VIDEO PREPROCESSING BASED ON HUMAN PERCEPTION FOR TELESURGERY

by

Jian Xu

B.S. in C.S., Shanghai Jiao Tong University, China, 1998

M.S. in E.E., Shanghai Jiao Tong University, China, 2001

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Jian Xu

It was defended on

August 18, 2009

and approved by

Ching-Chung Li, Ph.D., Professor, Department of Electrical and Computer Engineering

Zhi-Hong Mao, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Allen C. Cheng, Ph.D., Assistant Professor, Department Electrical and Computer Engineering

Robert J. Sclabassi, Ph.D., M.D.

Dissertation Co-Director: Luis F. Chaparro, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

Mingui Sun, Ph.D., Professor, Department of Neurological Surgery, Bioengineering and Electrical and Computer Engineering

VIDEO PREPROCESSING BASED ON HUMAN PERCEPTION FOR TELESURGERY

Jian Xu, PhD

University of Pittsburgh, 2009

Video transmission plays a critical role in robotic telesurgery because of the high bandwidth and high quality requirement. The goal of this dissertation is to find a preprocessing method based on human visual perception for telesurgical video, so that when preprocessed image sequences are passed to the video encoder, the bandwidth can be reallocated from non-essential surrounding regions to the region of interest, ensuring excellent image quality of critical regions (e.g. surgical region). It can also be considered as a quality control scheme that will gracefully degrade the video quality in the presence of network congestion.

The proposed preprocessing method can be separated into two major parts. First, we propose a time-varying attention map whose value is highest at the gazing point and falls off progressively towards the periphery. Second, we propose adaptive spatial filtering and the parameters of which are adjusted according to the attention map. By adding visual adaptation to the spatial filtering, telesurgical video data can be compressed efficiently because of the high degree of visual redundancy removal by our algorithm. Our experimental results have shown that with the proposed preprocessing method, over half of the bandwidth can be reduced while there is no significant visual effect for the observer. We have also developed an optimal parameter selecting algorithm, so that when the network bandwidth is limited, the overall visual distortion after preprocessing is minimized.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisors Dr Luis F. Chaparro and Dr Mingui Sun for their invaluable support, encouragement, supervision and useful suggestions throughout my graduate study.

I am also highly thankful to other members on my committee, Dr Robert J. Sclabassi, Dr Ching-Chung Li, Dr Zhi-Hong Mao and Dr Allen C. Cheng, for their valuable suggestions throughout this research work. I am grateful for the selfless helps from my fellow colleagues, Dr Wenyan Jia, Dr Qiang Liu and Ning Yao.

Most importantly, none of this would have been possible without the love and patience of my family, to whom this dissertation is dedicated to. I thank my dearest husband Zhongshan Zhang for his love and support, for being there for me and encouraging me to overcome all the difficulties when writing this dissertation. I thank my parents in China for their constant love, support and patience during all these years. I am very proud to be their daughter. Last but not least, I thank my baby girl Jasmine for the happiness she has brought to my life.

1.0 INTRODUCTION

In this dissertation, we present a perceptually adaptive preprocessing method for telesurgical video which requires not only high fidelity but also low delay and high scalability in video streaming. We aim at reducing network delay and jitter on robotic telesurgical video transmission by adaptively prioritizing and pre-processing video contents according to the surgeon's focus of attention and the degree of network congestions. Integrated with the existing or future video codecs and QoS-enabled networking products, we hope to further enhance telesurgical performance.

1.1 VIDEO TRANSMISSION FOR TELESURGERY

Robotic telesurgery is performed by a surgeon at a site remote from the patient. Developments in the fields of robotics, medicine and advances in telecommunications have made robotic telesurgery a powerful treatment option in cases where surgeons are not locally available [1]. The first demonstration of trans-Atlantic telesurgery was reported in 2001 when surgeons in New York operated on a 68-year-old woman in Strasbourg, France using remote-controlled robots to resect her gall bladder by laparoscopy [2]. As illustrated in Figure 1.1, surgical tasks are directly performed by a robotic system controlled by the surgeon at the remote site. Teleoperation control

signals are sent to the surgical site, and force feedback, real-time audio and video signals are transmitted back to the surgeon site.

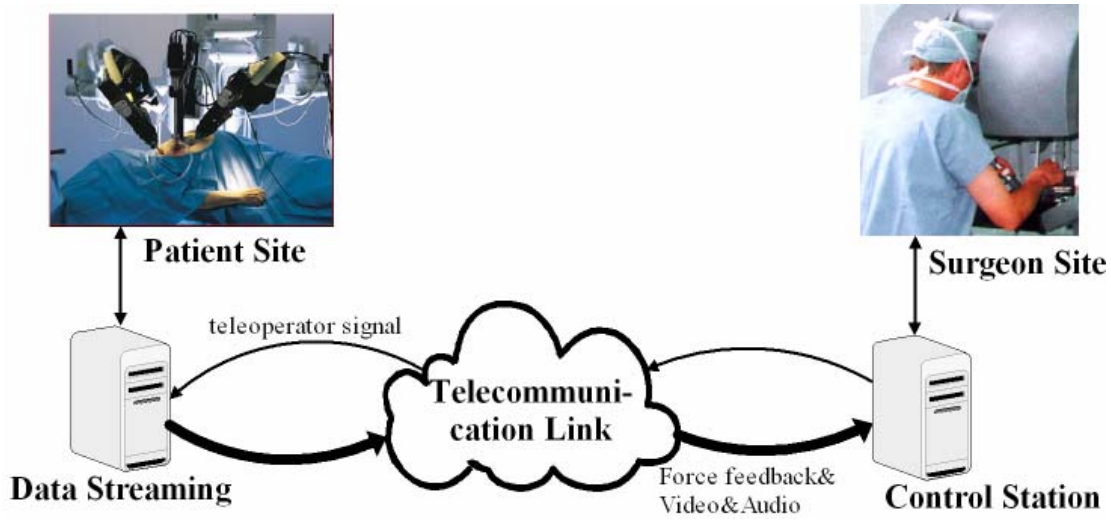


Figure 1.1 Illustration of Telesurgery

Currently, remote surgery can only be carried out between large hospitals and institutes where reliable communication links exist. Surgical data transmission via Internet, although extremely useful in many cases, is still problematic [3-6]. First, the bandwidth of current broadband Internet connections is often not enough for high-quality surgical data streaming. In the information pathway of telesurgery, video is the most essential part and requires by far the most bandwidth. Studies have shown that using a low-latency MPEG-2 encoder, transmitting a video with acceptable quality for surgery requires a bandwidth of at least 10Mbps [7]. Second, the IP-based Internet connection is a best-effort network, i.e., network conditions are dynamic and there is no quality of service (QoS) guarantee. During peak Internet use, network congestion usually causes delay and loss of information, which degrades the video playback quality. To solve this problem, one possible solution is to reengineer the network to provide necessary QoS support via resource reservation or service differentiation. However, the current state of QoS

deployment is only in an experimental stage, not yet ready for the mainstream application. In order to comply with the current architecture and capability of the Internet, telesurgery video transmission has to adapt to the dynamic network condition but still offer reasonable playback quality to the receivers. The conventional video CODECs are not yet satisfactory to this extent [7]. Apart from the unpredicable nature of the network transmission, an important reason is that the existing CODECs are designed to transmit general video content. With no knowledge of the observer's interest (e.g. focus region in the surgical landscape) and network traffic, video frames are encoded with uniform quality which wastes the valuable bandwidth.

1.2 PRIORITY VIDEO CODING

By understanding how surgeons visually perceive the surgical field, one can present them high quality pre-processed images only of the area that they are most interested in. Peripheral non-essential areas of the surgical field are encoded at a lower quality. As a result, the surgeon can receive the visual information necessary to perform a surgery, while the video data can be encoded and transmitted at lower bitrate. Priority image/video coding has been accepted for certain applications such as medical image compression and video conferencing [8-11].

1.2.1 ROI Video Coding

In many video applications, there exist one or more regions in a video frame that have greater importance than the rest of the frame. For instance, in a videoconferencing session or news cast, the face of the talking person is what people usually pay most attention to, rather than the

background. For such applications, the region of interest (ROI) coding technique aims at providing better fidelity in the ROI as opposed to the rest of the frame. ROI coding has been incorporated and supported by some of the latest image/video compression standards (JPEG 2000, H.263, MPEG-4) [8-12]. For example, MPEG-4 standard defines a video frame in terms of components. Each component, called a video object plane (VOP), consists of a snapshot of a video object. Each VOP can be treated separately in a coding/decoding session with a number of flexible choices, such as temporal, spatial, and quality scalabilities.

1.2.2 Foveated Video Coding

If we define the area(s) around the point(s) of fixation as the region of interest, then foveation-based image processing can be viewed as a special case of ROI image processing. The major difference with respect to traditional ROI processing is that the "interest" is continuously varying in the spatial domain and conforms to the human visual system (HVS) characteristics. This technique should provide potentially the greatest bandwidth saving. As we will discuss in Chapter 2.0 , the photoreceptors on the human retina are very nonuniformly distributed, by which only a small region around the center of gaze is captured at high resolution, with logarithmic resolution falloff with eccentricity.

1.2.3 Priority Regions Selection

Depending on the purpose of the video and the number of observers, the selection methods of priority regions vary. There are two main approaches: priority encoding for general-purpose video compression [13-16] and real-time interactive gaze-contingent video transmission [17-19].

General-purpose video compression assumes that a single compressed video stream will be viewed by many observers and at variable viewing distances without any eye tracking or user interaction. Using computer vision techniques, the perceptually important regions can be determined automatically based on some known properties of the human visual system (HVS), such as object size, contrast, shape, color, motion and novelty [13-15]. For example, in head-and-shoulder type image sequences, the human faces are always the most important areas to users. Thus priority regions can be determined via face tracking techniques such as skin-color detection [10]. These object/feature tracking methods generally require object segmentation, which sometimes is very difficult to implement.

A recent automatic foveation method focused on computing a topographic saliency map based on the human visual attention model [16]. Visually salient regions are selected based on nonlinear integration of low-level visual cues, such as color contrast, temporal flicker, intensity contrast and oriented motion. This method is applicable to automatic prioritization of arbitrary video contents. However it is computationally expensive and its improvement in terms of compression ratio is limited because the saliency map often consists of several regions of interest. If there is only one observer, a more effective way is to record the observer's focusing point at the receiving end (using an eye-tracking device or pointing devices such as a mouse), then apply a foveation filter to video contents at the source [19]. Thus, most of the bandwidth is allocated to high-fidelity transmission of the observer's current region of focus.

This gaze-contingent approach is more appropriate for the telesurgery video transmission because (1) surgery is a task requiring highly focused attention. The surgeon usually closely observes only a small part of the whole scene, purposely ignoring or deemphasizing the details in the surroundings; (2) the scene in surgical video usually consists of various types of biological

tissues and deformable substances (such as fluids). Instead of defining individual objects, it is more natural to define an *attention map* (AM) which is a smooth mask with its highest value at the surgeon's gazing point.

1.3 VIDEO PREPROCESSING

Given the attention map, we can either change the parameters (e.g. quantization) in a specific coding algorithm or preprocess the images before the encoding process. In this work, we choose the latter because video preprocessing methods can (1) effectively improve the performance of video coding; (2) cascade with any high-performance hardware video codec module; (3) allow switching video codec modules for different applications or networking environments.

Preprocessing algorithms improve the performance of a video codec by removing spurious noise and insignificant features from the original images. Statistical analysis of video signals indicates that there is a strong correlation both between successive frames and within the frame [20]. All standard video codecs are based on three fundamental redundancy reduction principles: spatial redundancy reduction, temporal redundancy reduction and entropy coding. Entropy is one of the most fundamental quantities that can be associated with stochastic information sequences. An accurate estimate of the entropy provides an indication of the amount of redundancy contained in the sequence and, consequently, an upper bound on the data compression possible. We can think of preprocessing as a process that modifies the original data in order to decrease the entropy.

When an input image is noisy or contains large amount of textures, the video encoder tends to allocate more bits. A common solution is to pre-filter the images prior to encoding.

There are many conventional spatial smoothing/denoising filters in the literature including the mean filter, median filter, Gaussian smoothing filter, etc. [21]. More advanced edge-preserving filters such as anisotropic diffusion filter [22], adaptive smoothing filter [23], and bilateral filtering [24] have attracted growing interest in recent years. (Detail in section 3.1) These nonlinear image filters target the preservation of important features, e.g. edges, during the denoising/smoothing process. This is a meaningful property for our application because edges are more easily perceived during saccadic or quick voluntary eye movements.

1.4 VIDEO QUALITY ASSESSMENT METHODS

Video preprocessing systems may introduce some distortion or artifacts in the video signal. For the application of telesurgery, videos are ultimately viewed by a surgeon, thus the only "correct" method of quantifying visual video quality is through subjective evaluation. However, an objective video quality metric is also important because it can be used to dynamically monitor and adjust video quality and it can also assist in the optimal design of the preprocessing algorithms.

1.4.1 Objective perceptual assessment methods

Objective video quality metrics can be classified according to the availability of an original (distortion-free) video, with which the distorted video is to be compared. Most existing approaches are known as of "full-reference" (FR), meaning that a complete reference video is

assumed to be known. In many practical applications, however, the reference video is unavailable, and a no-reference (NR) or "blind" quality assessment approach is desirable. Another type of approach called “reduced-reference” (RR) provides a solution that lies between FR and NR methods. The reference video is only partially available, in the form of a set of extracted features made available as the side information to help evaluate the quality of the distorted video.

The simplest and most widely used full-reference quality metric is the mean squared error (MSE), computed by averaging the squared intensity differences of distorted and reference image pixels, along with the related quantity of peak signal-to-noise ratio (PSNR). MSE is appealing because it is simple to calculate, has clear physical meanings, and is mathematically convenient in the context of optimization. But MSE and PSNR are not very well matched to perceived visual quality. It has been discovered that in the primary visual cortex, an input image is represented in a very different manner from pixel domain representations such as MSE. A number of important psychophysical and physiological features of the HVS are not accounted for by the MSE.

In the past three decades, a great deal of effort has been made on the development of quality assessment methods that take advantage of known characteristics of the HVS [25-29]. The majority of the proposed perceptual quality assessment models have followed a strategy of modifying the MSE measure so that errors are penalized in accordance with their visibility. The underlying principle of the error-sensitivity approach is that perceptual quality is best estimated by quantifying the visibility of errors. This is essentially accomplished by simulating the functional properties of early stages of the HVS, as characterized by both psychophysical and physiological experiments. Although this bottom-up approach to the problem has found nearly

universal acceptance [26-28], it is important to recognize its limitations. In particular, the HVS is a complex and highly nonlinear system, but most models of early vision are based on linear or quasi-linear operators that have been characterized using restricted and simplistic stimuli. Thus, error-sensitivity approaches must rely on a number of strong assumptions.

Since human visual perception is highly adapted for extracting structural information from a scene, an alternative complementary framework for distortion assessment was introduced by Wang based on the degradation of structural information [30]. Optimization problems utilizing both traditional and modern metrics will be formulated in Chapter 4.0 , and solved in Chapter 5.0 .

1.4.2 Subjective quality assessment methods

Subjective quality ratings form the benchmark for objective metrics. Formal subjective testing is defined in ITU-R Recommendation 500 [31], which suggests standard viewing conditions, criteria for observer and test scene selection, assessment procedures, and analysis methods. There are three common methods: Double Stimulus Continuous Quality Scale (DSCQS), Single Stimulus Continuous Quality Evaluation (SSCQE) and Double Stimulus Impairment Scale (DSIS). The DSIS method is better suited for evaluating clearly visible impairments such as artifacts caused by transmission errors, for example. In the DSIS test, viewers are shown multiple sequence pairs consisting of a “reference” and a “test” sequence, which are rather short (~10s). Then the viewers rate the amount of impairment in the test sequence on a discrete five-level scale ranging from “very annoying” to “imperceptible”. In Chapter 5.0 , we utilize the DSIS method for subjective evaluation.

1.5 INTRODUCTION OF THE PROPOSED METHOD

Given the gazing point of the surgeon at the receiving end, we will design a continuous map representing the degree of attention a surgeon pays on the corresponding pixel. This attention map is derived from the physical structure of the human eye, with the gazing region falling on the fovea in the retina. Utilizing the attention map, we simulate the eccentricity effects via adaptive spatial filtering. Thus, most bandwidth is allocated to high-fidelity transmission of a small region around the surgeon's focus region, while peripheral regions are highly degraded and transmitted with a smaller bandwidth.

1.6 OUTLINE OF THE DISSERTATION

In Chapter 2.0 we show how human visual system (HVS) process information, propose a time-varying attention map based on a human perception model. In Chapter 3.0 , we present a new adaptive scheme for bilateral filter utilizing the attention map. We show how to select the parameters of the bilateral filter automatically while minimizing visual distortion. In Chapter 4.0 , we investigate several methods of distortion assessment. A new distortion metric based on structure similarity is chosen to optimize our preprocessing algorithm. In Chapter 5.0 , our system implementation and experiment results are reported. In the last chapter, we conclude this dissertation and suggest future works.

2.0 PERCEPTION BASED ATTENTION MODEL

The human visual system (HVS) can be viewed as an information-processing system. In this chapter, we first present the basic anatomy of the eye and list some psychovisual aspects of the HVS. We then build a time-varying acuity map in order to mimic the "foveation process" of the human eye. For the application of telesurgery, the difficulties caused by network transmission delay are discussed, and two feasible solutions are provided.

2.1 THE HUMAN EYE

2.1.1 Basic anatomy and physiology of the eye

As illustrated in Figure 2.1 , light wave travels first through the *cornea*, which is the clear dome at the front of the eye. The light then progresses through the *pupil*, the circular opening in the center of the colored *iris*. Next, the light passes through the *lens*, which is located immediately behind the iris and the pupil. The *ciliary muscle* is a smooth muscle that affects the suspensory ligaments, enabling changes in lens shape for light focusing. The light continues through the *vitreous humor*, the clear gel that makes up about 80% of the eye volume, and then back to a clear focus on the *retina* which acts like a screen. The small central area of the retina is the *macula*, which provides the best vision of any location, and the point of central focus is called

fovea. Within the layers of the retina, light impulses are changed into electrical signals and then sent through the *optic nerve*, along with the visual pathway, to the occipital cortex at the posterior of the brain. The beginning of the optic nerve in the retina is called the *optic disk* (also known as the *blind spot*).

