposed	Gain (Proposed - Itti)	Itti	Proposed	Gain (Proposed - Itti)	Itti	Proposed	Gain (Proposed - Itti)
bistro01	0.59	0.7	**0.11**	0.08	0.19	0.11	5.32	4.73	-0.59
fishing	0.67	0.73	0.06	0.17	0.37	**0.2**	5.24	4.24	**-0.99**
park	0.74	0.76	0.02	0.35	0.39	0.04	4.42	3.80	-0.62
mainmall	0.73	0.74	0.01	0.27	0.42	0.15	4.97	4.79	-0.18
market	0.65	0.73	0.08	0.17	0.3	0.13	5.08	4.69	-0.39
bistro03	0.66	0.73	0.07	0.14	0.36	**0.22**	5.53	4.74	**-0.79**
balloon	0.80	0.81	0.01	0.43	0.49	0.06	4.65	4.39	-0.27
carousel	0.55	0.66	**0.11**	0.05	0.19	0.14	5.64	5.11	-0.54
playground	0.73	0.74	0.01	0.28	0.43	0.15	5.26	4.70	-0.56
bistro02	0.77	0.77	0.00	0.35	0.35	0	4.93	4.50	-0.44
Average	0.69	0.74	0.05	0.23	0.34	0.11	5.10	4.57	-0.53
ST dev.	0.08	0.05	-0.04	0.13	0.10	-0.03	0.38	0.36	-0.02

The top two improved sequences according to each metric are shown in bold.

Table 5.3 Average results of visual attention models on HDR videos.

Model	ROC		CC		KL	
	Ave.	ST dev.	Ave.	ST dev.	Ave.	ST dev.
Proposed ST	0.74	0.05	0.34	0.10	4.57	0.36
Proposed S	0.72	0.05	0.27	0.11	4.82	0.45
Proposed T	0.68	0.08	0.29	0.12	4.69	0.51
Itti-ST	0.69	0.08	0.23	0.13	5.10	0.38
Itti-S	0.68	0.08	0.21	0.12	5.08	0.45
Itti-T	0.66	0.07	0.20	0.13	5.29	0.20

ST stands for spatio-temporal, S stands for spatial and T stands for temporal.

5.3.3.3 Experiment 3 – learning optimal weights

In section 5.2.3, dynamic fusion, based on the finding that objects defined by two features are more likely to be noticed than single-feature objects [71], we propose to add the product term of the spatial and temporal saliency maps in the spatio-temporal map to represent and enhance those areas that are salient both spatially and temporally. Also, we propose to determine the relevant importance of the spatial and temporal saliency map to address the observation that the relative importance of spatial and temporal cues depends on the intensity of motion in the video [72]. These proposals are made based on psychology studies and observations. In this implementation, we take a mathematical perspective to quantify the relevant contribution of different saliency maps in predicting visual attention. We use the linear least square regression with constraints to determine the optimal weights of each term in equation (5.29) using eye movement data for the HDR video dataset. Let $S(t)$, $S(s)$ and $S(t) \cdot S(s)$ be the temporal saliency map, the spatial saliency map and the product term of these two maps, respectively. Let us denote $P = [S(t), S(s), S(t) \cdot S(s)]$, F as the fixation density map generated from eye tracking study, and $W = [Wt, Ws, Wts]$ as the weights of each channel in P. Then, the best solution of W is:

$$\arg\min_{W} ||P \times W - F||^2 \tag{5.33}$$

subject to

$$0 \leq \mathrm{W} \leq 1\,; \qquad W_t + W_s + W_{ts} = 1\,. \tag{5.34}$$

We learn optimal weights for each frame in the video dataset and average weights are reported in Table 5.4.

Table 5.4 Optimal weights for HDR videos.

	Temporal	**Spatial**	**Product**
bistro01	0.32	0.13	0.49
fishing	0.20	0.06	0.72
park	0.34	0.31	0.33
mainmall	0.41	0.02	0.53
market	0.08	0.03	0.87
bistro03	0.50	0.01	0.44
balloon	0.56	0.03	0.37
carousel	0.29	0.01	0.68
playground	0.09	0.04	0.84
bistro02	0.15	0.12	0.71

As shown in Table 5.4, for 7 out of 10 videos, the product term has the highest weight, which confirms that objects that are important both spatially and temporally stand out the most, and the fusion step should include the product term of the spatial and temporal saliency maps. However, in most of the models proposed, including Itti et al.'s model, the master saliency map only takes into consideration the average of saliency maps from each channel or the weighted sum [72], ignoring the product term. Besides, for all 10 videos, the optimal weights of the temporal saliency map are larger than that of spatial saliency map. This shows that temporal features contribute more to attention than spatial features in the context of videos.

