A Smart Video Coding Method for Time Lag Reduction in Telesurgery

Mingui Sun[1,2,3,4], Qiang Liu[1,4], Jian Xu[1], Amin Kassam[1], Sharon E. Enos[4]
Ronald Marchessault[5], Gary Gilbert[2,5], and Robert J. Sclabassi[1,2,3,4]

Laboratory of Computational Neuroscience
Departments of [1]Neurosurgery, [2]Electrical Engineering, [3]Bioengineering
University of Pittsburgh, Pittsburgh, PA 15260
[4]Computational Diagnostics, Inc., Pittsburgh, PA 15213
[5]Telemedicine and Advanced Technology Research Center (TATRC)
US Army Medical Research & Materiel Command (USAMRMC)
Fort Detric, Frederick MD 21702

Abstract

In the future war against terror and new types of offensive activities away from home, telemedical systems, including a telesurgical system, may become standard military medical equipment. In recent years, there have been significant technological advances in both telecommunications and robotics. These advances have made remotely operable telemedicine possible. However, a key technology that rapidly encodes, transmits, and decodes surgical video with the minimum round-trip delay and the least influence by network jitter (random fluctuation of delay) is not currently available. This paper presents a special-purpose video coding method to support telesurgery, telemonitoring, and teleconsultation, with special emphasis on telesurgery. Our method utilizes advanced image processing algorithms which prioritize the importance of the scene shown on the video screen. This prioritization is performed according to a gaze map constructed based on tracking the eye movements of the remote observer. During network congestion, our system processes video data more aggressively and transmits non-essential data at reduced data rates. As a result, the essential information necessary to perform surgery is protected against network deterioration.

1 Introduction

It has been a classic problem to provide severely injured soldiers with time-critical medical care, such as a sophisticated neurosurgery to stop an ongoing hemorrhage within the brain, within or near the battlefield. Recently, broadband telecommunication channels and computer networks have connected the theater of war to the Continental United States (CONUS) providing near instantaneous bi-directional communications. Also, robotic surgical systems (for example, the *da Vinci* Surgical System) are being developed and experience is being gained in their use. These technological advances in telecommunications make it possible to provide telemedicine services remotely and, when combined with surgical robots, to allow an expert surgeon in CONUS to operate on an injured soldier without being limited by geometrical distances and being exposed to the risk of injury in the battlefield.

Because of the high bandwidth requirement, video transmission plays the most critical role in the telesurgery information pathway. In the past twenty years, extensive research on video coding, which aims at data compression, has produced a number of widely utilized video coding standards, such as MPEG-2, MPEG-4, and H264[1, 2, 3, 4, 5]. These standards have been highly successful in the entertainment industry, leading to the development of technologies such as digital video disks (DVDs) and high definition video disks (HDVDs), and video broadcasting. However, interactive video applications, such as telesurgery, impose additional requirements on the video coding system, including low delay, high scalability in video streaming, and short encoding and decoding latencies. Unfortunately, these requirements are not fully supported by the existing standards beyond provisions of rudimentary rate distortion controls[6, 7].

Among the application-specific requirements, network delay and jitter have been identified to be the key problems

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 JAN 2008	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Smart Video Coding Method for Time Lag Reduction in Telesurgery	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Laboratory of Computational Neuroscience Departments of Neurosurgery, University of Pittsburgh, Pittsburgh, PA 15260**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002075., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**8**	