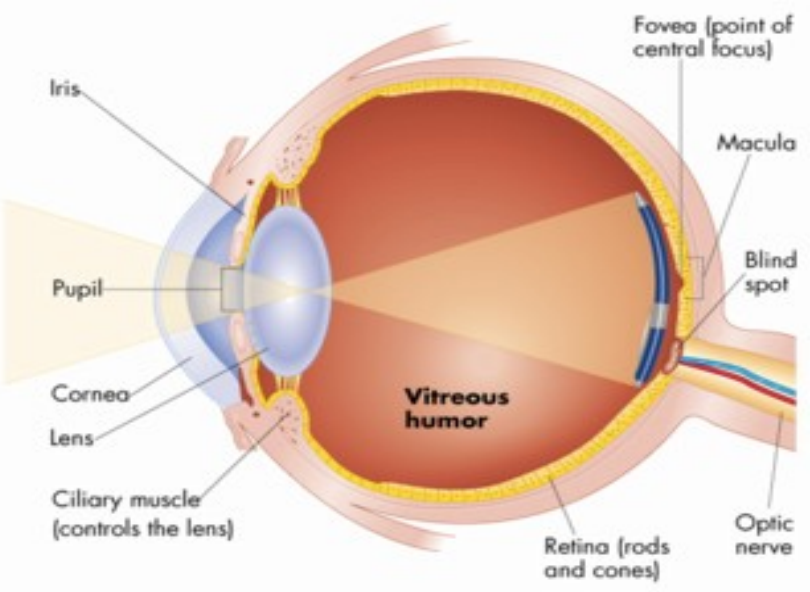


Figure 2.1 Structure of the eye (Adapted from www.artlex.com/ArtLex/s/see.html)

There are two kinds of photoreceptor cells in a human eye, which play a crucial sensory role for our vision: rods and cones. The rods are more sensitive to light and function primarily during night vision. Cones provide fine-grained spatial resolvability of the visual system. The resolution of the image projected on the retina is determined by the densities of the photoreceptor cells which are non-uniformly distributed with the greatest density in the central part of the retina, as shown in Figure 2.2. These photoreceptor cells connect to the ganglion cells and their axons form the optic nerve (illustrated in Figure 2.3(a)). The number of ganglion cells (about one million) that output the sensed signal to the visual cortex is much less than the number of photosensitive cells (about 100 million in total).

Human acuity perception is related to their mapping to the ganglion cells. The ganglion cells are also distributed densely at the central region of retina and coarsely outside. Within the fovea, the signals collected by the cone cells are integrated by ganglion cells at a resolution of 0.03 degrees. Outside the fovea, the rod cells are connected to ganglion cells at a resolution of 3 degrees, a hundred times coarser than those within the fovea. Therefore, detailed vision is only present at the central region of the eye, a fractional part of the entire visual field. The diameter of the highest acuity circular region subtends only 2 degrees, the parafovea (zone of high density) extends to about 4 to 5 degrees, and acuity drops off sharply beyond. At 5 degrees it is only 50%. Figure 2.3(b) shows an example of the perceived image at human visual cortex. Notice that our eye can only perceive the area around the focal point with high accuracy, while the rest are perceived at lower accuracy. This gives us the idea to mimic the "foveation process" by removing those unperceived details from the captured images before sending it to the viewers.

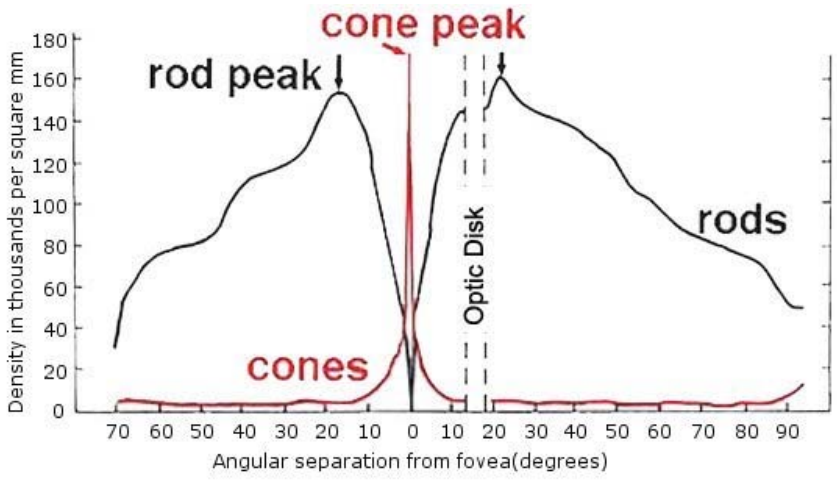
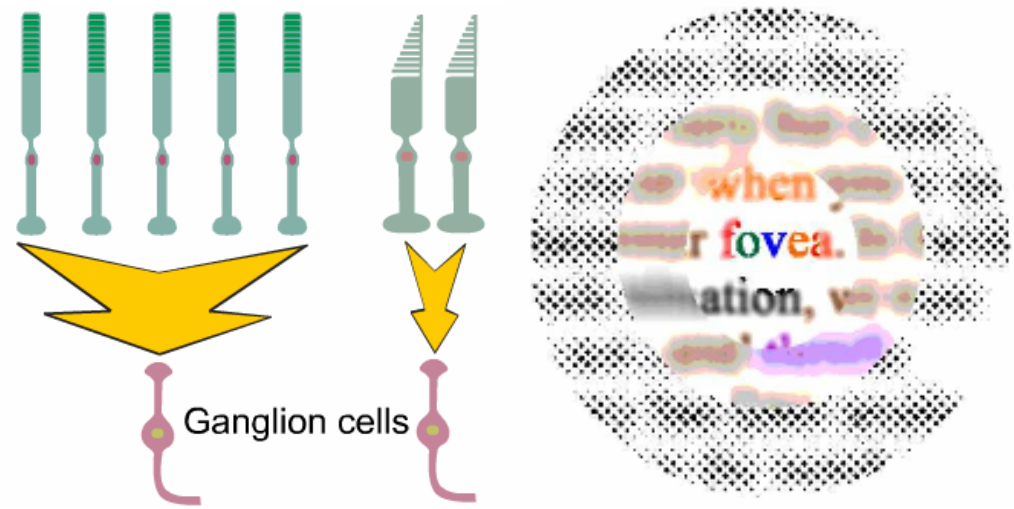


Figure 2.2 Density of photoreceptor cells (Osterberg, 1935)



(a) (b)

Figure 2.3 (a) Connection between ganglion cells and photoreceptor cells (cones and rods); (b) The perceived image at human visual cortex with the center of the image projected on fovea (Adapted from http://www.physpharm.fmd.uwo.ca/undergrad/sensesweb/L1Eye/)

2.1.2 Psychovisual aspects of human visual system

In this section, we describe some observations on pschovisual aspects of the human visual system. The objective is to extract and discriminate different properties of images that are of significance to our perception. We focus on the importance of edges to our visual perception.

In binocular vision, when two views are presented with two images that are different, both images will generally be seen at the same time superposed on one another in the field of view. Usually, in some locations of the field of view, one image dominates the other, and vice versa in other parts of the field. As an example, consider two bars, one vertical and one horizontal, as shown in Figure 2.4(a) and (b). When the vertical bar is seen by the left eye and the horizontal bar is seen by the right eye, the total effect will be an image similar to Figure

2.4(c). The four arms of the cross are perfectly black at their ends and almost entirely white near the center square, with transitions in between. Edge information arouses our visual perception.

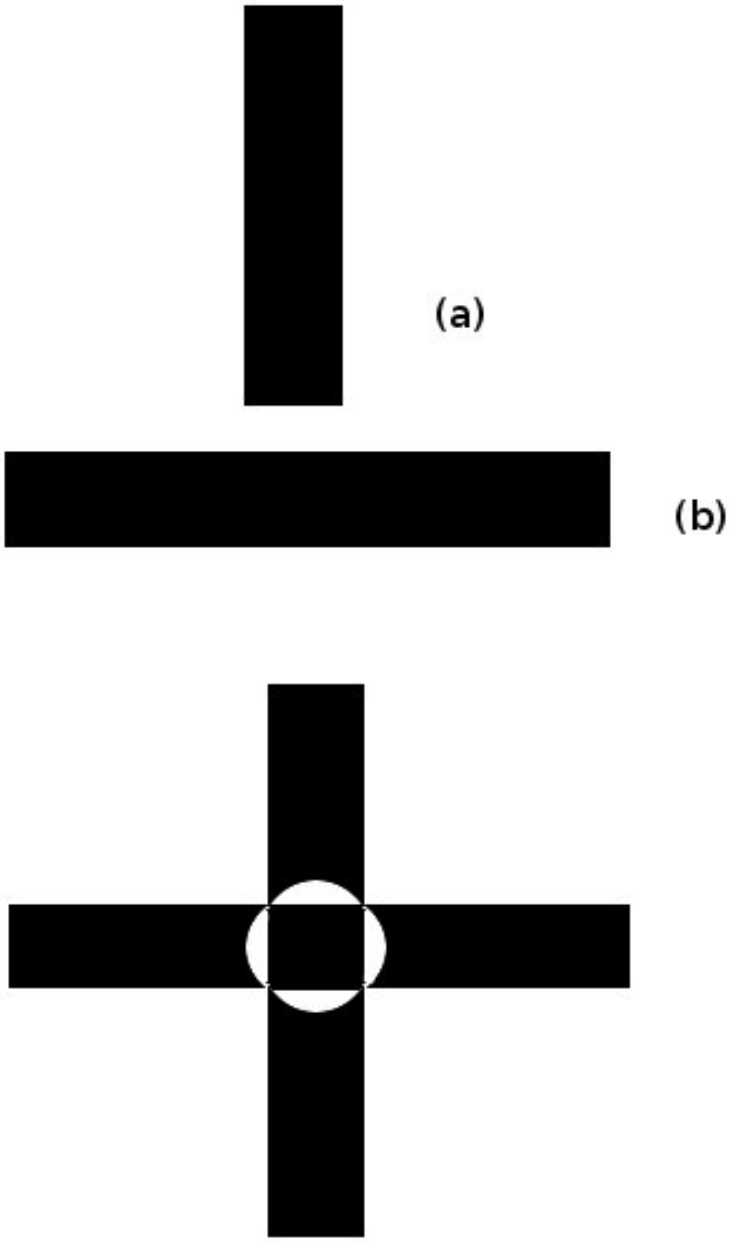
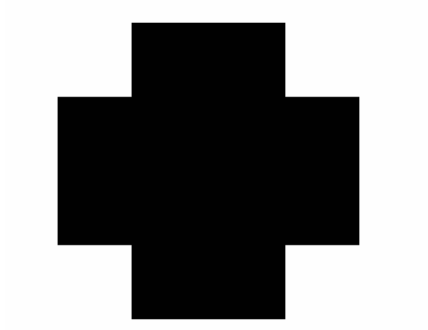
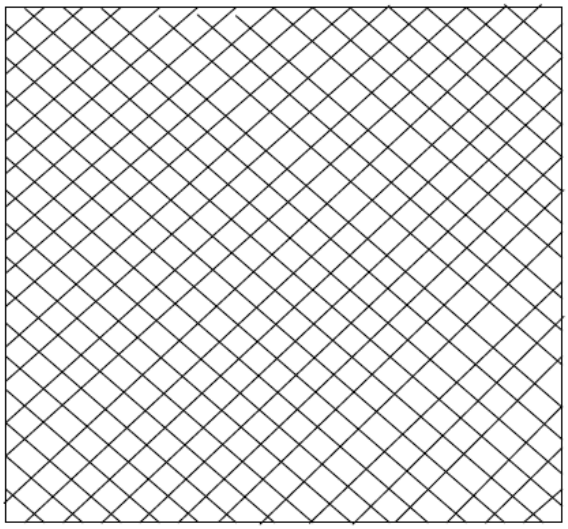


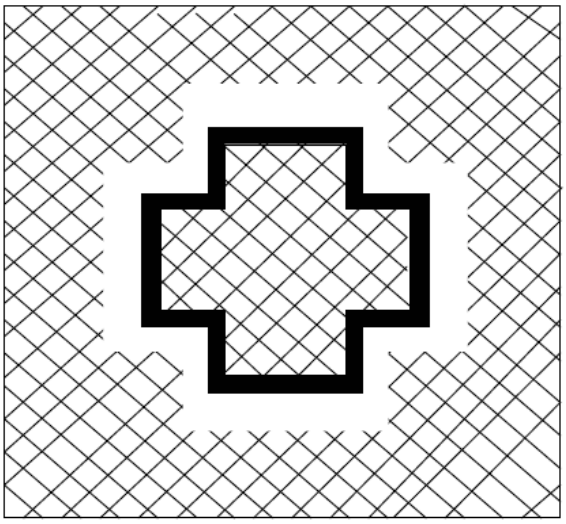
Figure 2.4 (a) Horizontal bar seen by the left eye; (b) Vertical bar seen by the right eye; (c) The binocular perception of images (a) and (b) (Adapted from [32])



(a)



(b)



(c)

Figure 2.5 (a) Cross bar image seen by the left eye; (b) Pattern image seen by the right eye; (c) Binocular perception of images (a) and (b) (Adapted from [32])

Another example is shown in Figure 2.5. A black cross is seen by the left eye (Figure 2.5 (a)), while a pattern image is seen by the right eye (Figure 2.5(b)). The binocular perception of these two images is shown in Figure 2.5(c).The edge information cannot be combined "linearly", thus the pattern around the edge of the cross disappear. We can see that edges of relatively high strength have stronger influence on our perception, and vice versa. This example showed that strong edges plays a more important role (for visual peception) than texture and smooth area. Smooth areas are characterized by slow changes in the intensity value, whiles edges are characterized by sharp intensity value changes along the border of an object. The term *texture* has been introduced to describe the irregularities observed on object surfaces, which often consists of wearker edges.

2.1.3 Eye movement

From section 2.1.1, we know that perception resolution decreases with increase in eccentricity. The foveal viewing area is small and to change the point of regard, our eyes move constantly so that the image of the point of regard is projected onto the fovea. There are three kinds of eye movements are closely related our research: saccades, smooth pursuit and fixation.

Saccades usually travel up to 15° arc before requiring head movements. Saccadic movements are considered programmed movements because of their ballistic characteristics (i.e., after planning the saccadic movement, it is automatically performed by the visual system without the ability to halt midway through the action). The planning phase of a saccade introduces latency and it is positively correlated with saccade amplitude. If the time and location of the appearance of a stimulus is known beforehand, the latencies decrease, and the peak velocity of the saccade increases. Such saccades are also known as express saccades. Another source of

latency in saccadic movements is refractoriness where a minimum amount of time (150ms) must elapse before another saccade can execute. During saccades, reduced visual input is received during a saccade due to the high speed known as saccadic suppression. In fact, visual acuity is also reduced during the planning stage of a saccade. Although visual input is reduced, some is still being processed, such as surface and edge information.

Smooth pursuit eye movements track moving targets to ensure that the image of the target remains on the fovea. The end result is to minimize target motion on the foveal area by moving it at the same speed as the moving object. Peripheral vision areas, however, become blurry from the eye motion. The effect is similar to panning a camera to track a moving race car. The brain, however, must differentiate between a moving object along the retina, and a stationary object that is moving along the retina because of eye movements. In both cases, the image is moving along the retina.

Fixations are eye movements to keep the eye steady at a stationary point of interest so that its retinal image is projected onto the fovea. Even though the image is perceived to be still, fixation eye movements are characterized by small eye movements such as tremor, drift and microsaccades. Tremors occur at high frequency and are relatively small. Drift is a gradual movement away from a fixation target that travels slower than tremors, but the distances in drift movements are larger than tremors. Microsaccades are corrective movements so that the image of the fixation target returns to the fovea after drift has occurred.

2.2 ATTENTION MODEL

In order to mimic the "foveation process" and predict how much attention the surgeon would give to certain area of the video, we first build an acuity map utilizing a model which is derived from psychovisual studies.

2.2.1 Acuity map

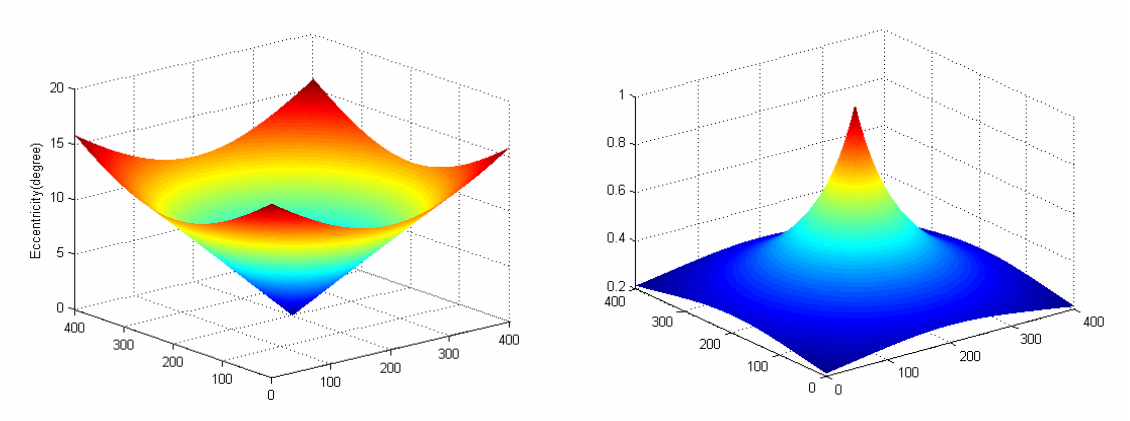
As mentioned before, the perception of acuity depends on the spatial distribution and mapping of cones, rods and ganglion cells, and the mapping of the visual fields across the visual cortex. The results from previous psychophysical tests suggest that visual acuity is nearly halved at 1 degree from the foveal center and decreased to one quarter at 5 degrees from the foveal center. A large number of functions have been suggested for contrast sensitivity function (CSF). Some of them are based on anatomical considerations and some are based on psychovisual empirical studies. Since these models have similar attenuating profiles, we implement a simple model [33]. The acuity $A(x, y)$ at pixel (x, y) can be calculated as follows:

$$A(x, y) = \frac{1}{1 + k\theta(x, y)} \tag{2.1}$$

Where $\theta(x, y)$ is the eccentricity in visual angle and k is a constant. The value of k can be obtained by fitting the dataset of empirical results. Given the fixation point (x_C, y_C) and the viewing distance D, we can calculate the eccentricity:

$$\theta(x,y) = \frac{180}{\pi}\arctan\left(\frac{\sqrt{(x-x_C)^2+(y-y_C)^2}}{D}\right) \tag{2.2}$$

Once $A(\vec{x})$ is determined, we will convert it to an attention map according to specific physical parameters (screen size, viewing distance, etc.) of the *da Vinci* robotic surgical system available at the University of Pittsburgh Medical Center, which will be described further in section 5.3. Figure 2.6 shows an example of the eccentricity map and the acuity map.



(a) (b)

Figure 2.6 (a) The eccentricity map with D=1000 pixels (b) The spatial acuity model with a hypothetic k=0.24.