5.4 Conclusion

In this chapter, a spatio-temporal model simulating the bottom-up visual attention has been proposed for HDR images and videos. Considering the HVS's properties related to HDR's characteristics of wide luminance ranges and rich color gamut, the proposed model addresses limitations of state-of-the-art models. The spatio-temporal saliency map is obtained combining the low level visual features including color, orientation, intensity and motion. A dynamic fusion method is used to control the relative weights of spatial map and temporal map based on the intensity of motion in video. Quantitative evaluation has been conducted with data collected from eye tracking study. The proposed method shows better performance on both image and video datasets.

In terms of application, the proposed method could be utilized in a tone-mapping algorithm which locally adjusts the contrast of HDR image and video according to area of interest provided by saliency map [76]. With the information of areas of interests, compression methods for HDR could be more effective and efficient by allocating the bit rate resources more to important areas and less to the rest of frame.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

The emerging HDR technology is transforming the digital media industry with its superior picture quality and life-like visual experience. It allows representation of scenes with values corresponding to real-world light levels. As all other new technologies, HDR provides opportunities as well as challenges in all stages of digital imaging pipeline. With the goal to optimize the HDR imaging pipeline and contribute to enabling this exciting technology, this thesis presents a few attempts to assess and improve the quality of HDR using subjective and objective approaches.

Chapter 3 presents two in depth studies focusing on HDR compression platform and HDR video quality metrics. The findings confirm that HEVC could be used as a platform for a codec to compress HDR videos. The studies also provide insights in available quality metrics for HDR videos by comparing subjective opinions and objective results. We found that HDR-VDP-2 and VIF outperform all other tested metrics, but in order to apply VIF (originally developed for LDR) to HDR videos, an extending method like PU encoding or multi-exposure is necessary.

Chapter 4 presents a HDR video eye tracking dataset collected from naïve participants viewing HDR videos in a few viewing task. This dataset is the only publicly available HDR eye tracking dataset at the time of this writing. The comparative study shows a clear subjective reference for HDR display when individuals are given a choice between HDR and LDR displays.

Chapter 5 presents a new computational approach to predict visual attention for HDR content. The method leverages state-of-the-art bottom-up framework as well as traits of the HVS to provide visual attention prediction for both HDR images and videos.

Besides, the HDR video datasets, *DML-HDR* [37], and the HDR eye tracking dataset, *DML-iTrack-HDR* [47], generated from this thesis could be used in many HDR research areas such as compression, quality assessment, visual attention, and video resizing.

6.2 Future Work

The visual attention model developed in this thesis could be used to design a quality metric for HDR content which incorporates the visual attention information. Utilizing the saliency map provided by the visual attention model, it is able to address the fact that visual distortions appearing in less salient areas are less visible and less annoying when designing the quality metric. Currently in the HEVC standard, the quality metric used is the full reference PSNR which does not take into account the visual attention information. A full reference quality metric using the aforementioned design could be implemented in the compression standard and improve its compression efficiency. A non-reference quality metric could also be designed using a similar approach. In the non-reference quality metric, the saliency map from the visual attention model could be used to derive a weighting function or modulate the contribution of each pixel to the final quality score. Such a non-reference quality metric is valuable in many applications where a reference image does not exist, such as HDR capturing (i.e., in cameras) or set-top boxes.

Bibliography

[1] E. Reinhard, G. Ward, S. Pattanaik, P. Debevec, W. Heidrich and K. Myszkowski, High dynamic range imaging: acquisition, display, and image-based lighting, Morgan Kaufmann, 2010.

[2] H. Seetzen, W. Heidrich, W. Stuerzlinger, G. Ward, L. Whitehead, M. Trentacoste, A. Ghosh and A. Vorozcovs, "High dynamic range display systems," *ACM Transactions on Graphics (TOG),* vol. 23, no. 3, pp. 760-768, 2004.