affecting the performance of telesurgery, leading to instability in the control of remote surgical robots[8, 9, 11]. Two major sources of delay have been observed in the transmission of surgical video[10]. The first source is the data processing time required for digitization, encoding, decoding, and rendering. This type of delay can be controlled using advanced computational architectures (e.g., pipeline or parallel processing) and dedicated hardware such as FPGA and DSP chips. In general, an increase in the computational complexity requires additional processing time. However, this increase in time does not necessarily cause an increase in delay because digital video rendering can be considered as a series of discrete events occurring at equally separated frame intervals. For example, all data processing tasks would produce the same 33.3 ms delay if the total amount of processing can be completed within 33.3 ms, assuming a 30 frames/sec frame rate. The second source of delay is due to the network. This delay can be subdivided into three categories[12]: *propagation*, *serialization*, and *queuing*. The *propagation delay* is simply due to the travel time of the electrical signal between Point A and Point B (e.g., 66.7 ms round-trip between Los Angeles and Beijing). Without a significant routing change, this delay is close to a constant, implying that the variation in this delay due to propagation is small. The *serialization delay* is the time spent on transmitting a single packet. It depends on the bandwidth of the network connection and the size of the packet. For example, a 10M bps Ethernet connection transmits a 188 byte packet in 0.15 ms (assuming 100% efficiency) while it takes a 56K modem (at bit rate 203K bps) 16.7 ms for the same packet (with 10 synchronization bits for each byte)[14]. For a fixed packet size and stable routing, this delay is near constant and the corresponding jitter is small. The last type of delay, *queuing delay*, is the time that a packet spends waiting within the router queues. This delay is determined by the network traffic. Without congestion this delay is negligible. With heavy congestion this delay becomes substantial. Therefore, the queuing delay is usually the most significant delay component in today's IP based network and the jitter resulting from the variation in this queuing delay is dominant. Both the queuing delay and jitter are key factors affecting the performance of telesurgery[10, 12, 15].

From the above analysis , it can be observed that the delay and jitter cannot be totally eliminated because a minimum time for data processing and transmission is required and the rates of processing and transmission are not always constant. There are two obvious solutions. First, a dedicated connection can bring the queuing delay close to zero if the bandwidth is high enough. Second, a quality-of-service (QoS) enabled connection, which assigns a higher priority to the telesurgical data, would substantially reduce delay and jitter. However, the availability of these options cannot be assumed in the theater of war, especially in the cases where an inter-continental network connection through many countries is required.

In this work, we present a novel approach to reduce both the serialization and queuing delays, as well as jitter, in a different perspective. Using an eye tracking technique, we detect the visual attention of the surgeon at the remote site and send the information to the operating site. Since the data size of the ROI information is only a small number of bytes and the information is usually predictable using the previously received data, the delay caused by this transmission is minimal or can be eliminated. In the next step, a region of interest (ROI), which reflects a real-valued gaze map indicating the importance of each pixel, is determined. The video data at the operating site are preprocessed to preserve the quality of data according the ROI and the network traffic condition. In this way, the required bandwidth is dynamically controlled adapting to the transmission channel while the critical information necessary to perform surgery is nearly invariant. Our experimental results indicate an over 50% reduction in bandwidth without affecting the essential visual fidelity in the field where surgery is performed. Using this context based video coding approach, the number of packets that must be transmitted through the network can be scaled down significantly, so is the likelihood of excessive queuing delay.

2 Methods

Traditionally, the quality of a video is measured by the accuracy of pixels in the entire image with respect to the physical scene. As a result, all pixels in the image are expected to be reconstructed as close as possible to their original values after the encoding, transmission, reception, and decoding cycle. This traditional approach requires large and stable bandwidth and this bandwidth cannot be reduced substantially by any video codec available today.

2.1 Basic Concepts

In telemedical applications, especially in telesurgery, the sole purpose of transmitting the video is for the surgeon to view and manually operate within a surgical landscape. If he/she chooses not to view a part of the video screen closely, high-quality transmission of that part becomes unnecessary. By understanding how a surgeon visually perceives the surgical field, one can present him/her preprocessed images with high-quality content only within the area where he/her is attending. Outside of this region, lower-quality content is provided. As a result, the surgeon collects the same amount of visual information in his/her

brain necessary to perform surgery, while the data can be transmitted more rapidly.

2.2 Eye Gazing

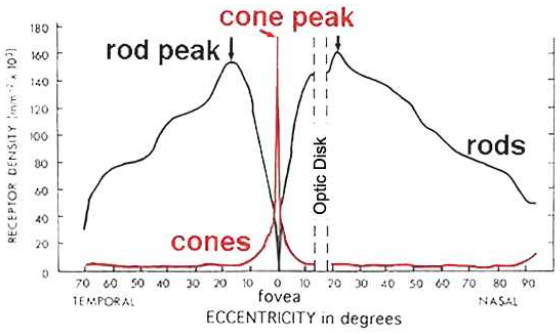

Figure 1: Eccentricity (in degrees) vs. preceptor density for cone and rod visual sensory cells (Osterberg, 1935).