2.2.2 Eye tracking

In order to determine the time varying attention map of the surgeon's gaze, an eye tracking system can be used. To date, there have been three methods developed to track eye movements [34]. In the first method, a small device is attached to the eye, such as a special contact lens with an embedded mirror or magnetic field sensor. This method seems to provide the most accurate

results, but is cumbersome to use. The second method relies on the Electo-Oculogram (EOG) measured from several specially arranged skin-surface electrodes and it is based on the fact that the eye has a standing electrical potential, with the cornea being positive relative to the retina. However, this method is somewhat unreliable for measuring slow eye movements and fixed gaze positions, thus not suitable for our application. The third method utilizes natural or infrared light to be reflected from the eye and sensed by a digital video camera. Digital image processing techniques are then applied to extract eye rotation from changes in reflections [34-37]. This method appears to be the most suitable method in our case because it does not require any attachment to the eye or the skin. We plan to use a small infrared digital camera to acquire eye images and investigate the tracking algorithms using three different pairs of automatically detected features: 1) the corneal reflection and the center of the pupil; 2) reflections from the front of the cornea and the back of the lens; and 3) reflections from a number of selected retinal blood vessels. The most feasible method will be determined with respect to the telesurgical application. In cases where more sensitive eye tracking is required in the study of the eye motion model, the approach using a special contact lens will be considered in our research.

2.2.3 Impact of feedback delay

Feedback delay is the period of time between the instance of the eye position is detected by an eye tracker, and the moment when a perceptually encoded frame is displayed. This delay should be taken into consideration during the real time telesurgery video preprocessing because future eye movements should fall within the highest quality region of the video. Otherwise, a surgeon would notice the degradation of the video. The length of the feedback delay ranges from 20 ms to a few seconds, depending on the properties of the transmittion network. Saccades can move

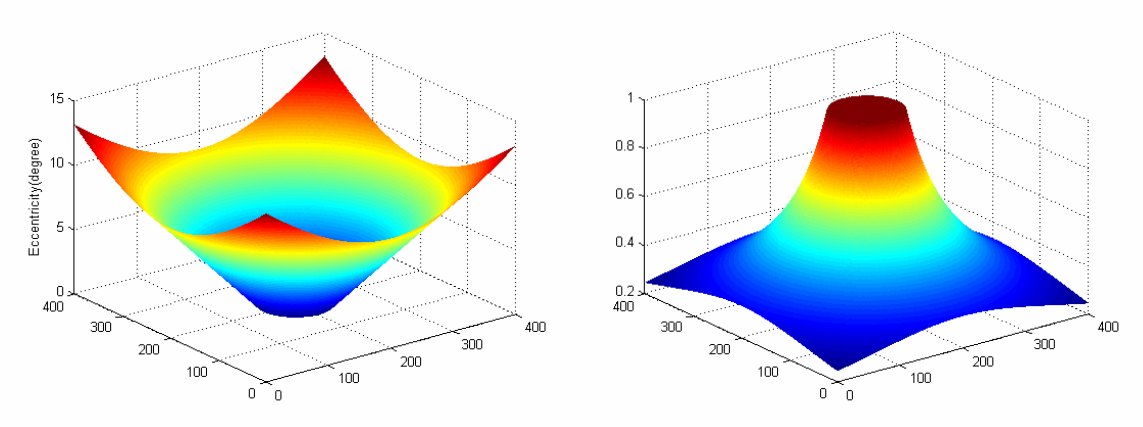
the eye position more than 10-100 degrees during that time. In order to slove this problem, we provide two feasible solutions:

Solution 1: we create an eye gazing window which represents an area where eye-gazes are containted. Within the ellipse shaped eye gazing window, the acuity value is set to 1, while the acuity value outside of the window can be calculated as follows:

$$A(x,y,t)=\begin{cases} 1 & \text{when } \sqrt{\left(\dfrac{x-x_C}{R_x(t)/R_y(t)}\right)^2+(y-y_c)^2} < R_y(t) \\ \dfrac{1}{1+k\cdot\theta(x,y,t)}, & \text{otherwise} \end{cases} \tag{2.3}$$

$$\theta(x,y,t)=\frac{180}{\pi}\arctan\left(\frac{\sqrt{\left(\dfrac{x-x_C}{R_x(t)/R_y(t)}\right)^2+(y-y_c)^2}-R_y(t)}{D}\right) \tag{2.4}$$

where $R_x(t), R_y(t)$ are the radial components of the eye gazing window which centered at (x_C, y_C). $R_x(t), R_y(t)$ are determined by the feedback delay at time t. Komogortsev [38] introduces a mathematical model of the human eye that uses anatomical properties of the Human Visual System to predict eye movement trajectories. The eye mathematical model is transformed into a Kalman filter form to provide continuous eye position signal prediction during all eye movement types. In Figure 2.7, the eccentricity map and acuity map are modified as we set $R_x(t), R_y(t)$ equal to 50 pixels.



(a) (b)

Figure 2.7 (a) Modified eccentricity map with D=1000 pixels, (b) Modified spatial acuity model with a hypothetic k=0.24. In both cases, $R_x = R_y = 50$ pixels.

Solution 2: We replace eye tracker with an automatic surgery tool tracking system. In surgery, the attention focus in the surgical field-of-view is usually located within a neighborhood centered at the tip of an active surgical instrument [39]. Thus the eye gazing window can be determined by tracking the surgical instrument within the video. Since a surgical instrument is usually a rigid metallic object with distinct characteristics from biological tissue, we utilized an object tracking algorithm based on edge, color, and motion information extracted from each video frame and compare this information to that of a known template. The shape of the eye gazing window is designed to vary progressively to reduce the total area of window while providing a natural visualization of the surgical video, as shown in Figure 2.8. Within the estimated eye gazing window, the acuity value is set to be 1.

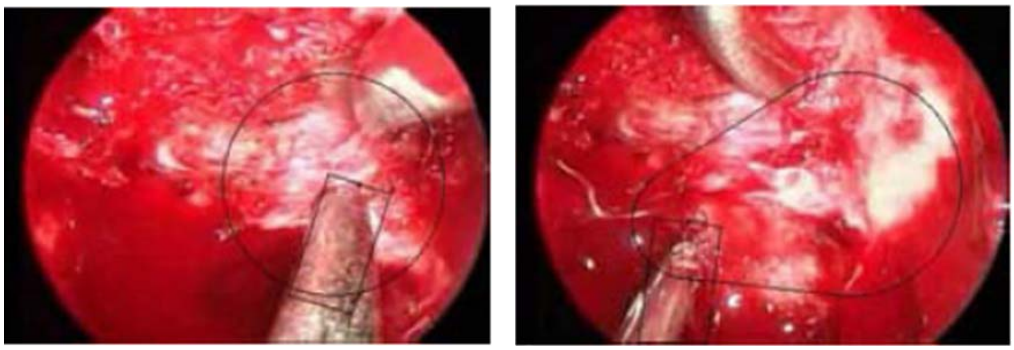


Figure 2.8 Results of eye gazing area tracking. (a) A circular gazing window is centered at the detected tip of an active surgical instrument (a suction tool). (b) The eye gazing window after 152 ms when the suction tool quickly moves to the lower left. The shape of the eye gazing window is progressively changed to allow observation of both the previous and current locations along with the trajectory of tool motion.

3.0 VIDEO PREPROCESSING ALGORITHM

The human vision provides a tremendous insight for perceptual data reduction in robotic surgery. Only a small region of 2 to 5 degrees of the visual angle around the fixation point is captured at high resolution, with logarithmic resolution falloff with eccentricity [40]. The purpose of preprocessing is to remove the unnecessary contents outside the region of interest where small details are deemphasized by the surgeon. In this chapter, we first investigate several image smoothing techniques focusing on edge-preserving nonlinear filters. Then, we propose a spatially-variable preprocessing technique that will be adaptive to both the aforementioned acuity map and the network available bandwidth. Finally, we form an optimization problem which will minimize the visual distortion.

3.1 IMAGE SMOOTHING FILTERS

Our preprocessing is based on image smoothing (blurring), which decreases the information resulting from the discrete cosine transform (DCT). A spatially-variable preprocessing can be applied to the input frames so that less critical regions are strongly blurred.

We first investigate the Gaussian filter. The Gaussian filter outputs a “weighted average” of each pixel's neighborhood, with the average weighted more towards the value of the central pixels. This is in contrast to the mean filter's uniformly weighted average. Because of this, a

Gaussian provides gentler smoothing and preserves edges better than a similarly sized mean filter. A Gaussian filter has identical forms in the spatial domain and in the frequency domain (the Fourier transform of a Gaussian is also a Gaussian). The Gaussian is the only real-valued function that minimizes the product of spatial and frequency domain spreads, according to the Heisenberg uncertainty principle [41]. A Gaussian filter is given by:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{3.1}$$

where σ is the standard deviation of the Gaussian and it controls the extent of smoothing. The result of Gaussian smoothing is shown is Figure 3.1. We notice that Gaussian filtering produced blurred edges because it does not take into account intensity variations within the image.

Figure 3.1 (a) Original image (b) Gaussian filtered image with σ =2 and (c) σ =4

In a wide variety of applications, it is necessary to smooth images while preserving the edges. This is also a meaningful property for our application because edges are more easily perceived during saccadic or quick voluntary eye movements. Most visual neurons in the primary areas of the visual cortex in mammals react to intensity jumps or edges which have a particular orientation, and especially so when these edges move across the image at right angles to their orientation [42]. Therefore, strong edges should be preserved after preprocessing. Three

nonlinear edge-preserving smoothing filters are investigated here: (a) anisotropic diffusion (b) adaptive smoothing, and (c) bilateral filtering.

(a) Anisotropic Diffusion:

This method was developed by Perona and Malik in the early 1990's [22]. Diffusion algorithms remove noise from an image by modifying the image via a partial differential equation (PDE). For example, consider applying the isotropic diffusion equation (the heat equation) given by $\frac{\partial I(x,y,t)}{\partial t} = div(\nabla I)$, using the original image I(x,y,0) as the initial condition. Modifying the image according to this isotropic diffusion equation is equivalent to filtering the image with a Gaussian filter.

Perona and Malik replaced the diffusion equation with

$$\frac{\partial I(x,y,t)}{\partial t} = div[g(\|\nabla I\|)\nabla I]) \tag{3.2}$$

Where $\|\nabla I\|$ is the gradient magnitude, and $g(\cdot)$ is an "edge-stopping" function. This function is chosen to satisfy $g(x) \to 0$ when $x \to \infty$ so that the diffusion is "stopped" across edges. Perona and Malik discretized their anisotropic diffusion equation as follows:

$$I_s^{t+1} = I_s^t + \frac{\lambda}{|\eta_s|}\sum_{p\in\eta_s} g(\nabla I_{s,p})\nabla_{s,p} \tag{3.3}$$

Where I_s^t is a discretely sampled image, s denotes the pixel position in a discrete, two-dimensional (2-D) grid, and t now denotes discrete time steps (iterations). The constant $\lambda \in \Re^+$ is

a scalar that determines the rate of diffusion, η_s represents the spatial neighborhood of pixel s, and $|\eta_s|$ is the number of neighbors (usually four, except at the image boundaries). Image gradient in a particular direction was linearly approximated as

$$\nabla I_{s,p} = I_p - I_s^t, \quad p \in \eta_s . \tag{3.4}$$

Qualitatively, the effect of anisotropic diffusion is to smooth the original image while preserving brightness discontinuities. The choice of $g(\cdot)$ can greatly affect the extent to which discontinuities are preserved. Perona and Malik suggested two g-functions $g_1(x)$ and $g_2(x)$.

$$g_1(x) = \frac{1}{1+\frac{x^2}{k^2}} \tag{3.5}$$

$$g_2(x) = e^{-(x^2/k^2)} \tag{3.6}$$

Black etc. developed a statistical interpretation of anisotropic diffusion from the point of view of robust statistics [43]. They showed that the g-functions are closely related to the error norm and influence function in the robust estimation framework. For example, $g_1(x)$ is related to the Lorentzian error norm and $g_2(x)$ is related to Leclerc error norm. They also proposed an alternative g-function based on Tukey's biweight robust error norm, called Tukey's function.

$$g_3(x) = \begin{cases} \frac{1}{2}\left[1-\frac{x^2}{k^2}\right]^2, & |x| \le k \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

It was shown [43] that Tukey function is more robust, and produces sharper boundaries. There are two parameters which controls the degree of smoothing: k in the g-function, and the number of iterations, N. As k goes to infinity, g(x,y) will approach a constant value. The anisotropic diffusion filter becomes the Gaussian smoothing filter. Meanwhile, as the number of

iterations increases, the diffusion will affect a larger neighborhood area, thus removing more details. In order to adapt the blurring effect to human perception, those two parameters should vary spatially—smaller within the area of high acuity vision and larger outside this area.

(b) Adaptive Smoothing:

The general idea behind adaptive smoothing is to apply a versatile operator to adapt itself to the local topography of the image [23]. Given an image $I^{(t)}(x,y)$, where (x,y) denotes space coordinates, an iteration of adaptive smoothing yields:

$$I^{(t+1)}(x,y) = \frac{\sum_{i=-1}^{1}\sum_{j=-1}^{1} I^{(t)}(x+i,y+j)w^{(t)}(x+i,y+j)}{\sum_{i=-1}^{1}\sum_{j=-1}^{1} w^{(t)}(x+i,y+j)} \tag{3.8}$$

where the convolution mask:

$$w^{(t)}(x,y) = f(d^{(t)}(x,y)) = \exp\left(-\frac{\left|d^{(t)}(x,y)\right|^2}{2k^2}\right) \tag{3.9}$$

where k is the variance of the Gaussian mask and $d^{(t)}(x,y)$ is defined as the magnitude of the gradient $d^{(t)}(x,y) = \sqrt{G_x^2 + G_y^2}$. The gradient is defined as $(G_x, G_y)^T = (\frac{\partial I^{(t)}(x,y)}{\partial x}, \frac{\partial I^{(t)}(x,y)}{\partial y})^T$.

The parameter k determines the magnitude of the edges to be preserved during the smoothing process. If k is chosen to be large, all discontinuities disappear, and the result is the same as if Gaussian smoothing were used. If k is chosen to be small, then all the discontinuities are preserved, and no smoothing is performed. Similar to the previously discussed anisotropic diffusion method, the number of iterations N also affects the degree of smoothing. Given the acuity map $A(x,y)$, we will adapt k and N to the local acuity value.

(c) Bilateral Filtering:

Bilateral Filter is a non-iterative method which combines domain and range filtering. It was developed by Tomasi and Manduchi in 1998 [24] as an alternative to anisotropic diffusion. Given an input image $\mathbf{I}(\vec{x})$, using continuous representation, the output image $\mathbf{J}(\vec{x})$ is obtained by:

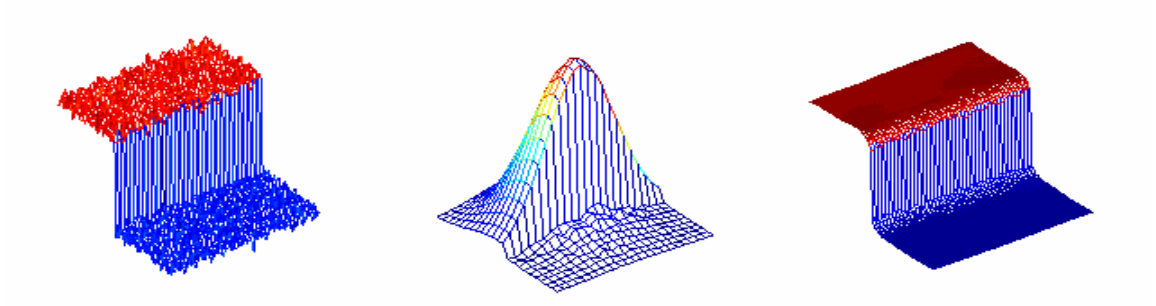
$$\mathbf{J}(\vec{x}) = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{I}(\vec{\xi})c(\vec{\xi},\vec{x})s(\mathbf{I}(\vec{\xi}),\mathbf{I}(\vec{x}))d\vec{\xi}}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(\vec{\xi},\vec{x})s(\mathbf{I}(\vec{\xi}),\mathbf{I}(\vec{x}))d\vec{\xi}} \tag{3.10}$$

where $\vec{x} = (x_1, x_2)$, $\vec{\xi} = (\xi_1, \xi_2)$ are space variables and $\mathbf{I} = (I_1, I_2, I_3)$ is the intensity value of a color pixel. The convolution mask is the product of the function $c(\cdot)$ and $s(\cdot)$, which represent "closeness" in space and "similarity" in intensity, respectively. The discrete version of bilateral filtering can be written as:

$$\mathbf{J}(\vec{x}) = \frac{\sum_{i=-S}^{S}\sum_{j=-S}^{S} \mathbf{I}(x_1+i, x_2+j)w(\vec{x},\vec{\xi})}{\sum_{i=-S}^{S}\sum_{j=-S}^{S} w(\vec{x},\vec{\xi})} \tag{3.11}$$

where $w(\vec{x},\vec{\xi}) = c(\vec{x},\vec{\xi})s(\vec{x},\vec{\xi})$. For simplicity, we use Gaussian filter to form the kernel function.

$$w(\vec{x},\vec{\xi}) = \exp\left(\frac{-(\|\vec{\xi}-\vec{x}\|)^2}{2\sigma_D^2}\right)\exp\left(\frac{-(\|\mathbf{I}(\vec{\xi})-\mathbf{I}(\vec{x})\|)^2}{2\sigma_R^2}\right) \tag{3.12}$$



(a) **(b)** **(c)**

Figure 3.2 (a) An edge (intensity step 100) perturbed by Gaussian noise with standard deviation equal to 10 (in intensity). (b) Combined similarity weights for a 23x23 neighborhood centered at a pixel slightly (two pixels) to the right of the step. (c) The edge in (a) after bilateral filtering with $\sigma_R = 50$ (in intensity) and $\sigma_D = 5$ (in pixels).

Figure 3.2 gives an example of how the bilateral filter will work. Consider a sharp boundary between a dark and a bright region, as in Figure 3.2(a). When the bilateral filter is centered, say, on a pixel on the bright side of the boundary, the similarity function $s(\cdot)$ assumes values close to 1 for pixels on the same side, and values close to 0 for pixels on the dark side. The convolution mask is shown in Figure 3.2(b) for a 23x23 filter support centered two pixels to the right of the step in Figure 3.2(a). As a result, the filter replaces the bright pixel at the center by an average of the bright pixels in its vicinity, and essentially ignores the dark pixels. Conversely, when the filter is centered on a dark pixel, the bright pixels are ignored instead. Thus, as shown in Figure 3.2(c), noise was removed and a sharp edge was preserved at the same time.