[3] J. A. Ferwerda, "Elements of early vision for computer graphics," *IEEE Computer Graphics and Applications,* vol. 21, no. 5, pp. 22-33, 2001.

[4] P. E. Debevec and J. Malik, "Recovering High Dynamic Range Radiance Maps from Photographs," *Proc. the 24th annual conference on Computer graphics and interactive techniques. (Proc. of SIGGRAPH 97),* pp. 369-378, 1997.

[5] RED, "High dynamic range video with HDRX," RED Inc., 09 May 2013. [Online]. Available: http://www.red.com/learn/red-101/hdrx-high-dynamic-range-video. [Accessed Sep 2014].

[6] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling and H. Brendel, "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays," *IS&T/SPIE Electronic Imaging, San Francisco,* 2014.

[7] J. Fröhlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling and H. Brendel, "HdM-HDR-2014 Project," 5 Feb 2014. [Online]. Available: https://hdr-2014.hdm-stuttgart.de/. [Accessed Feb 2014].

[8] Y. Olivier, D. Touzé, C. Serre, S. Lasserre, F. L. Léannec and E. François, "Description of HDR sequences," *ISO/IEC JTC1/SC29/WG11 MPEG2014/m31957,* pp. San Jose, USA, Oct 2014.

[9] SIM2 Multimedia, "HDR real time LCD - LED display," [Online]. Available: http://www.sim2.com/HDR/hdrdisplay/hdr47e_s_4k. [Accessed Oct 2014].

[10] BrightSide, "BrightSide DR37-P HDR display," [Online]. Available: http://www.bit-tech.net/hardware/2005/10/03/brightside_hdr_edr/1. [Accessed Oct 2014].

[11] R. Mantiuk, A. Efremov, K. Myszkowski and H.-P. Seidel, "Backward compatible high dynamic range MPEG video compression," *ACM Transactions on Graphics,* vol. 25, no. 3, pp. 713-723, 2006.

[12] "H.265 : High efficiency video coding," ITU. 2013-06-07.

[13] J. Boyce, J. Chen, Y. Chen, D. Flynn, M. M. Hannuksela, M. Naccari, C. Rosewarne, K. Sharman, J. Sole, G. J. Sullivan, T. Suzuki, G. Tech, Y.-K. Wang, K. Wegner and Y. Ye, "Edition 2 Draft Text of High Efficiency Video Coding (HEVC), Including Format Range

(RExt), Scalability (SHVC), and Multi-View (MV-HEVC) Extensions," JCT-VC, 2014.

[14] W. James, The Principles of Psychology, New York: Holt, 1891.

[15] D. Kahneman, Attention and effort, New Jersey: Prentice-Hall, 1973.

[16] O. Le Meur, P. L. Callet, D. Barba and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *Pattern Analysis and Machine Intelligence, IEEE Transactions,* vol. 28, no. 5, pp. 802-817, 2006.

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

[18] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott and A. M. Wichansky, "Eye tracking in web search tasks: design implications," *Proceedings of the 2002 symposium on Eye tracking research & applications, ACM,* pp. 51-58, 2002.

[19] T. van Gog and K. Scheiter, "Eye tracking as a tool to study and enhance multimedia learning," *Learning and Instruction,* vol. 20, no. 2, pp. 95-99, 2010.

[20] H. Yu, J. Li, Y. Tian and T. Huang, "Automatic interesting object extraction from images using complementary saliency maps," *Proceedings of the international conference on Multimedia, ACM,* pp. 891-894, 2010.

[21] G. Zhang, Z. Yuan, N. Zheng, X. Sheng and T. Liu, "Visual saliency based object tracking," *Computer Vision–ACCV,* pp. 193-203, 2009.

[22] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *IEEE Transactions on Robotics,* vol. 25, no. 4, pp. 861-873, 2009.

[23] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing,* vol. 13, no. 10, pp. 1304-1318, 2004.

[24] Y.-F. Ma, X.-S. Hua, L. Lu and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia,* vol. 7, no. 5, pp. 907-919, 2005.

[25] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Transactions on graphics (TOG),* vol. 26, no. 3, pp. 1-10, 2007.