"Gaze" was originally a psychological concept of visual culture describing how people look at each other in order to effectively gather interpersonal or social information. In the context of this technical paper, we use "eye gazing" to describe the action of attending to a portion of a visual scene for a certain amount of time. During "gazing", the eyes project a scene onto the retina. The gazing point is projected to its center (the fovea). Studies in visual science have shown that the image projected onto the retina is not perceived with a uniform resolution. This is because the cones and rods are not distributed uniformly. The density of the light-sensitive receptors is the highest at the fovea and decreases precipitously towards the outer rim, as shown in Fig. 1. The cones which are sensitive to color concentrate at the fovea. This center or vision covers less than 2 degrees of the visual field, compared with an entire visual field of 160 degrees in adult human. The rods, sensitive to luminance, have their largest density slightly outside the fovea and monotonically decreasing density in the outer boarder.

The resolution of the image projected on the retina is determined by the densities of the cone and rod cells. This resolution, however, is not the same as that of the image perceived by the visual cortex. The ganglion cells are also distributed densely in 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.

2.3 Spatial Acuity Modeling

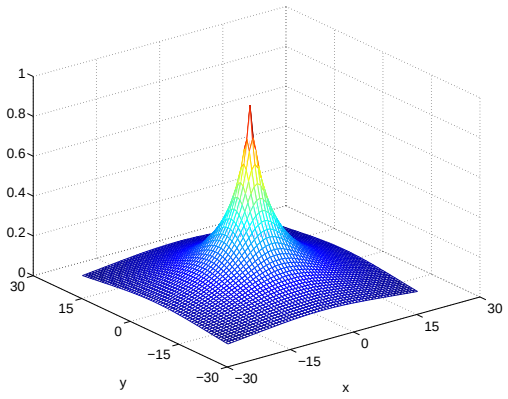

Figure 2: Spatial acuity model with parameter k=0.2

We utilize the basic concepts in visual science to construct the gaze map in the development of the novel adaptive video codec. Previous psychophysical tests suggest that visual acuity is nearly halved at one degree from the foveal center and decreased to one quarter at five degrees from the foveal center [16]. Several acuity models have been proposed based on empirical studies [17, 18]. Since these models have similar attenuating profiles, we utilize the following simple model[19]:

$$A(\bar{x}) = \frac{1}{1 + k\theta(\bar{x})} \tag{1}$$

where $A(\bar{x})$ and $\theta(\bar{x})$ are, respectively, the acuity and eccentricity angle at pixel location $\bar{x}$ (see Fig. 2), and k is a constant. The value of k can be obtained empirically according to video acquisition system utilized[20, 21].

2.4 Measurement of Eye Movements

The "gazing" map to be constructed is a time-varying function with a dynamically changing location of it's peak. This location must be determined by eye tracking. Although we have not construct an eye tracking system because of the availability of commercial systems (e.g., LC Technologies, Inc., Fairfax, Virginia), three commonly used methods are briefly described here. 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. This methods suffers from the same problem as the first method and the results seem to be unreliable. 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 movements from changes in reflections. This method appears to be promising in our case because it does not require any attachment to the eye or the skin.

2.5 Pre-Processing Algorithms

We have investigated a content-based video coding algorithm using adaptive bilateral pre-processing. When compared to alternative algorithms which combine data processing and coding procedures, such as the adaptive wavelet-MPEG coding[13], the bilateral pre-processing approach is attractive in our application because it provides freedom for the users to choose a video coding standard after pre-processing. This allows an easy cascade of our pre-processing module to any high-performance hardware encoding engine. It also allows switches among a number of encoding engines for different applications or networking environments. In the hardware implementation of the new video coding method, we suggest to use a cascade of the pre-processing and encoding modules, each running in a pipeline fashion. With optimized computational algorithms and dedicated hardware, we expect to complete the pre-processing task of each image within one frame interval. Because the current encoding module usually requires longer processing time than the pre-processing module, the delay caused by pre-processing is expected to be small and the gain in transmission speed due to reduced data size is expected to be much greater than the time lag due to processing delay.