The bilateral filtering was originally proposed as an intuitive tool without theoretical connection to the classical approach [24]. Recently, Barash [44] has shown that the bilateral filtering represents a large class of nonlinear image filters and the relationship between the bilateral filtering and nonlinear diffusion equation was explored. It has been pointed out that the

nature of bilateral filtering resembles that of anisotropic diffusion, while adaptive smoothing serves as a link between the two approaches. In anisotropic diffusion, several iterations of adaptive smoothing are performed. In bilateral filtering, the kernel is extended to become globally dependent on intensity, which leads to the increase of window size and the abandon of the need for iterations. Furthermore, Elad [45] bridged the bilateral filtering with anisotropic diffusion (AD), weighted least squares (WLS) and robust estimation (RE) by stating that the bilateral filtering emerges from the Bayesian framework. This was derived by introducing a penalty function which penalizes nonsmoothness with distant neighbors as well. The penalty function can be minimized using the Jacobi algorithm or the diagonal normalized steepest descent (DNSD). It was shown that the bilateral filter can be obtained after first iteration of the Jacobi algorithm [46]. The output can be expressed as a time varying convolution, and the coefficients also include the geometric part and the photometric part.

After comparing all three edge preserving image smoothing filters, we decided to adopt the bilateral filtering in our preprocessing algorithm because it is easy to understand and it is non-iterative.

3.2 ADAPTIVE BILATERAL FILTERING

The bilateral filtering is known to be locally adaptive to image features such as textures and edges. However it is not adaptive to human perception. The peak contrast sensitivity falls off quickly as the eccentricity increases, which means that the eye can not perceive some of the details (such as texture) when the pixels are located away from the gazing point. We approach

this problem by proposing adaptive bilateral filtering. The parameters of the bilateral filtering are adjusted globally to match the contrast sensitivity variance of the human eye.

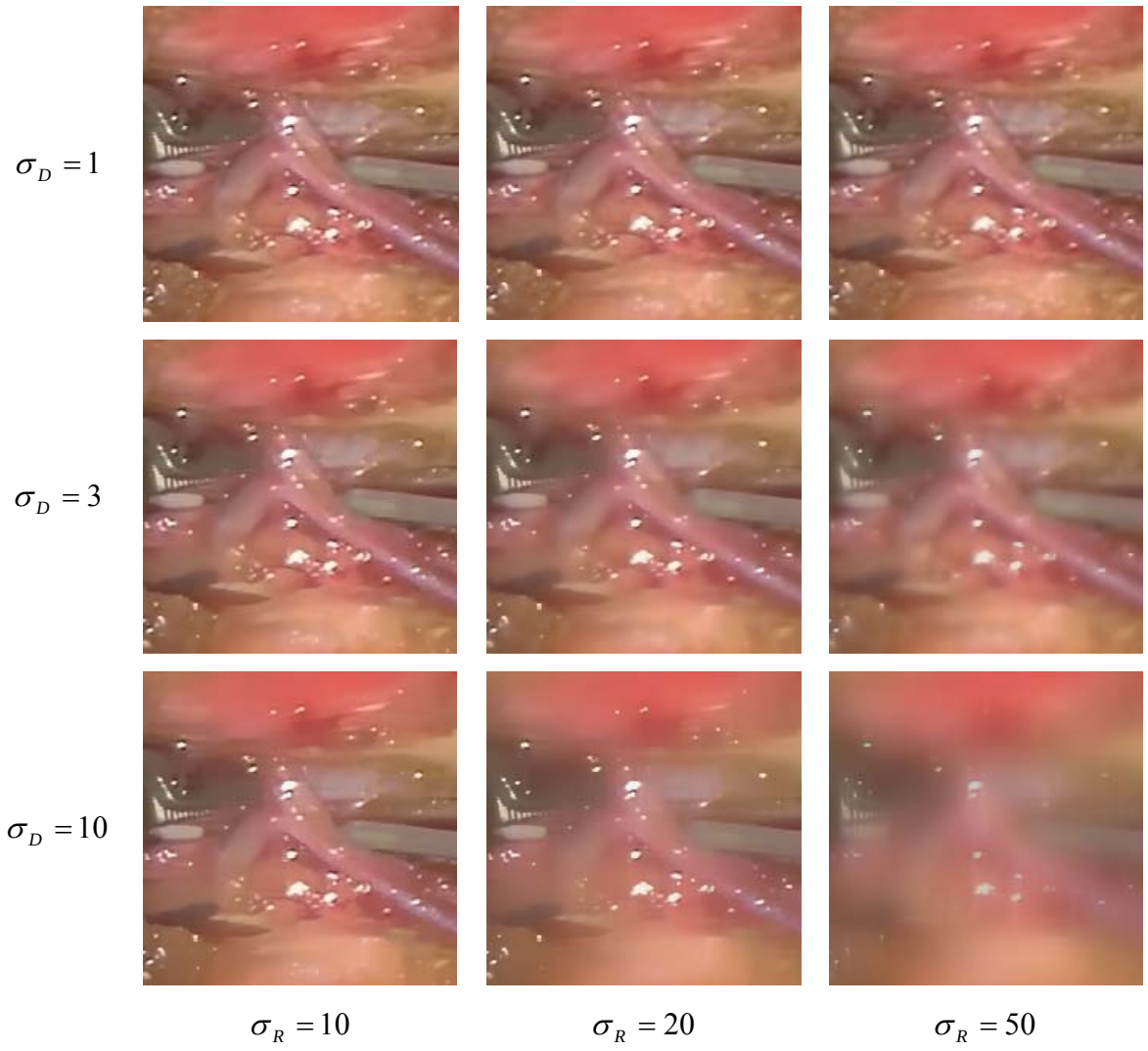


Figure 3.3 Center of the surgery frame processed with bilateral filters with various geometric and photometric spread values.

The bilateral filter is controlled by two parameters σ_D and σ_R (see Eq. 3.12). The geometric spread σ_D is a constant determined by the desired amount of low-pass filtering. A large σ_D results in more of a blur effect since more neighbors are combined for diffusion. The photometric spread σ_R is chosen to achieve the desired amount of combination of pixel values.

Pixels with values much closer to each other than σ_R are mixed together, while those with values much more distant than σ_R are not. For very large (infinity) values we get a simple uniform non-adaptive filtering which is known to degrade edges, using too small values reduces the smoothing effect. The effect of different values of the parameters is shown in Figure 3.3. Rows correspond to different value of σ_D, while columns correspond to different value of σ_R. For smaller σ_R's, the similarity weight dominates because it preserves edges. As σ_R increases, detail is lost but edges are still preserved. When σ_R is large with respect to the overall range of values in the image, the bilateral filter acts as a standard Gaussian filter because all the pixels in the neighborhood have about the same weight from the similarity function. This effect can be seen in the last column of Figure 3.3. As of the geometric spread σ_D, when the value is small, the smoothing effect is reduced because only a few neighbors are combined. On the contrary, when σ_D is large, the output image appears to be hazy.

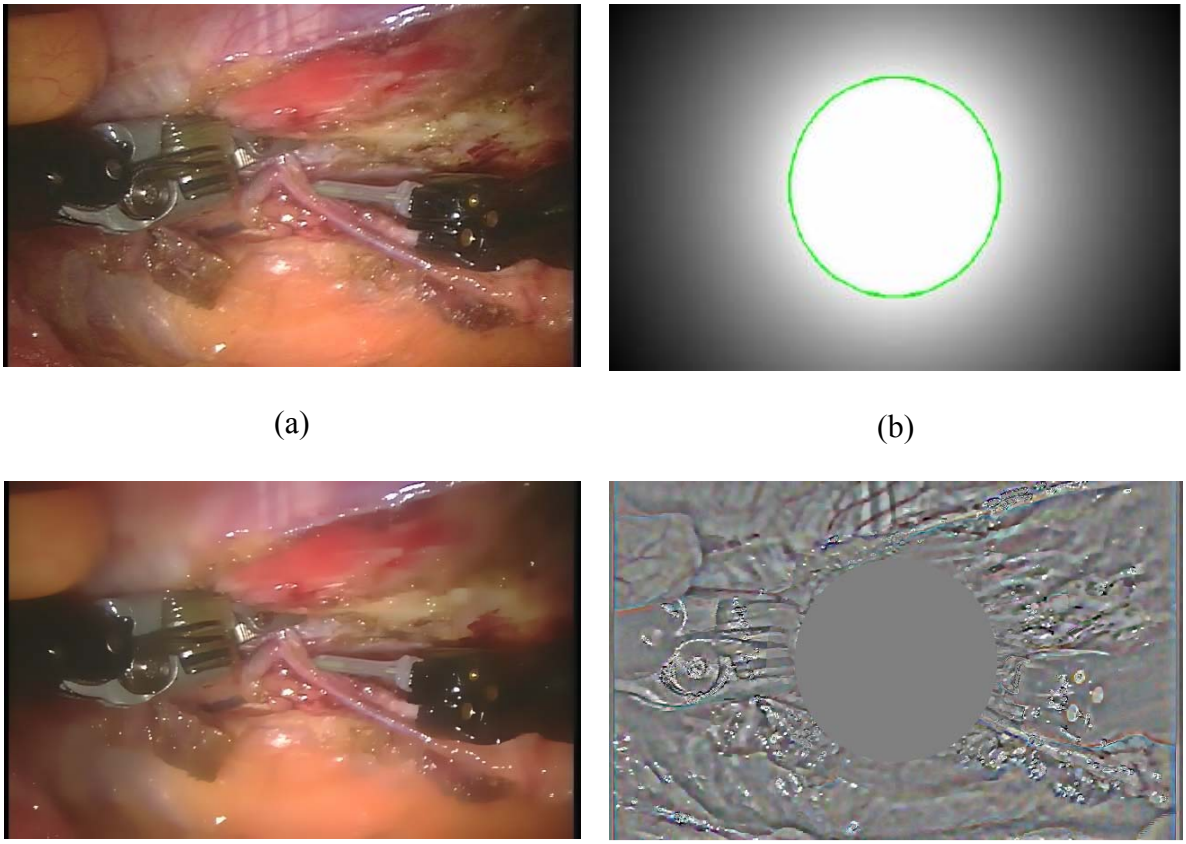
In our approach, we adjust the parameters of the bilateral filter according to the acuity model of human eye (see section 2.2.1). Close to the gazing point, a smaller geometric spread σ_D is utilized which produces less smoothing effect and vice versa. The photometric spread σ_R is adapted in a similar fashion, in that a smaller σ_R is used at the pixels closer to the gaze point to maintain the contrast ratio. Figure 3.4 shows the result from our preliminary experiment. One frame from the original video is shown in Figure 3.4(a), which represents a snapshot of a robotic prostatectomy using the *da VinCi* surgical system. The resolution and frame rate of this video were, respectively, 640x400 pixels and 30 fps. In the experiment, we predefined three sets of parameters for regions with different importance levels, denotes as ROI, transit, and non-ROI, respectively, as shown in Table 3.1. We assume that the gaze point is at the center of the frame,

and the simulated attention map is shown in Figure 3.4(b). The circle indicates the ROI within which the quality of the video is to be preserved. Figure 3.4(c) shows the preprocessed frame based on the attention map.

Table 3.1 Pre-defined Parameters for Regions with Different Importance Value

Regions	Geometric Spread(pixels)	Photometric Spread (intensity levels)
ROI(A(x)=1)	1	1
Transit(0.5<=A(x)<1)	5	7
Non-ROI (0<=A(x)<0.5)	10	20

We compressed the original video frames (200 frames in total) with and without the adaptive bilateral filtering. In our experiment, we chose Microsoft Media 9 codec, which provides support for a wide range of bit rates for high-quality video streaming or downloading. At the same quality settings (VBR 90), the average bitrate for the un-preprocessed segment was 2707.08 Kbps, while the average bitrate for the preprocessed segment was 1372.83 Kbps, a more than 50% reduction compare with the un-preprocessed segment. The difference between the original and filtered images is shown in Figure 3.4(d) which has been scaled to the values between 0 and 255 to facilitate display. Notice that a considerable amount of texture was removed from the original frame outside the ROI.



(a) (b)

(c) (d)

Figure 3.4 (a) Original frame; (b) Attention map; (c) Preprocessed frame; (d) Difference between (a) and (c).

We have already shown that given the observer's current attention map, we were able to save about 50% bandwidth by preprocessing the video frames, while the preprocessed frame and the original frame are visually indistinguishable within the ROI. In our next experiment, we modify the shape of the attention map so that it is adaptive to the traffic condition of the network. We design a new map called **control map** which is a function of the attention map $A(\vec{x},t)$ and effective bandwidth $B(t)$:

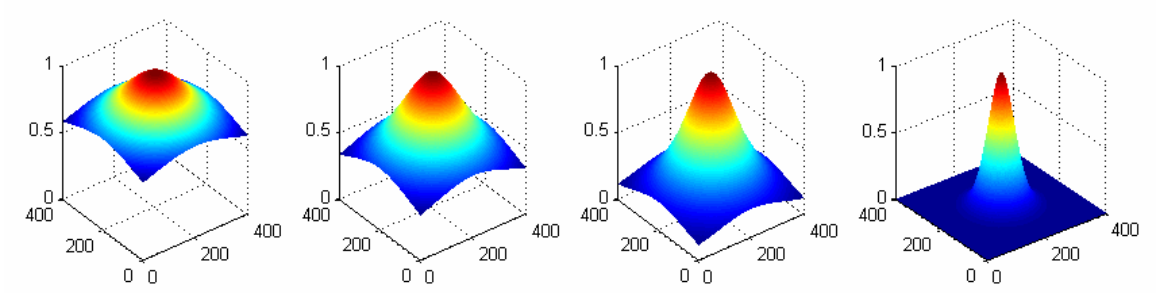
$$C(\vec{x},t) = A(\vec{x},t)^{\frac{B_0}{B(t)}} \tag{3.13}$$

where B_0 is the necessary bandwidth to represent the video frame at the acceptable visual quality. When $B(t) >> B_0$, the control map is almost flat and equal to 1, as shown in Figure 3.5(a). When $B(t)$ decreases, the control map becomes shaper, as shown in Figure 3.5(b)-(d). The control map can be used to determine the local parameters for the bilateral filter, according to:

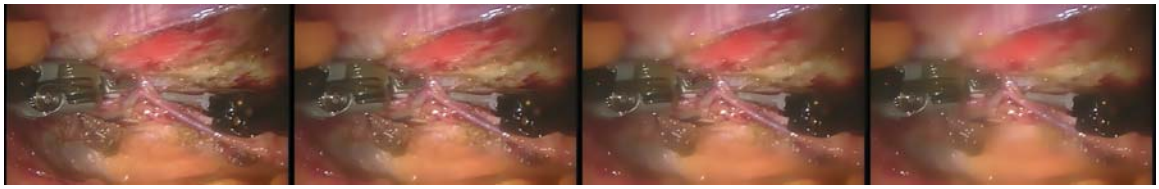
$$\sigma_D^{'}(\vec{x},t) = \sigma_{DU} - (\sigma_{DU} - \sigma_{DL})C(\vec{x},t) \tag{3.14}$$

$$\sigma_R^{'}(\vec{x},t) = \sigma_{RU} - (\sigma_{RU} - \sigma_{RL})C(\vec{x},t) \tag{3.15}$$

where σ_{DU} and σ_{DL} are the pre-selected upper and lower limit of the geometric spread σ_D and σ_{RU} and σ_{RL} are those of the photometric spread σ_R, respectively.



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



(e) (f) (g) (h)

Figure 3.5 (a)-(d): Control maps for decreasing effective bandwidth B(t). (e)-(h): Resulting preprocessed frames using control maps (a)-(d).

Figure 3.5(e)-(h) show the preprocessed frames using the control maps of panels (a)-(d). Again, we compressed the preprocessed video segments using Windows Media Encoder 9. The resulting average bandwidths were 2230.39 Kbps, 1810.83 Kbps, 1339.94 Kbps and 1085.42 Kbps, respectively. (Compared to 2707.08 Kbps original bandwidth)

3.3 AUTOMATIC PARAMETER SELECTION

In practical applications, a video frame often needs to be transmitted with a target number of bits. Thus, for given target bitrate, we derive a set of parameters for the adaptive bilateral filter which generates output frames with the minimum visual distortion.

Suppose that the current frame is segmented into N subimages. Let $\omega_n = (\sigma_D(n), \sigma_R(n))$ denote the local parameters for the *n*th subimage, $r(\omega_n)$ be the rate and $d(\omega_n)$ be the distortion. The bilateral filter parameters for N blocks form a parameter vector $\vec{\Omega} = (\omega_1, \omega_2 ... \omega_N)$. Assuming that R0 target bits are assigned to the current frame, the optimal parameter control is to find the parameter vector which minimizes the overall distortion subject to the constraint that the total number of bits is equal to R0.

$$\vec{\Omega}^* = (\omega_1{}^*, \omega_2{}^* ... \omega_N{}^*) = \underset{\substack{\omega_1, \omega_2 ... \omega_N \\ \sum_{n=1}^{N} r(\omega_n) = R0}}{\arg\min} \sum_{n=1}^{N} d(\omega_n)\beta_n \tag{3.16}$$

Note that the total distortion is a weight summation of the distortions of all the subimages. The weight of the *n*th subimage β_n depends on the importance level of that subimage, which can be calculated as follows:

$$\beta_n = \sum_{\vec{x}_j \in \Phi_n} A(\vec{x}_j) \tag{3.17}$$

where $A(\vec{x}_j)$ is the attention value at pixel $\vec{x}_j$ and Φ_n denotes the pixels set of the *n*th subimage.

By introducing a Lagrange multiplier $\lambda \geq 0$, the constrained problem can be solved. An optimal parameter vector $\vec{\Omega}^*$ is obtained which minimizes the Lagrange cost function $L(\vec{\Omega}, \lambda)$.

$$\begin{aligned}\vec{\Omega}^*, \lambda^* &= \arg\min_{\vec{\Omega},\lambda} L(\vec{\Omega}, \lambda) \\ &= \arg\min_{\omega_1,\omega_2,\ldots\omega_N,\lambda} \{\sum_{n=1}^{N} d(\omega_n)\cdot\beta_n + \lambda\cdot(\sum_{n=1}^{N} r(\omega_n) - R_0)\}\end{aligned} \tag{3.18}$$

Although the solution can not be written in a close form function, this optimization problem can be solved numerically. In next chapter, we will focus on finding a suitable perceptual distortion metric. In Chapter 5.0 both the rate function and the distortion function will be modeled as polynomial functions via data fitting. We then solve the nonlinearly constrained problem by implementing the sequential quadratic programming (SQP).

4.0 DISTORTION METRICS AND MINIMIZATION

During preprocessing, the images/videos are subject to distortions, which may result in a degradation of visual quality. For the applications such as telesurgery, in which the images/videos are viewed by a human observer, the only "correct" method of quantifying visual distortion is through subjective evaluation. However, in order to optimize our preprocessing algorithm, we need to find suitable objective distortion metrics which can automatically predict perceived image/video quality. In this chapter, we compare the traditional distortion metrics and the structure-based distortion metric. Subjective quality assessment methods for system evaluation purpose will also be discussed.