[26] J. Wang, D. M. Chandler and P. L. Callet, "Quantifying the relationship between visual salience and visual importance," in *Proc. SPIE 7527, Human Vision and Electronic Imaging XV, 75270K* , 2010.

[27] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience,* vol. 2, no. 3, pp. 194-203, 2001.

[28] B. J. Stankiewicz, N. J. Anderson and R. J. Moore, "Using performance efficiency for testing and optimization of visual attention models," in *Proc. SPIE 7867, Image Quality and System Performance VIII, 78670Y* , 2011.

[29] U. Engelke, H. Liu, J. Wang, P. L. Callet, I. Heynderickx, H.-J. Zepernick and A. Maeder, "Comparative study of fixation density maps," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1121-1133, 2013.

[30] R. J. Jacob and K. S. Karn, "Eye tracking in human–computer interaction and usability research: ready to deliver the promises (section commentary)," in *The mind's eye: cognitive and applied aspects of eye movement research*, vol. 2, Amsterdam, Elsevier Science, 2003, p. 573–605.

[31] D. Hansen and P. Majaranta, "Basics of camera-based gaze tracking," in *Gaze interaction and applications of eye tracking: advances in assistive technologies*, Medical Information Science Reference, Hershey, 2012, p. 21–26.

[32] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior research methods*, vol. 45, no. 1, pp. 251-266, 2013.

[33] Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, "New Test Sequences in the VIPER 10-bit HD Data," JVT-Q090, October, 2005.

[34] Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, "Donation of Tone Mapped Image Sequences," JVT-Y072, October, 2007.

[35] K. McCann, B. Bross, W.-J. Han, S. Sekiguchi and G. J. Sullivan, "High Efficiency Video Coding (HEVC) text specification draft 5," JCTVC-G1102, November, 2011.

[36] ISO/IEC JTC1/SC29/WG11, "Joint Call for Proposals on Video Compression Technology," JCTVC-N1113, January, 2010.

[37] Digital Multimedia Lab, "DML-HDR dataset," [Online]. Available: http://dml.ece.ubc.ca/data/DML-HDR. [Accessed Sep 2014].

[38] G. Ward, "Real Pixels," *Graphics Gems II*, p. 80–83, 1992.

[39] E. Reinhard, M. Stark, P. Shirley and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 267-276, 2002.

[40] ITU, "Recoomendation ITU-R BT.500-13: Methodology for the subjective assessment of the quality of television pictures," 2012.

[41] T. O. Aydın, R. Mantiuk and H.-P. Seidel, "Extending quality metrics to full luminance range images," in *Proc. SPIE 6806, Human Vision and Electronic Imaging XIII, 68060B*, 2008.

[42] J. Munkberg, P. Clarberg, J. Hasselgren and T. Akenine-Möller, "High dynamic range texture compression for graphics hardware," *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 698-706, 2006.

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

[44] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing,* vol. 15, no. 2, pp. 430-444, 2006.

[45] R. Mantiuk, K. J. Kim, A. G. Rempel and W. Heidrich, "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Transactions on Graphics (TOG),* vol. 30, no. 4, 2011.

[46] P. Lopez, E. Fran çois, P. Ying, T. Lu, K. M. A. Luthra and J. C. S. Lee, "Generating of anchors for the explorations for HDR/WCG Contetn Distribution and Storage," ISO/IEC JTC1/SC29/WG11 MPEG2014/M34077, Saporo, Japan, July 2014.

[47] Digital Multimedia Lab, "DML-iTrack-HDR," [Online]. Available: http://dml.ece.ubc.ca/DML-iTrack-HDR.

[48] R. Boitard, K. Bouatouch, R. Cozot, D. Thoreau and A. Gruson, "Temporal coherency for video tone mapping," in *SPIE Conference on Applications of Digital Image Processing XXXV,* 2012.

[49] SensoMotoric Instruments, "SMI products," [Online]. Available: http://www.smivision.com/en/gaze-and-eye-tracking-systems/products/red-red250-red-500.html. [Accessed March 2014].

[50] SensoMotoric Instrucments (SMI), "Experiment center 2 manual," Version 2.4, 2010.

[51] U. Engelke, H. Liu, J. Wang, P. L. Callet, I. Heynderickx, H.-J. Zepernick and A. Maeder, "A comparative study of fixation density maps," *IEEE Transactions on Image Processing,* vol. 22, no. 3, pp. 1121-1133, 2013.