2.6 Mathematics Formulation

The purpose of pre-processing is to remove the unnecessary contents outside the region of interest where small details are intentionally ignored by the surgeon. Our pre-processing will be adaptive according to: 1) the gazing map constructed by eye-tracking, and 2) the network condition fed back via the transmission channel.

The bilateral filter is a nonlinear image filter which can smooth image regions without blurring the edges. This is meaningful for our application because edges are more easily perceived during saccadic or quick voluntary eye movements. Therefore, strong edges should be best preserved after pre-processing. To achieve this goal, a bilateral filter is designed to perform a weighted averaging in a neighborhood centered at a reference pixel. Higher weights are assigned to the pixels that are closer in both space and intensity to the reference pixel. Mathematically, given an input image $\bar{I}(\bar{x})$, the output image $\bar{J}(\bar{x})$ is obtained by:

$$\bar{J}(\bar{x}) = \frac{\sum_{i=-S}^{S}\sum_{j=-S}^{S}\bar{I}(x_1+i, x_2+j)w(\bar{x},\bar{\xi})}{\sum_{i=-S}^{S}\sum_{j=-S}^{S}w(\bar{x},\bar{\xi})} \quad (2)$$

where $w(\bar{x},\bar{\xi}) = c(\bar{x},\bar{\xi})s(\bar{x},\bar{\xi})$ is a kernel function, $\bar{x} = (x_1, x_2)$ and $\bar{\xi} = (\xi_1, \xi_2)$ are, respectively, space variables of the current and reference pixels, and $\bar{I} = (I_1, I_2, I_3)$ is the intensity value of a color pixel. Note that the kernel function w in the convolution is the product of the functions c and s, which represent the "closeness" in the domain (space) and in the range (intensity), respectively. For simplicity, we utilize Gaussian functions to form the kernel function

$$w(\bar{x},\bar{\xi}) = exp\left(\frac{-||\bar{\xi}-\bar{x}||^2}{2\sigma_D^2}\right) exp\left(\frac{-||\bar{I}(\bar{\xi})-\bar{I}(\bar{x})||^2}{2\sigma_R^2}\right) \quad (3)$$

which is controlled by two parameters, σ_D and σ_R . The former, called geometric spread, is determined by the desired amount of low-pass filtering. Given the acuity map, this amount is defined by the acuity of the centered pixel in the neighborhood. A larger σ_D results in a more blurring effect as more neighboring pixels are involved for the averaging. The latter, called photometric spread, is chosen to compensate the blurring effect on the pixels representing edges. Pixels that resembles the test pixel in intensity have heavier weights in the averaging. The smaller the σ_R, the more restrictive this constraint is. In case where the test pixel is on an edge, the discontinuity of the intensity in the neighborhood confines the averaging to the neighboring pixels that are also on the edge, therefore preventing the excessive blurring.

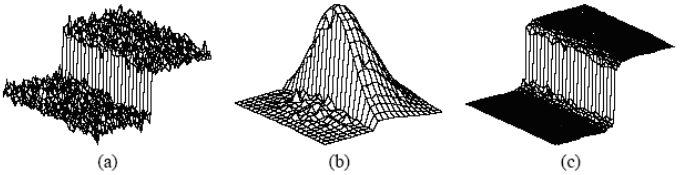

Figure 3: (a) An edge (intensity step 100) perturbed by Gaussian noise with standard deviation equal to 10 in intensity. (b) Combined similarity weights for a 23x23 neighborhood centered at a pixel slightly (two pixels) to the right of the step. (c) The edge in (a) after bilateral filtering with σ_R = 50 (in intensity) and σ_D = 5 (in pixels) (courtesy of [22]).