4.1 INTRODUCTION TO DISTORTION METRICS

Distortion metrics can be classified according to the availability of an original image, with which the distorted image is to be compared. Most existing approaches are of the full-reference (FR) type [26-28, 47, 48], meaning that a complete reference image is assumed to be known. While in some situation, the reference image is unavailable or too expensive to be transported (e.g., in a transmission with a limited bandwidth), a no-reference (NR) [49-52] metric or a reduced-reference (RR) [53] metric is desirable. Usually, the FR metrics are more likely to provide good

visual quality prediction. In the application of telesurgery video preprocessing, we can apply FR metrics since the original images are fully accessible.

The traditional FR distortion metrics for images or video are based on merely pixel differences, e.g., mean absolute error (MAE), mean square error (MSE), peak signal-to-noise ratio (PSNR), which can be calculated as follows:

$$MAE = \frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y}\left|o(x,y)-d(x,y)\right| \tag{4.1}$$

$$MSE = \frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y}\left(o(x,y)-d(x,y)\right)^2 \tag{4.2}$$

$$PSNR = 10\log\frac{A^2}{MSE} \tag{4.3}$$

where $o(x,y)$ and $d(x,y)$, respectively denote the original and the processed image pixels at position (x,y); X and Y denote the image dimensions; and A represents the maximum greylevel of the image (A=255 for 8-bit representation). These metrics are appealing because they are simple to calculate, have clear physical meanings, and are mathematically convenient in the context of optimization. But they are not very correspondent to the perceived visual quality.

Since the human visual system (HVS) is the ultimate receiver of most visual signals after various processing (compression, restoration, enhancement, etc.), there has been substantial research effort to incorporate relevant perceptual characteristics for visual distortion/quality evaluation [26-28, 47-52]. The majority of the proposed perceptual distortion metrics have followed a strategy of modifying the MSE measure so that errors are penalized in accordance with their visibility. The underlying principle of the error-sensitivity approach is that perceptual quality is best estimated by quantifying the visibility of errors. This is essentially accomplished

by simulating the functional properties of early stages of the HVS, as characterized by both psychophysical and physiological experiments.

From the viewpoint of methodology, there are two major strategies of perceptual distortion metrics:

- **Bottom-up strategy** --- Perceptual models are built according to the understanding of the HVS and the psycho-visual experiments. The HVS characteristics being modeled typically include contrast sensitivity function (CSF), luminance adaptation, and various masking effect. These metrics are usually based on temporal/spatial/color decomposition with parameters calibrated by basic psychophysical data. Building an accurate bottom-up model is difficult because of the complex nature and limited understanding of the human eye and brain, and it is also computationally expensive for many applications.
- **Top-down strategy** --- Image data are analyzed for lumination/color difference and common visual artifacts. The HVS is treated as a black box, and only the input-output relationship is of the concern. The top-down approach is more effective and efficient if some prior knowledge about the dominant artifacts (e.g., blurring in our preprocessing algorithm) is incorporated [29, 30].

In the next section, we will focus one of the most recent and successful top-down approaches, the *structure similarity* approach.

4.2 STRUCTURE SIMILARITY BASED DISTORTION METRIC

The human visual system is highly adapted to extract structural information from the viewing field. Thus, a measure of structural information change can provide a good approximation to perceived image distortion. In [54], the structural information in an image is defined as those attributes that represent the structure of objects in the scene, independent of the average luminance and contrast.

The system separates the task of similarity measurement into three comparisons: luminance, contrast and structure.

(1) Luminance Comparison--- The mean intensity of the original and the distorted image can be calculated as follows:

$$\mu_o = \frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y} o(x,y)^2$$

$$\mu_d = \frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y} d(x,y)^2$$

The luminance comparison function $l(o(x,y),d(x,y))$ is a function of μ_o and μ_d, with a value range of [0,1]. It measures how close the mean luminance is between x and y. It equals 1 if and only if $\mu_o = \mu_d$.

$$l(o(x,y),d(x,y)) = \frac{2\mu_o\mu_d + C_1}{\mu_o^2 + \mu_d^2 + C_1} \tag{4.4}$$

where the constant C_1 is included to avoid instability when $\mu_o^2 + \mu_d^2$ is very close to zero.

$$C_1 = (K_1 L)^2$$

where L is the dynamic range of pixel value (255 for 8-bit grayscale images), and $K_1 << 1$ is a very small constant. Similar considerations also apply to the contrast comparison and structural comparison described later.

(2) Contrast Comparison --- The mean luminance is removed from the image and the standard deviation is used as the estimate of the image contrast. An unbiased estimate in discrete form is given by

$$\sigma_o = \left(\frac{1}{XY-1} \sum_{x=1}^{X} \sum_{y=1}^{Y} (o(x,y) - \mu_o)^2 \right)^{\frac{1}{2}}$$

$$\sigma_d = \left(\frac{1}{XY-1} \sum_{x=1}^{X} \sum_{y=1}^{Y} (d(x,y) - \mu_o)^2 \right)^{\frac{1}{2}}$$

The contrast comparison $c(o(x,y), d(x,y))$ is then the comparison of σ_o and σ_d.

$$c(o(x,y), d(x,y)) = \frac{2\sigma_o \sigma_d + C_2}{\sigma_o^2 + \sigma_d^2 + C_2} \tag{4.5}$$

It also ranges from 0 to 1, where the best value 1 is achieved if and only if $\sigma_o = \sigma_d$.

(3) Structure Comparison --- The image signal is normalized by its own standard deviation, so that the two images being compared have unit standard deviation. The structure comparison $s(o(x,y), d(x,y))$ is conducted after luminance subtraction and variance normalization. The correlation between two unit matrices $(o(x,y) - \mu_o)/\sigma_o$ and $(d(x,y) - \mu_d)/\sigma_d$ is an effective measure to quantify the structural similarity. Notice that the correlation between $(o(x,y) - \mu_o)/\sigma_o$ and $(d(x,y) - \mu_d)/\sigma_d$ is equivalent to the correlation coefficient between $o(x,y)$ and $d(x,y)$. The structure comparison function is defined as follows:

$$s(o(x,y),d(x,y)) = \frac{\sigma_{od} + C_3}{\sigma_o \sigma_d + C_3} \tag{4.6}$$

Finally, the three comparisons are combined into one similarity measure, which is called **Structural SIMilarity (SSIM)** index, between images $o(x,y)$ and $d(x,y)$.

$$\begin{aligned} SSIM(o(x,y),d(x,y)) = &\left|l(o(x,y),d(x,y))\right|^{\alpha} \cdot \left|c(o(x,y),d(x,y))\right|^{\beta} \\ &\cdot \left|s(o(x,y),d(x,y))\right|^{\gamma} \end{aligned} \tag{4.7}$$

where $\alpha > 0, \beta > 0$ and $\gamma > 0$ are parameters used to adjust the relative importance of the three components. Let $\alpha = \beta = \gamma = 1$ and $C_3 = C_2 / 2$, we have

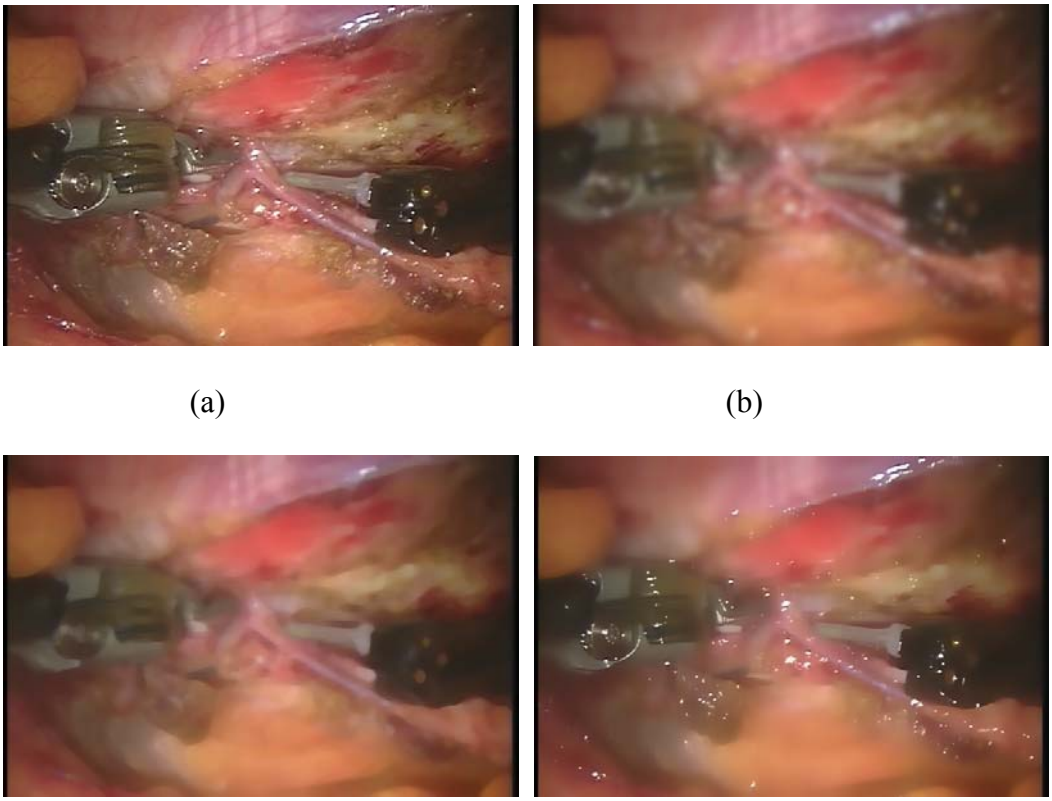
$$SSIM(o(x,y),d(x,y)) = \frac{(2\mu_o\mu_d + C_1)(2\sigma_{od} + C_2)}{(\mu_o^2 + \mu_d^2 + C_1)(\sigma_o^2 + \sigma_d^2 + C_2)} \tag{4.8}$$

Since image statistical features are usually highly spatially nonstationary, it is useful to apply the SSIM index locally rather than globally. The local statistics μ_x, σ_x and σ_{xy} are computed within a local 8x8 square window, which moves pixel by pixel over the entire image. In practice, one usually requires a single overall quality measure of the entire image. The mean SSIM (MSSIM) index is used to evaluate the overall image quality.

$$MSSIM(O,D) = \frac{1}{M}\sum_{j=1}^{M} SSIM(o_j(x,y), d_j(x,y)) \tag{4.9}$$

An example of how PSNR and MSSIM perform on evaluating image smoothing filters is shown in Figure 4.1. The original image is shown in Figure 4.1(a). We applied three different image smoothing filters to the original image and the output images are shown in Figure 4.1(b)-(d). The SSIM quality maps of these processed images are plotted in Figure 4.2(a)-(c), and the corresponding absolute error maps are shown in Figure 4.2(d)-(f) (contrast-inverted for easier comparison to the SSIM maps).

Finally, there PSNR and MSSIM values were calculated and listed in Table 4.1. Notice that the PSNR values for the three filtered images are very close, while the MSSIM values varies. The image processed with bilateral filtering has much higher MSSIM value than those filtered with Gaussian and Median filter.



(a) (b)

(c) (d)

Figure 4.1 (a) original image (b) processed image using Gaussian filter ($\sigma = 4$) (c) processed image using Median filter (w=13) (d) processed image using bilateral filter ($\sigma_D = 7, \sigma_R = 21$)

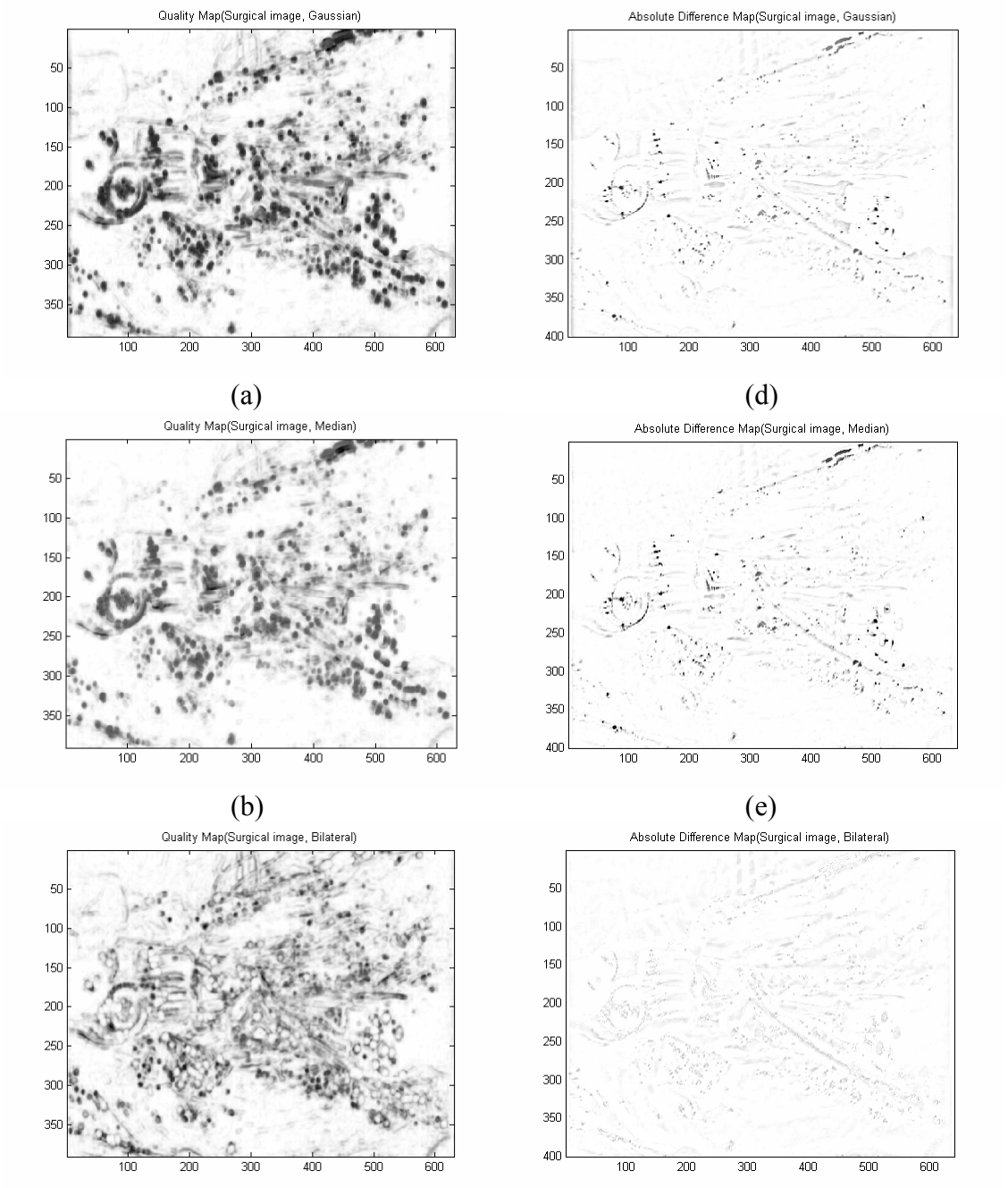


Figure 4.2 (a) SSIM map of processed image using Gaussian filter; (b) SSIM map of processed image using Median filter; (c) SSIM map of processed image using bilateral filter; (d) Absolute error map of processed image using Gaussian filter; (e) Absolute error map of processed image using Median filter; (f) Absolute error map of processed image using bilateral filter

Table 4.1 PSNR and MSSIM comparison for image smoothing filters

	PSNR(dB)	MSSIM
Gaussian	35.45	0.7953
Median	35.45	0.8010
Bilateral filter	35.49	0.8720

4.3 DISTORTION MIMINIZATION

In order to take into account the spatial variation of visual resolution according to the gazing point, we utilize the attention map as the foveal weighting metric and modify the MSSIM into the *foveated SSIM* index(we call it **FSSIM**)index.

$$FSSIM(O,D) = \frac{\sum_{j=1}^{M} f_j \cdot SSIM(o_j(x,y), d_j(x,y))}{\sum_{j=1}^{M} f_j} \tag{4.10}$$

where f_j is the foveal weighting metric, which is a function of the attention map.

We then substitute the MSE with the FSSIM into the nonlinearly constrained optimization problem, Eq. 3.16 now becomes:

$$\vec{\Omega}^* = (\omega_1^*, \omega_2^* \ldots \omega_N^*) = \underset{\substack{\omega_1, \omega_2 \ldots \omega_N \\ \sum_{n=1}^{N} r(\omega_n) = R0}}{\arg\max} \sum_{n=1}^{N} s(\omega_n)\beta_n \tag{4.11}$$

where $r(\omega_n)$ is the number of bits of the n[th] subimage, $s(\omega_n)$ is FSSIM between the original and the filtered subimage, β_n is the importance level of the n[th] subimage based on the attention map. As we mentioned before, an optimal parameter vector is obtained by minimizing the Lagrange cost function $L(\vec{\Omega},\lambda)$:

$$\begin{aligned}\vec{\Omega}^*,\lambda^* &= \arg\min_{\vec{\Omega},\lambda} L(\vec{\Omega},\lambda) \\ &= \arg\min_{\omega_1,\omega_2,\ldots\omega_N,\lambda} \{-\sum_{n=1}^{N} s(\omega_n)\cdot\beta_n + \lambda\cdot(\sum_{n=1}^{N} r(\omega_n) - R_0)\}\end{aligned} \tag{4.12}$$

4.4 SUBJECTIVE QUALITY ASSESSMENT

In order to evaluate our perception based video preprocessing method, subjective quality assessment is necessary. Subjective quality ratings also act as benchmarks for objective metrics. There are three commonly used procedures:

1. Double Stimulus Continuous Quality Scale (DSCQS): Viewers are shown multiple sequence pairs consisting of a "reference" and a "test" sequence, which are rather short (typically 10 seconds). The reference and test sequence are presented twice in alternating fashion, with the order of the two chosen randomly for each trial. Subjects are not informed which is the reference and which is the test sequence. They rate each of the two separately on a continuous quality scale ranging from "bad" to "excellent". Analysis is based on the difference in rating for each pair, which is often calculated from an equivalent numerical scale from 0 to 100.

2. Double Stimulus Impairment Scale (DSIS): As opposed to the DSCQS method, the reference is always shown before the test sequence, as shown in Figure 4.3, and neither is repeated. Subjects rate the amount of impairment in the test sequence on a discrete five-level scale ranging from "very annoying" to "imperceptible" as shown in Figure 4.4.
3. Single Stimulus Continuous Quality Evaluation (SSCQE): Instead of seeing separate short sequence pairs, viewers watch a program of typically 20-30 minutes duration which has been processed by the system under test; the reference is not shown. Using a slider, the subjects continuously rate the instantaneously perceived quality on the DSCQS scale from "bad" to "excellent".