[52] 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, 1998.

[53] L. Itti, N. Dhavale and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *Optical Science and Technology, SPIE's 48th Annual Meeting. International Society for Optics and Photonics,* pp. 64-78, 2004.

[54] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 1, pp. 185-207, 2013.

[55] R. Br émond, J. Petit and J.-P. Tarel, "Saliency maps of high dynamic range images," *Trends and Topics in Computer Vision,* pp. 118-130, 2012.

[56] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Matters of Intelligence,* pp. 115-141, 1987.

[57] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology,* vol. 195, no. 1, pp. 215-243, 1968.

[58] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Visual Cognition,* vol. 12, no. 6, pp. 1093-1123, 2005.

[59] Y.Zhai and M. Shah, "Visual attention detection in video sequences using saptiotemporal cues," *Proceedings of the 14th annual ACM international conference on Multimedia,* pp. 815 - 824, 2006.

[60] R. W. G. Hunt, The Reproduction of Colour, 6th ed, John Wiley, 2004.

[61] M. H. Kim, T. Weyrich and J. Kautz, "Modeling human color perception under extended luminance levels," *ACM Transactions on Graphics (TOG),* vol. 28, no. 3, 2009.

[62] M. R. Luo, A. A. Clarke, P. A. Rhodes, A. Schappo, S. A. Scrivener and C. J. Tait, "Quantifying colour appearance. Part I. LUTCHI colour appearance data," *Color Research & Application,* vol. 16, no. 3, pp. 166-180, 1991.

[63] O. E. Uscanga, "On the fundamental data-base of normal and dichromatic color vision," Ph.D thesis, Univerisyt of Amsterdam, 1979.

[64] R. Mantiuk, K. Myszkowski and H.-P. Seidel, "Lossy compression of high dynamic range images and video," *Proceedings of SPIE,* vol. 6057, pp. 311-320, 2006.

[65] S. J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology. International Society for Optics and Photonics,* pp. 2-15, 1992.

[66] T. O. Aydin, R. Mantiuk, K. Myszkowski and H.-P. Seidel, "Dynamic range independent image quality assessment," *ACM Transactions on Graphics (TOG),* vol. 27, no. 3, p. 69, 2008.

[67] R. Mantiuk, S. J. Daly, K. Myszkowski and H.-P. Seidel, "Predicting visible differences in high dynamic range images: model and its calibration," *Electronic Imaging 2005. International Society for Optics and Photonics,* pp. 204-214, 2005.

[68] B. Horn and B. G. Schunck, "Determining optical flow," *Artificial Inteliigence,* vol. 17, pp. 185-203, 1981.

[69] T. Vaudrey, A. Wedel, C.-Y. Chen and R. Klette, "Improving Optical Flow Using Residual and Sobel Edge Images," *In Arts and Technology: First International Conference, ArtsIT 2009,* vol. 30, 2010.

[70] C. a. R. M. Tomasi, "Bilateral filtering for gray and color images," *Sixth International Conference on Computer Vision,* pp. 839-846, 1998.

[71] H.-C. Nothdurft, "Salience from feature contrast: additivity across dimensions," *Vision research,* vol. 40, no. 10, pp. 1183-1201, 2000.

[72] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," *Proceedings of the 14th annual ACM international conference on Multimedia, ACM,* pp. 815-824, 2006.

[73] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior research methods,* vol. 45, no. 1, pp. 251-266, 2013.

[74] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of vision*, vol. 11, no. 3, 2011.

[75] J. Harel, "A Saliency Implementation in MATLAB," [Online]. Available: http://www.vision.caltech.edu/~harel/share/gbvs.php. [Accessed Januarary 2014].

[76] W.-C. Lin and Z.-C. Yan, "Attention-based high dynamic range imaging," *The Visual Computer*, vol. 27, no. 6-8, pp. 717-727, 2011.

[77] A. Luthra, E. Fran ̧ois, H. Thoma and S. Hattori, "Draft requirements for High Dynamic Range (HDR) and Wide Colour Gamut (WCG) video coding for Broadcasting, OTT, and Storage Media," ISO/IEC JTC1/SC29/WG11 MPEG2014/N14278, San Jose, USA, Jan 2014.