The above-mentioned methods generally have different applications: DSCQS is the preferred methods when the quality of test and reference sequence are similar, because it is quite sensitive to small differences in quality. The DSIS method is better suited for evaluating clearly visible impairments such as artifacts caused by transmission errors, for example. Both DSCQS and DSIS method are not suited to the evaluation of such long sequences because of the recency phenomenon, a bias in the ratings toward the final 10-20 seconds due to limitations of human working memory. SSCQE relates well to the time-varying video quality of compressed video. However, SSCQE scores of different tests are harder to compare because of the lack of a reference. We apply the DSIS method to evaluation our proposed video preprocessing method, because (1) artifacts such as the blurriness cause by the filter and the blockness caused by the compression are visible. (2) experimental video clips are quite short.

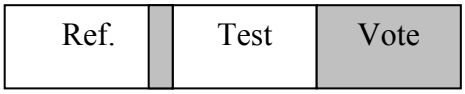


Figure 4.3 Presentation sequence of the DSIS method

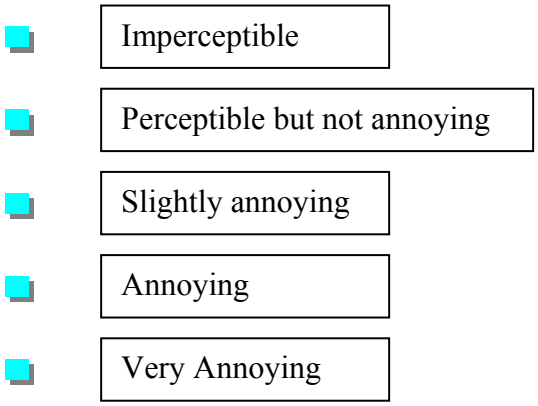


Figure 4.4 Rating scale of the DSIS method

5.0 IMPLEMENTATION AND EXPERIMENT

Our proposed preprocessing system consists of three major units: the preprocessing unit, an importance map generator, and a parameter selector, as shown inside the dashed square in Figure 5.1. The preprocessing unit utilizes the proposed adaptive bilateral filtering to remove unperceived small detail contents. The acuity map generator collects real-time eye gazing positions of the viewer, and output time-varying acuity map into the parameter selector, which can automatically select parameter sets for the adaptive bilateral filtering algorithm.

In this chapter, we first model the relationship between the bilateral filtering parameters and the distortion metrics and encoding bitrate. Then, we solve the nonlinearly constrained optimization problem utilizing the most successful method called sequential quadratic programming (SQP) [55-57]. Finally, experimental results are presented and implementation methods are discussed.

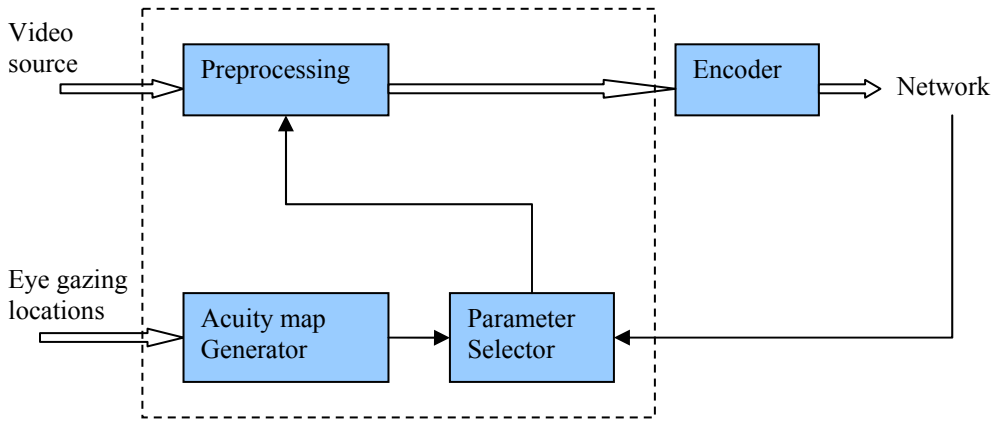
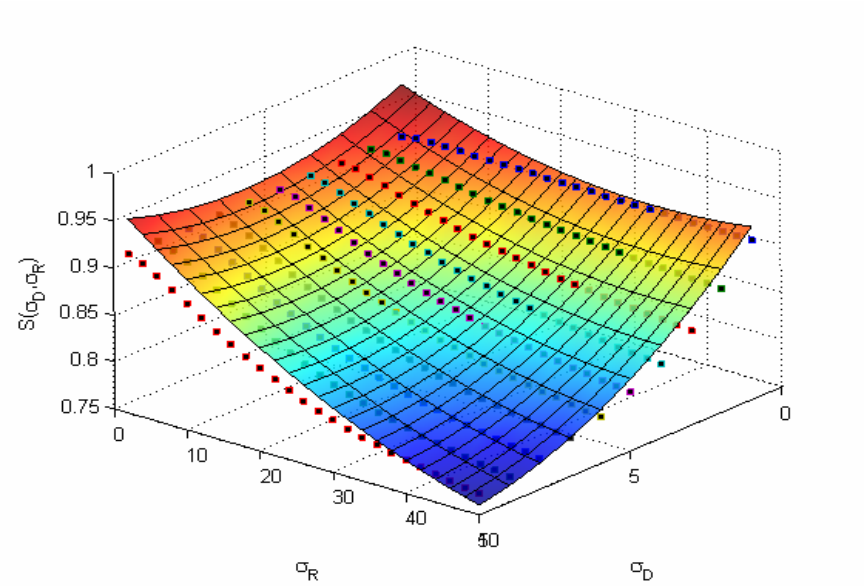


Figure 5.1 System design of the proposed video preprocessing system

5.1 DISTORTION AND BITRATE MODELING

In order to solve the optimization problem, we model the effect of the bilateral filtering parameters on the distortion and encoding bitrate. We used the same surgery video segment as in Chapter 3.0 for our experiments. Each video frame was decomposed into macroblocks of 16 x 16 pixels, then each macroblock was filtered repeatedly using different parameter pairs within the range($1 \le \sigma_D \le 10,\ 2 \le \sigma_R \le 50$). First, the averages of the mean structure similarity index (MSSIM) between the original and the filtered macroblock were calculated (plotted in Figure 5.2). The data were then fitted into a polynomial:

$$S(\sigma_D, \sigma_R) = a_1 + b_1\sigma_D + c_1\sigma_R + d_1{\sigma_D}^2 + e_1{\sigma_R}^2 + f_1\sigma_D\sigma_R \qquad (5.1)$$

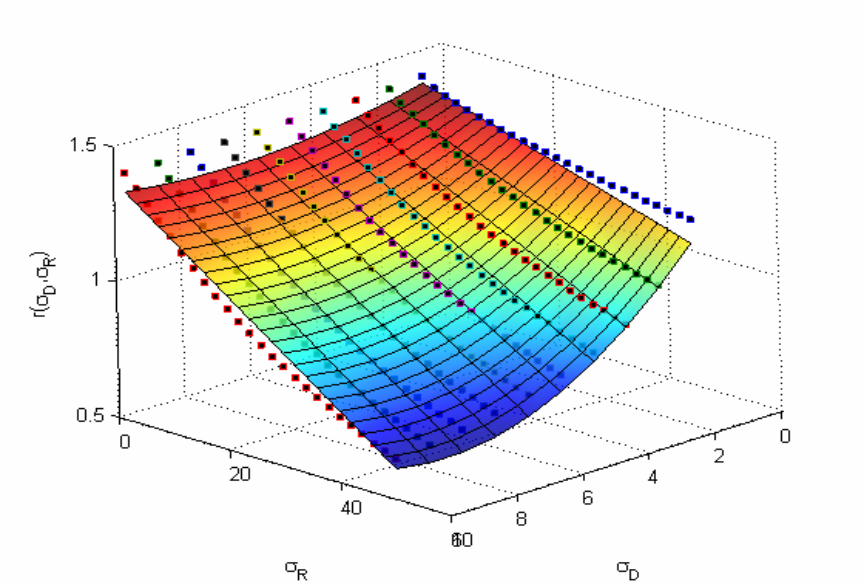


a1 = 9.9814695362319517E-01 b1 = -1.7965263636363589E-02

c1 = -2.7526344481610208E-03 d1 = 1.4979242424242689E-03

e1 = 4.2433946488290872E-05 f 1= -3.5271608391607991E-04

Figure 5.2 Approximation of the Distortion Model (SSIM). Scattered dots: calculated MSSIM values, Surface: approximated polynomial function



a2 = 1.5695128884058203E+00 b2 = -5.8141106060606337E-02

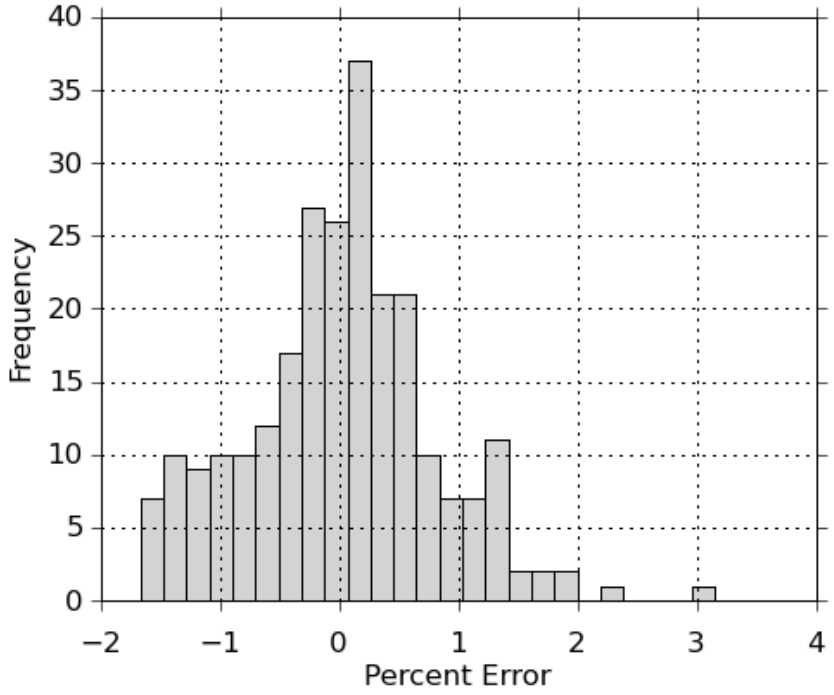
c2 = -1.4737299851357209E-02 d 2 = 5.1746212121212424E-03

e2= 1.5852568747677848E-04 f 2= -1.0780296037295827E-03

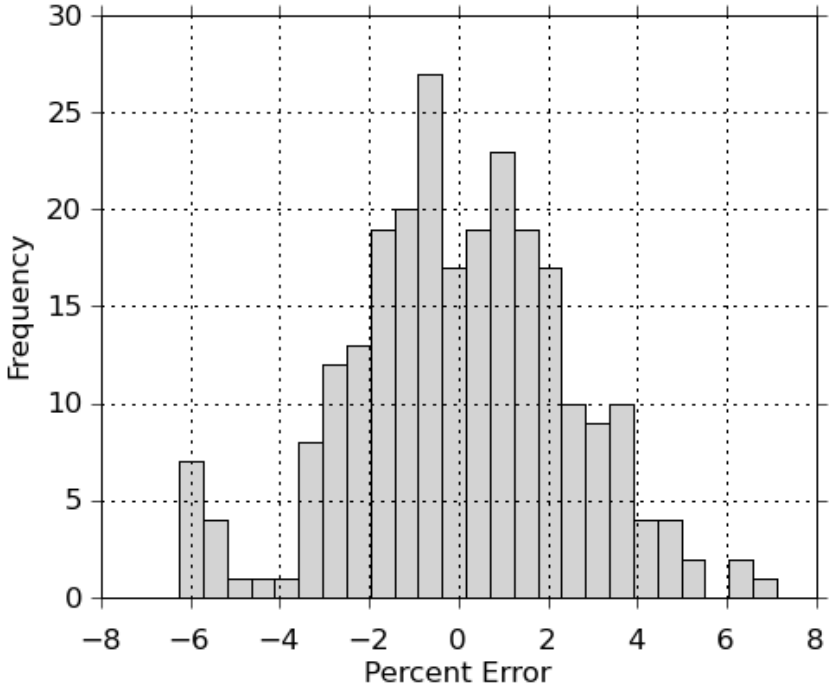
Figure 5.3 Approximation of the Entropy Model. Scattered dots: calculated entropy values, Surface: approximated polynomial function.

Suppose that the preprocessed frame is sent to a DCT-based encoder with uniform scalar quantizer of step size Q, the bitrate is bounded below by the empirical entropy of the Q-quantized coefficients, denoted as $H(Q)$. Let Q=6, we estimated the bitrate for preprocessed video frames with various parameter pairs. The results are plotted in Figure 5.3, where the scattered square dots denote the calculated empirical entropy values. The entropy data was again fitted into a polynomial:

$$R(\sigma_D,\sigma_R) = a_2 + b_2\sigma_D + c_2\sigma_R + d_2{\sigma_D}^2 + e_2{\sigma_R}^2 + f_2\sigma_D\sigma_R \qquad (5.2)$$



(a)



(b)

Figure 5.4 Data fitting percent error histograms of (a) $S(\sigma_D, \sigma_R)$ (b) $R(\sigma_D, \sigma_R)$

The histograms of the fitting errors for both $S(\sigma_D, \sigma_R)$ and $R(\sigma_D, \sigma_R)$ are plotted in Figure 5.4(a) and (b), respectively. From the histograms, we can see that most of the fitting errors are bounded between -5% and 5%.

5.2 IMPLEMENTATION OF THE SQP

In order to solve the nonlinear optimization problem, we adopt the sequential quadratic programming (SQP) approach. The SQP attempts to solve a nonlinear problem directly rather than convert it to a sequence of unconstrained minimization problems. The method is based on solving a series of subproblems designed to minimize a quadratic model of the objective subject to a linearization of the constraints. The basic idea is analogous to Newton's method for unconstrained minimization: At each step, a local model of the optimization problem is constructed and solved, yielding a step toward the solution of the original problem. The details of the SQP are described in Appendix.

When the number of subimages increases, the dimension of the parameter vector also increases, which may cause computational difficulties for the optimization algorithm. Instead of searching for the optimal solution, we utilized a new method which greatly speeds up the computational process. First, we divide image into two areas (inner and outer area). Second, run SQP to obtain one pair of parameters for each area. Then, subdivide the outer area into two areas, repeat step 2 until the total number of areas equal to N. Although the result is suboptimal, it is computational more efficient.

5.3 EXPERIMENT RESULTS

5.3.1 Automatic parameter selection

First, we set the number of areas N equals to 3, which means the SQP optimization algorithm calculate 3 pairs (a total of 6) parameters for the adaptive bilateral filter. The segmentation of the three areas is illustrated in Figure 5.5, with importance weights $\beta_1 = 1$, $\beta_2 = 0.6$, $\beta_3 = 0.4$. Given different target bitrate R0, the optimization results are listed in Table 5.1.

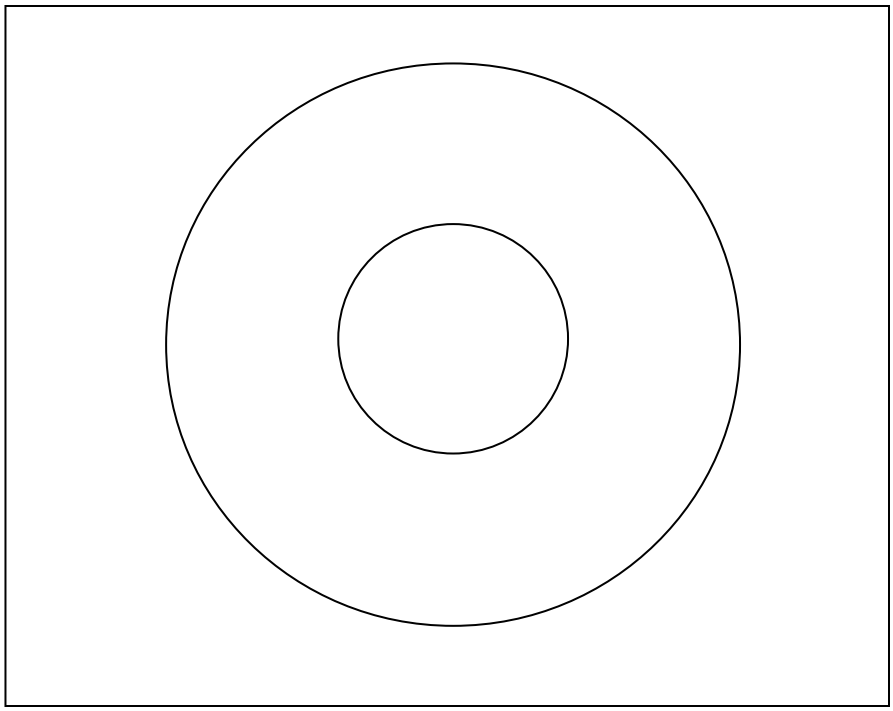


Figure 5.5 Illustration of area division when N=3 (Importance weights for each area is set to be $\beta_1 = 1 \quad \beta_2 = 0.6 \quad \beta_3 = 0.4$)

Table 5.1 Automatically selected parameter sets and FSSIM value (N=3).

R_0	σ_{D1}, σ_{R1}	σ_{D2}, σ_{R2}	σ_{D3}, σ_{R3}	FSSIM
1.2	(1.00, 2.51)	(1.00, 10.72)	(3.74, 22.80)	0.9221
1.0	(1.00, 5.93)	(2.25, 25.06)	(4.59, 40.47)	0.8826
0.8	(1.00, 8.01)	(4.98, 32.37)	(8.21, 49.82)	0.8295

Second, we set N=4 and repeat the experiment. The importance values for each area are set as $\beta_1 = 1$, $\beta_2 = 0.6$, $\beta_3 = 0.4$ and $\beta_4 = 0.2$. The automatically selected parameter sets are listed in Table 5.2. Comparing the results from both tables, we notice that for the same bitrates, the FSSIM value increase as the number of areas increases.

Table 5.2 Automatically selected parameter sets and FSSIM value (N=4).

R_0	σ_{D1}, σ_{R1}	σ_{D2}, σ_{R2}	σ_{D3}, σ_{R3}	σ_{D4}, σ_{R4}	FSSIM
1.2	(1.00, 2.28)	(1.00, 4.07)	(1.00, 13.32)	(4.25, 39.55)	0.9394
1.0	(1.00, 2.51)	(1.00, 19.00)	(3.40, 28.83)	(8.31, 39.92)	0.8947
0.8	(1.00, 8.83)	(4.93, 19.94)	(7.98, 41.94)	(10.00, 50.00)	0.8380

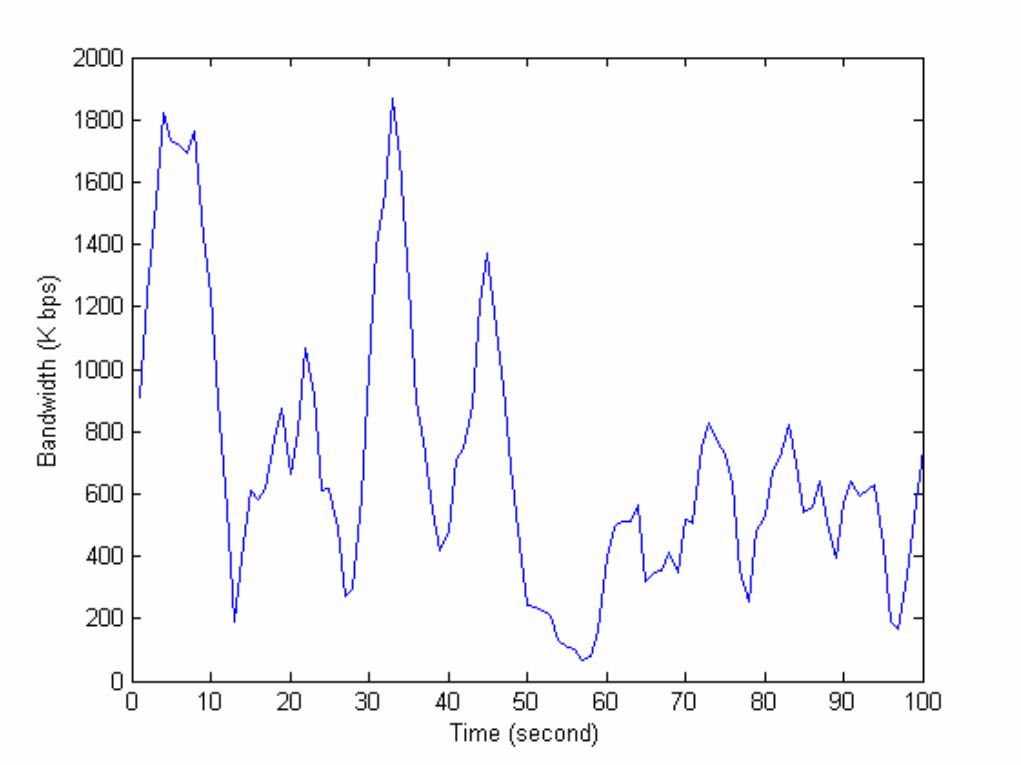


Figure 5.6 Recorded available bandwidth over 100 seconds

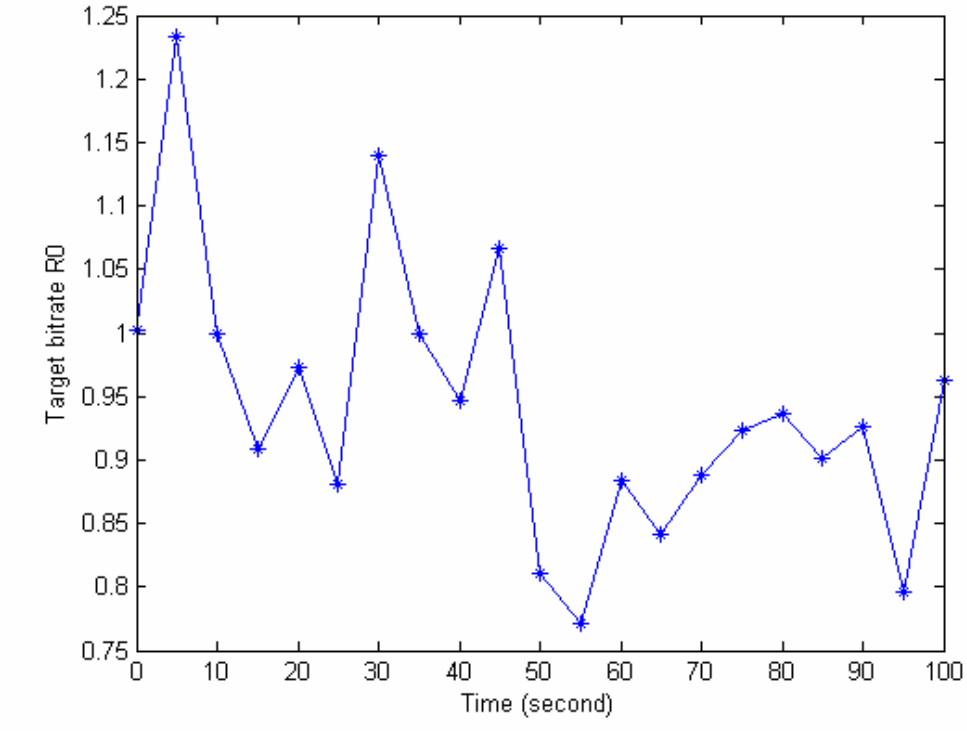
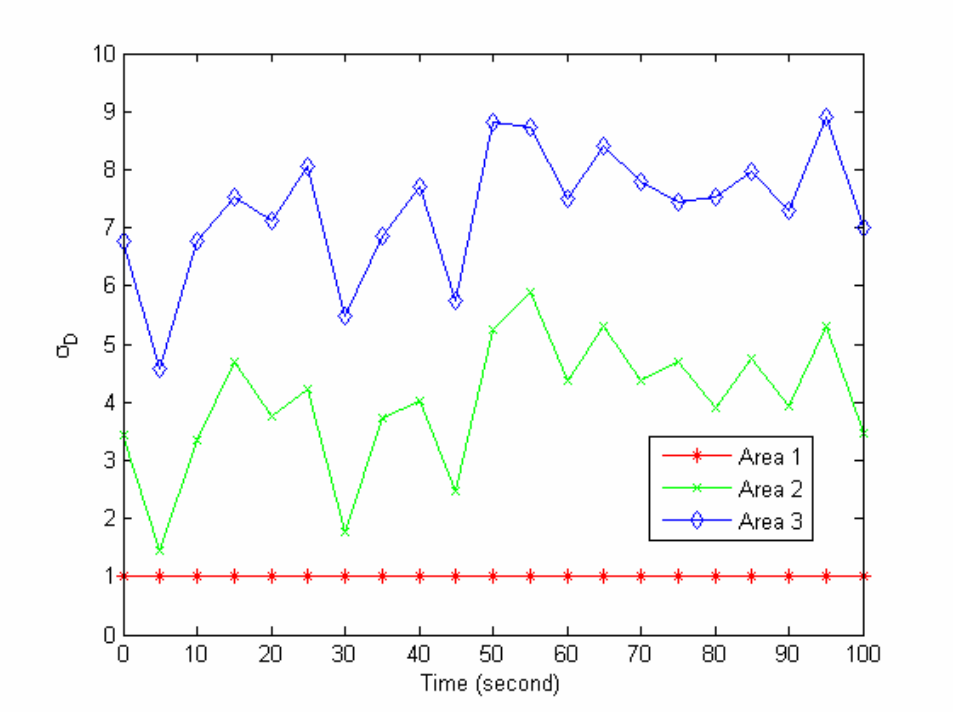
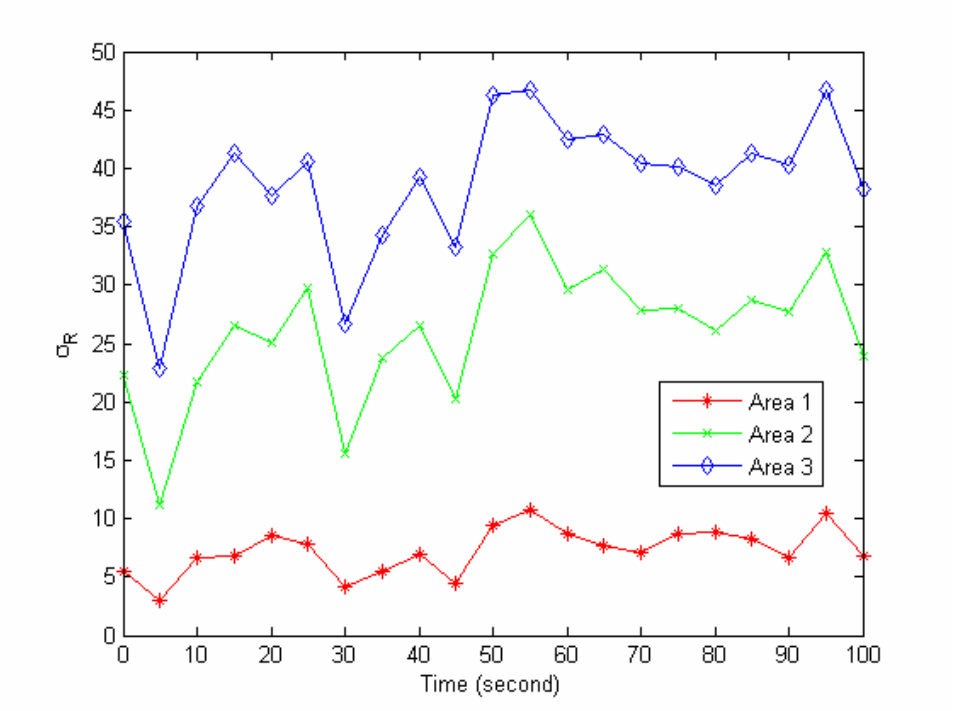


Figure 5.7 Target bitrates over 100 seconds



(a)



(b)

Figure 5.8 (a) Selected geometric spread σ_D ; (b) Selected photometric spread σ_R over time

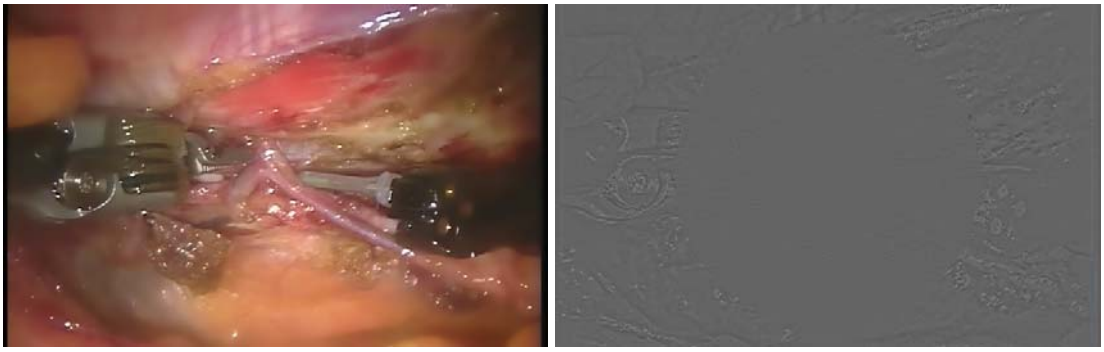
Since the network conditions of the Internet is dynamic, the parameters for the preprocessing algorithm must be updated as the network bandwidth fluctuates. We utilized the network monitoring software called **Bandwidth Monitor** to record the network traffic rates, and the recording interval was set to be 1 second. Figure 5.6 showed the recorded traffic data for the duration of 100 seconds. It has been shown that the Internet traffic condition was very unstable, and the available bandwidth varied from less than 100Kbps to about 1.8Mbps. We assume that the target bitrate R0 is a function of the available bandwidth, and should be updated every T seconds. Let T equals 5 seconds, the resulting target bitrate values R0(t) were plotted in Figure 5.7. We again divided each frame into three areas, and applied importance weights $\beta_1 = 1,\ \beta_2 = 0.6$, $\beta_3 = 0.4$, as illustrated in Figure 5.5. The optimal parameter sets for each target bitrate value were calculated, and the resulting geometric spread σ_D s were plotted in Figure 5.8(a), while the photometric spread σ_R s were plotted in Figure 5.8(b). From Figure 5.8, we notice that the value of σ_D and σ_R for the most important region remained relatively small over time, which means that the details within that region would be kept after preprocessing. On the other hand, the parameters for peripheral region fluctuate along with the target bitrate.

5.3.2 Telesurgery video preprocessing performance

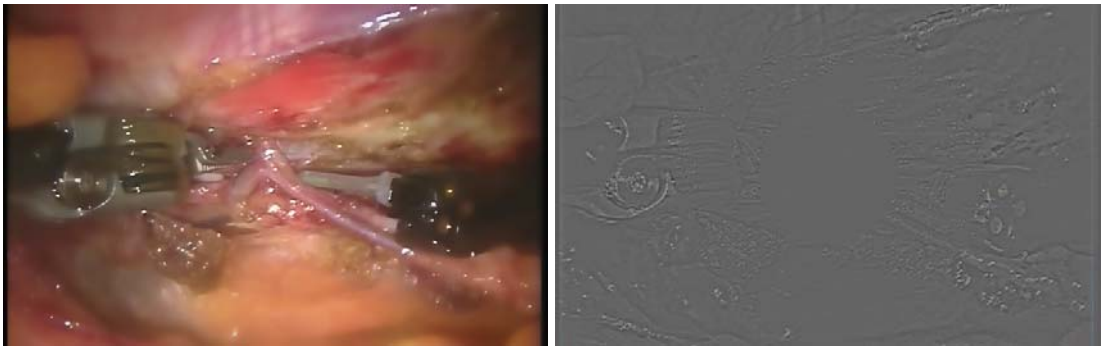
Our experiment videos are from a robotic prostatectomy using the *da VinCi* surgical system (provided by Intuitive Surgical, Inc), the video resolution is 640x400 pixels and frame rate is 30 fps. Each video clip contains 200 frames. We applied the automatically selected parameters (listed in Table 5.1) to our preprocessing algorithm, the output image frames of the adaptive

bilateral filtering algorithm are shown in Figure 5.9(a)-(c), and the difference frames are shown in Figure 5.9(d)-(f). Notice that when the target bitrate was relatively high, the bilateral filtering parameters chosen for all areas were small, thus the difference between original and preprocessed image frames were barely noticeable (as shown in Figure 5.9(d)). On the other hand, when the target bitrate was relatively low, the parameters chosen for the peripheral areas were much larger which result in heavy blurriness in the those areas, however the center area remained as sharp as before.

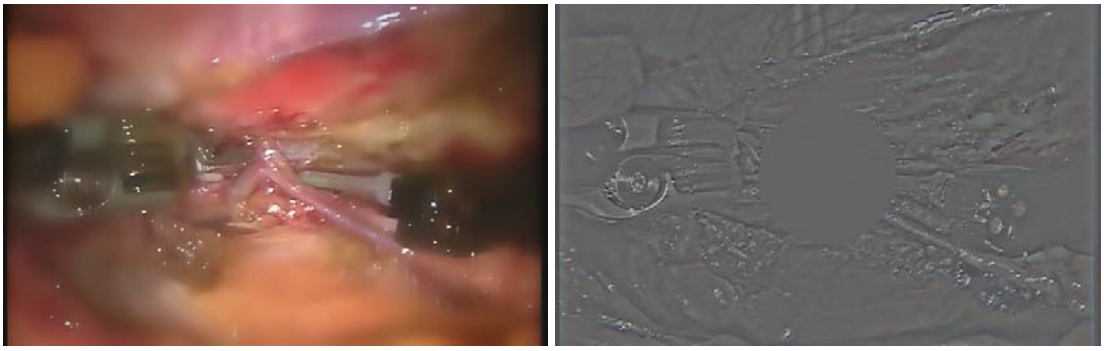
We then compressed all three preprocessed image sequences along with the original image sequence using the Microsoft Media 9 codec, which provides support for a wide range of bitrates for high-quality video streaming or downloading. We apply four different quality settings ranging from the Highest Quality (VBR100) to the Medium Quality (VBR75). The bitrates of the compressed videos are listed in Table 5.3. "Filter120" corresponds to the preprocessed video using parameters list in the first row of Table 5.1. "Filter100" and "Filter80" correspond to those preprocessed videos using parameters listed in second and the third row of that table, respectively. It can be seen that transmitting the highest quality telesurgery video without preprocessing will need a bandwidth of greater than 8M bps. Instead, we only need a bandwidth of around 5M bps to transmit the preprocessed video with almost same perceived quality. At around 2.77Mbps bandwidth, the preprocessed video "Filter120" was compressed using a higher quality setting (VBR 97), while the unfiltered video was compressed at a lower quality setting (VBR 95). It has been shown from these experimental results that our preprocessing algorithm can significantly improve the video encoding performance.



(a) (d)



(b) (e)



(c) (f)

Figure 5.9 (a)-(c) The resulting preprocessed frames for different target frame rates; (d)-(f) the difference between the original frame and preprocessed frames. (The parameters for adaptive bilateral filter is listed in Figure 5.1)

Table 5.3 Compression result in bitrates (kbps)

Quality	Original	Filter120	Filter100	Filter80
Highest(VBR100)	8281	5768	4602	3704
Higher(VBR97)	4294	2767	2095	1596
High(VBR95)	2779	1820	1403	1085
Medium(VBR75)	673	533	442	353

5.3.3 Subjective evaluation

In our research, we apply the Double Stimulus Impairment Scale (DSIS) to evaluation our proposed video preprocessing method. In the DSIS test, viewers are shown multiple sequence pairs consisting of a "reference" and a "test" sequence, which are rather short. Subjects rate the overall amount of impairment in the test sequence on a discrete five-level scale ranging from "imperceptible" to "very annoying" [31]. Subsequently, the Mean Opinion Scores (MOS) were computed. To facilitate numerical analysis and plots, the DSIS ratings are mapped onto a MOS scale from 1 to 5, where 1 indicates the worst quality ("very annoying"), and 5 the best ("imperceptible").

The evaluation was conducted among a group of 10 viewers. The monitor used in our evaluation is about 16 inches wide and 9 inches tall. The viewing distance was set to be around 16 inches. Each subject was given 10 pair of short video clips. Right after viewing each pair, he/she was asked to rate the test video clip (from "imperceptible" to "very annoying"). The MOS results were calculated and listed in Table 5.4. As we mentioned before, "Filter120", "Filter100" and "Filter80" denote the preprocessed video with parameter sets shown in Table 5.1. The MOS

for videos which were compressed using the highest quality setting (VBR100) were shown in the top row of Table 5.4, and those of compressed videos using medium quality setting (VBR75) were listed in the bottom row. The MOS score for the video "Filter120" compressed at highest quality is 4.8, which means the degradation by the preprocessing algorithm was imperceptible. Compare the bandwidth required transmitting this video and the original required bandwidth, we can see a reduction of around 30%. More aggressive preprocessing can provide more bandwidth reduction, however, the blurriness in the peripheral regions became somehow annoying to the viewers. For the test videos which were compressed at medium quality setting (VBR75), the MOS scores were significantly lower, as the viewers noticed some major artifacts such as the blocky effect produced by the video encoder.

Table 5.4 MOS value for compressed videos

	Filter120	Filter100	Filter80
VBR100	4.8	4.1	3.3
VBR75	3.8	3.5	2.6

We also compared the MOS for both non-filtered video and filtered video when compressed at lower bandwidths (500Kbps and 250Kbps). Compressed at 500 Kbps, the filtered video get a MOS score of 3.7, when the non-filtered video get a MOS score of 3.5. Compressed at 250 Kbps, the filtered video get a MOS score of 2.7, when the non-filtered video get a MOS score of 2.4. These results proved that our preprocessing algorithm can improve overall video perceived quality when the available network bandwidth is very low.

5.4 IMPLEMENTATION METHODS

For real-time applications such as telesurgery, we need to construct a hardware system to automatically control bit rate according to both extracted attention information and networking condition. A practical solution is to pre-compute the optimal parameter sets for multiple initial conditions (available bandwidth, surgeon's viewing distance, number of subimages), and store the values into a look-up table (LUT), which can be easily implemented by hardware. The savings in terms of processing time is significant, since retrieving a value from LUT is much faster than running the optimization algorithm. Before initializing the LUT, there are four steps:

(1) Initial attention map for given viewing distance and average network feedback delay is computed utilizing the method described in section 2.2.

(2) Given the number of subimages (i.e. N=4), the importance weights for each subimages are computed based on the initial attention map.

(3) Lower bound B_L and upper bound B_U of the available network bandwidth was statistically estimated from the network traffic records. Select L target bitrates R_i ($i = 1,...,L$) evenly from the range [B_L , B_U], where L is the length of the look-up table.

(4) Run the optimization algorithm for each target bitrates R_i ($i = 1,...,L$), then record the output parameter sets into the look-up table.

Once the LUT is load into memory, the appropriate parameters for the adaptive bilateral filtering algorithm can be found directly or interpolated from the table. The adaptive bilateral filtering algorithm can be implemented by a DSP chip, such as Texas Instruments TMS320C6X. First, the computational routines need to be optimized using a DSP simulation software package. Next, our routines will be compiled and loaded into the DSP board from a host computer. Finally,

the input, output, power supply, and other supporting components will be integrated and packed into a complete video preprocessing system. This system will be cascaded with a high-performance commercial hardware encoding engine, such as the advanced HaiVision video coding system (Montreal, Canada). The average time of encoding a single frame through the HaiVision video encoder is about 40 ms, close to the inter-frame interval of 33.3 ms (assuming a 30 frames/second video). We expect our preprocessing system to run faster because preprocessing is less computationally demanding than video compression. The total delay should be less than the time needed for encoding one frame.

6.0 CONCLUSION AND FUTURE WORK

6.1 SUMMARY OF CONTRIBUTIONS

In this thesis, we have presented a bio-inspired adaptive video preprocessing method for telesurgery video transmission based on the physiology of human visual system. We show by experimental results that the coding efficiency is improved significantly by the preprocessing method.

Firstly, we present a novel visual perception based video preprocessing method for telesurgery. Our preprocessing method utilized bilateral filtering is to remove unperceived details from the surgical video. The parameters of the bilateral filter are adjusted according to a human acuity model and the network traffic condition. By adding visual adaptation to the bilateral filter, we were able to remove higher degree of visual redundancy from the surgical video data. Secondly, we provide a natural way of reallocating bandwidth across the video frame. We have developed an automatic parameter selecting algorithm that can minimize the overall visual distortion when the network bandwidth is limited. Thirdly, we have shown by experiment that using the observer's attention map, approximately 50% bandwidth can be saved by preprocessing the raw video frames, while the preprocessed and the original video data are visually equivalent for surgical manipulation.

6.2 FUTURE WORK

Some future work of this research is suggested as follows. First, temporal filtering should be considered in addition to spatial filtering. Our preprocessing method only reduces the entropy of the intra-frames, while temporal filtering targets at entropy reduction for inter-frames. Temporal filtering can be carried out with or without motion-compensation.

Second, efforts should be made on developing a fast algorithm to reduce computational complexity. Our adaptive bilateral filtering adopted Gaussian kernel function. However, there exist a variety of kernel functions that could replace the Gaussian kernel. Separable kernels are desirable, because that would speed up the filtering process greatly, especially when the window size is large.

Third, the scientific basis and special characteristics of eye gazing during robotic surgery should be investigated. The feedback delay when transmitting video over the Internet should be studied statistically. The solutions we provided in section 2.2.3 need to be implemented and compared.

APPENDIX A

SOLVING NLP VIA SQP

A.1 NONLINEAR PROGRAMMING PROBLEM FORMULATION

A constrained nonlinear programming problem can be described as follows:

$$\textit{Minimize} \quad f(x), \quad x \in F \subseteq S \subseteq R^n \tag{1}$$

Subject to:

$$h_i(x) = 0, \quad i = 1,\ldots, p.$$
$$g_j(x) \le 0, \quad j = p+1,\ldots,q,$$

Given that $a_k \le x_k \le b_k$, $k = 1,\ldots,n$. $x = [x_1,\ldots,x_n]$ is a vector of n variables. $f(x)$ is the objective function, $h_i(x)$ ($i = 1,\ldots,p$) is the i[th] equality constraint, and $g_j(x)$ ($j = p+1,\ldots,q; q < n$) is the j[th] inequality constraint. S is the whole search space and F is the feasible search space. The a_k and b_k present the lower ad upper bounds of the variable x_k ($k = 1,\ldots,n$), respectively.

A.2 SEQUENTIAL QUADRATIC PROGRAMMING (SQP)

The Sequential Quadratic Programming (SQP) algorithm is a powerful technique for solving nonlinear constrained optimization problems [55]. SQP allows you to closely mimic Newton's method for constrained optimization just as is done for unconstrained optimization. At each iteration, an approximation is made of the Hessian of the Lagrangian function using a quasi-Newton updating method. The result is then used to generate a Quadratic Programming (QP) subproblem whose solution is used to form a search direction for a line search procedure [56, 57].

In studying the SQP we will adopt the full description of the methodology presented in [58]. SQP is an iterative method which solves at the k iteration a QP of the following form:

$$Minimize \quad \frac{1}{2} d^T H_k d + \nabla f(x_k)^T d, \tag{2}$$

Subject to:

$$\nabla h_i(x_k)^T d + h_i(x_k) = 0, \qquad i = 1, \ldots, p,$$
$$\nabla g_j(x_k)^T d + g_j(x_k) \le 0, \qquad j = p+1, \ldots, q,$$

where d is defined as the search direction and H_k is a positive definite approximation to the Hessian matrix of Lagrangian function of the problem. The Lagrangian function can be described as:

$$L(x, \gamma, \beta) = f(x) + \sum_{i=1}^{p} \gamma_i h_i(x) + \sum_{j=p+1}^{q} \beta_j g_j(x) \tag{3}$$

where γ and β are the Lagrangian multipliers. The developed quadratic subproblems can then be solved using the active set strategy. The solution x_k at each iteration is updated as follows:

$$x_{k+1} = x_k + \alpha_k d_k \tag{4}$$

where α is defined as the step size and takes values in the interval [0, 1]. After each iteration the matrix H_k is updated based on the Newton Method. One known methods to update the matrix H_k is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) methods [59-61]. Thus:

$$H_{k+1} = H_k + \frac{y_k y_k^t}{s_k y_k^t} - \frac{H_k s_k s_k^t H_k}{s_k^t H_k s_k} \tag{5}$$

where:

$$s_k = x_{k+1} - x_k$$
$$y_k = \nabla L(x_{k+1}, \gamma_{k+1}, \beta_{k+1}) - \nabla L(x_{k+1}, \gamma_k, \beta_k)$$

BIBLIOGRAPHY

[1] E. Flynn, “Telesurgery in the United States”, Homeland Defense J, p. 24-28. 2005

[2] J. Marescaux, J. Leroy, M. Gagner, F. Rubino, D. Mutter, M. Vix, S. E. Butner and M. K. Smith, "Transatlantic robot-assisted telesurgery"*, Nature*. Vol. 413, pp. 379, 2001.

[3] A. Rafiq, J. A. Moore, X. Zhao, C. R. Doarn, R. C. Merrell, "Digital video capture and synchronous consultation in open surgery", *Annal surgery*, Vol. 239, No. 4, pp. 567-573, 2003.

[4] J. C. Rosser Jr, R. L. Bell, B. Harnett, E. Rodas, M. Murayama, R. C. Merrell, "Use of mobile low-bandwidth telemedical techniques for extreme telemedicine applications", *J Am Coll Surg*, Vol. 189, No. 4, pp. 397-403, 1999.

[5] R. C. Merrell, B. E. Jarell, N. S. Schenkman, B. Schoener, "Telemedicine for the operating room of the future", *Semin Laparosc Surg,* Vol. 10, No. 2, pp. 91-94, 2003.

[6] J. M. Thompson, M. P.Ottensmeyer, T. B. Sheridan, "Human factors in telesurgery: effects of time delay and asynchrony in video and control feedback with local manipulative assistance", *J. Telemedicine,* Vol. 5, No. 2, pp. 129-137, 1999.

[7] S. Naegele-Jackson, P. Holleczek, T. Rabenstein, J. Maiss, E.G. Hahn, M. Sackmann, "Influence of compression and network impairments on the picture quality of video transmissions in tele-medicine", in *Proceedings of the 35th Hawaii International Conference on System Sciences*. 2002: Hawaii.

[8] G. K. Anastassopoulos and A. N. Skodras, "JPEG 2000 ROI coding in medical imaging applications", in *Visualization, Imaging, and Image Processing*, 2002: Marbella, Spain.

[9] T. Adiono, T. Isshiki, K. Ito, T. Ohtsuka, D. Li, C. Honsawek, H. Kunieda. "Face focus coding under H.263+ video coding standard." in *Circuits and Systems 2000 IEEE APCCAS 2000*.

[10] M. Chen, M. Chi, C. Hsu, and J. Chen, "ROI video coding based on H.263+ with robust skin-color detection technique", *IEEE Trans. on Consumer Electronics*, Vol. 49, No. 3, pp. 724-730, 2003.

[11] N. Doulamis, A. Doulamis, D. Kalogeras, S. Kollias, "Low bit-rate coding of image

[12] W. Lai, X. Gu, R. Wang, W. Ma, H. Zhang, "A content-based bit allocation model for video streaming", in *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*. 2004.

[13] W. Osberger and A. J. Maeder, "Automatic identification of perceptually important regions in an image using a model of human visual system", in *Proc. Int. Conf. Pattern Recognition*. 1998.

[14] X. Yang, K. Ramchandran, "A low-complexity region-based video coder using backward morphological motion field segmentation", *Image Processing, IEEE Transactions on,* Vol. 8, No. 3, pp. 332-345, 1999.

[15] F. Stentiford, "An estimator for visual attention through competitive novelty with application to image compression", in *Proc. Picture Coding Symp.* 2001.

[16] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention", *IEEE Trans. Image Processing,* Vol. 13, No.10, pp. 1304-1318, 2004.

[17] W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication", *Proc. SPIE,* Vol. 3299: pp. 294-305, 1998.

[18] S. Lee, A.C. Bovik, Y. Y. Kim, "Low delay foveated visual communications over wireless channels", *in Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on.* 1999.

[19] Philip Kortum and Wilson Geisler "Implementation of a foveated image coding system for image bandwidth reduction", in *SPIE Proceedings*. 1996.

[20] M. Ghanbari, "Video coding: an introduction to standard codes", *IEE Telecommunications Series. London: The Institution of Electrical Engineering,* pp.264, 1999.

[21] E. R. Davies, "Machine vision: theory, algorithms", *practicalities* 2nd ed. ed. San Diego: *Academic Press*, pp.750, 1997.

[22] P. Perona and J. Malik., "Scale-space and edge detection using anisotropic diffusion", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol.12, No.7, pp. 629-639, 1990.

[23] P. Saint-Marc, J.S. Chen, G. Medioni, "Adaptive smoothing: a general tool for early vision", *IEEE. Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, No. 6, pp. 514 -529, 1991.

[24] C. Tomasi, R. Manduchi, "Bilateral filtering for gray and color images", in *Proceedings of the 1998 IEEE International Conference on Computer Vision,* Bombay, India,1998.

[25] S. Winkler, "Issues in vision modeling for perceptual video quality Assessment", *Signal Processing,* Vol. 78, No.2, pp. 231–252, 1999.

[26] A. B. Watson, J. H., and J. F. McGowan III, "DVQ: a digital video quality metric based on human vision", *J. Electron. Imaging*, Vol.10, No.1, pp. 20-29, 2001.

[27] J. Lubin, "A visual discrimination model for imaging system design and evaluation", E. Peli, Ed. Singapore: *World Scientific. Vision Models for Target Detection and Recognition,* pp. 245-283, 1995.

[28] Z. Yu, H. R.Wu, S.Winkler, and T. Chen, "Vision-model-based impairment metric to evaluate blocking artifacts in digital video", *Proc. IEEE ICIP,*Vol. 90, No. 1, pp. 154-169, 2002.

[29] W. Lin, L. Dong and P. Xue, "Visual distortion gauge based on discrimination of noticeable contrast change", *IEEE Trans. on Circuits and Systems for Video Technology,* Vol. 15, No. 7, 2005.

[30] Z. Wang, A.C. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", *IEEE Trans. On Image Processing*, Vol. 13, No. 4, pp. 13, 2004.

[31] ITU-R Recommendation BT.500-11: Methodology for the subjective assessment of the quality of television pictures. 2002.

[32] X. Ran and N. Farvardin, "A perceptually motivated three component image model. Part I: Description of the model". *IEEE Trans. On Image Processing,* Vol. 4, No. 4, pp. 15, 1995.

[33] S. Daly, K. Matthews, and J. Ribas-Corbera, "Visual eccentricity models in face-based video compression", in *SPIE Human Vision and Electronic Imaging IV,* San Jose, CA,1999.

[34] Alex Poole, Linden J. Ball, "Eye tracking in human-computer interaction and usability research: current status and future prospects", in *Encyclopedia of Human Computer Interaction*, C. Ghaoui, Editor. Idea Group, 2005.

[35] K. R. Gegenfurtner, D. Xing, B. H. Scott, M. J. Hawken, "A comparison of pursuit eye movement and perceptual performance in speed discrimination", *Journal of Vision*, Vol. 3, No.11, pp. 865-876,2003.

[36] S. Amarnag, R.S. Kumaran, J.N. Gowdy, "Real time eye tracking for human computer interfaces", in *Multimedia and Expo, 2003. ICME '03. Proceedings. International Conference on*. 2003

[37] Lawrence H. Yu, Moshe Eizenman, "A new methodology for determining point-of-gaze in head-mounted eye tracking systems", *IEEE Trans. on Biomedical Engineering*, Vol. 51, No.10, pp. 1765-1773, 2004.

[38] Oleg V. Komogortsev, Javed I. Khan, "Eye movement prediction by Kalman filter with integrated linear horizontal oculomotor plant mechanical model", in *Proceedings of the 2008 symposium on Eye tracking research & applications.* ACM: Savannah, Georgia, 2008.

[39] B. Liu, M. Sun, Q. Liu, A. Kassam. C.C. Li and R.J. Sclabassi, "Automatic detection of region of interest based on object tracking in neurosurgical video", in *Proc. IEEE EMBS.* Shanghai, China, 2005.

[40] William R. Hendee and Peter N.T. Wells, "The perception of visual information", 2e. Springer-Verlag New York, Inc. pp.409, 1997.

[41] W. Heisenberg, "Physikalische prinzipien der quantentheorie (Leipzig: Hirzel)". (English translation: The physical principles of quantum theory), Chicago: University of Chicago Press, 1930.

[42] H. A. Mollot, "Computational vision: information processing in perception and visual behavior", *2000, Cambridge, MA: The MIT Press.* pp.296.

[43] M. J. Black, G. Sapiro, D. Marimont, and D. Heeger, "Robust anisotropic diffusion", *IEEE Transactions on Image Processing*, Vol. 7, No. 3, pp. 421-432, 1998.

[44] D. Barash, "A fundamental relationship between bilateral filtering, adaptive smoothing, and the nonlinear diffusion equation", *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 24, No. 8, pp. 844-847, 2002.

[45] M. Elad, "On the origin of the bilateral filter and ways to improve it". *IEEE Trans. On Image Processing*, Vol. 11, No. 10, pp. 1141-1151, 2002.

[46] N. Black, S. Moore and E. W. Weisstein, "Jacobi method", From *MathWorld*--A Wolfram Web Resource. Available from: http://mathworld.wolfram.com/JacobiMethod.html

[47] S. Winkler, "Visual fidelity and perceived quality: towards comprehensive metrics". in *Proc. SPIE*, 2001.

[48] S. Winkler, "A perceptual distortion metric for digital color video", in *Proc. SPIE Human Vision and Electronic Imaging IV*. 1999.

[49] H. R. Wu and M. Yuen, "A generalize block-edge impairment metric for video coding", *IEEE Signal Process. Lett.,* Vol. 4, No. 11, pp. 317-320,1997.

[50] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric", *Proc. IEEE ICIP*, Vol. 3, pp. 57-60, 2002.

[51] J. Caviedes and S. Gurbuz, "No-reference sharpness metric based on local edge kurtosis", *Proc IEEE ICIP*, Vol. 3, 2002, pp. 53–56.

[52] J. Dijk, et al, "A new sharpness measure based on Gaussian lines and edges", *Proc. Int. Conf. Computer Analysis on Images and Patterns (CAIP)*, 2756: pp. 149-156, 2003.

[53] S. Wolf, "Measuring the end-to-end performance of digital video systems", *IEEE Trans. Broadcast*, Vol. 43, No. 3, pp. 320-328, 1997.

[54] Z. Wang and A.C. Bovik, "Modern image quality assessment. Synthesis lectures on image, video, and multimedia processing", [*San Rafael, Calif.]: Morgan & Claypool Publishers,* 2006.

[55] M. C. Biggs, "Constrained minimization using recursive quadratic programming in towards global optimization", L.C.W Dixon, Editor. *1975: North-Holland*.pp. 341-349.

[56] S. P. Han, "A globally convergent method for nonlinear programming", *J. Optimization Theory and Applications*, Vol. 22, pp. 297, 1977.

[57] M.J.D. Powell, "A fast algorithm for nonlinearly constrained optimization calculations", in *Numerical Analysis*, G.A. Watson, Editor. Springer Verlag, 1978.

[58] M.J.D. Powell, "A Fortran subroutine for solving systems of nonlinear algebraic equations in numerical methods for nonlinear algebraic equations", P. Rabinowitz, Editor,1970.

[59] K. Schittkowski, "NLQPL: A FORTRAN-Subroutine Solving Constrained Nonlinear Programming Problems". *Annals of Operations Research*, Vol. 5: pp. 485-500, 1985.

[60] C. G. Broyden, "The convergence of a class of double-rank minimization algorithms", *J. Inst. Maths. Applics*, Vol. 6: pp. 76-90,1970.

[61] D. F. Shanno, "Conditioning of quasi-Newton methods for function minimization", *Mathematics of Computing*, Vol. 24: pp. 647-656,1970.