An example shown in Fig. 3 explains this concept. Consider now a sharp boundary between a dark and a bright region, as in panel (a). When the bilateral filter is centered, say, at a pixel on the bright side of the boundary, the similarity function s assumes values close to one for pixels on the same side, and values close to zero for pixels on the dark side. The convolution mask is shown in (b) for a

23x23 filter support centered at a pixel slightly away (two pixels to the right) from the edge in (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 at a dark pixel, the bright pixels are ignored instead. Thus, as shown in (c), noise was removed while the crisp edge was preserved.

2.7 Adaptive bilateral filtering

In video processing, the geometric spread σ_D is utilized to adapt to the acuity map. At any location closer to the gazing point, a smaller 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. To be specific, both spreads are functions of the acuity map: $\sigma_D(\bar{x}) = \sigma_{DU} - (\sigma_{DU} - \sigma_{DL})A(\bar{x})$ and $\sigma_R(\bar{x}) = \sigma_{RU} - (\sigma_{RU} - \sigma_{RL})A(\bar{x})$, where σ_{DU} and σ_{DL} are the pre-selected upper and lower limit of the geometric spread and σ_{RU}, and σ_{RL} are those of the photometric spread, respectively.

Besides adapting to the acuity map, we also dynamically reconfigure the bilateral filter adaptively to the traffic condition of the network. In practice, the lag time in the feedback information from the transmission channel can serve as an indicator of the most recent network status to which the output bitrate of the codec must adapt. Clearly, this bitrate can be gracefully controlled by properly reconfiguring the bilateral filter which controls the values of σ_D and σ_R. Since this adaptation is built on top of the acuity map, we define the values of the lower and upper limits, σ_{DL}/σ_{DU} and σ_{RL}/σ_{RU}, such that σ_D and σ_R are adjusted accordingly. Specifically, if we denote the current network capacity by N_c, the combined adaptation to both the network condition and the acuity map will be $\sigma_D(\bar{x}) = \sigma_{DU}(N_c) - (\sigma_{DU}(N_c) - \sigma_{DL}(N_c))A(\bar{x})$ and $\sigma_R(\bar{x}) = \sigma_{RU}(N_c) - (\sigma_{RU}(N_c) - \sigma_{RL}(N_c))A(\bar{x})$, where $\sigma.(N_c)$ denotes that the lower/upper limits are functions of the network capacity.

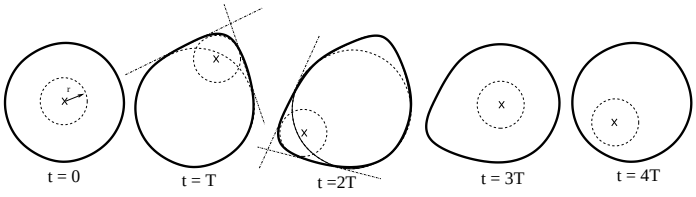

Figure 4: ROI shape changes with respect to a constant update interval T

3 Experimental Results

We have performed experiments to evaluate the video coding performance using the described context-based approach.

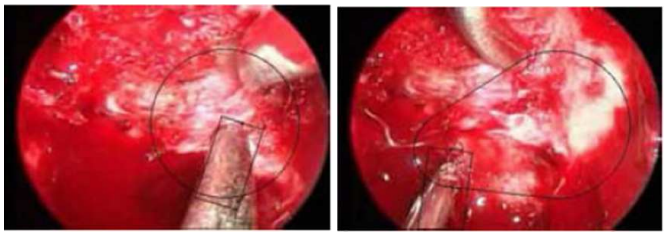

Figure 5: Results of ROI tracking. LEFT: A circular ROI is centered at the tip of an active surgical instrument (a suction tool). RIGHT: The ROI after 152 ms when the suction tool moves to the lower left. The shape of the ROI is progressively changed to allow observation of both the previous and current locations along with the trajectory of tool motion.

3.1 Gaze Map and Its Updates

We assume that a gaze map is in the basic form of Eq. 1 reflecting the visual attention of the operating surgeon. During surgery, the gaze map must be updated constantly, not only by the detected point of fixation, but also by the network condition as discussed previously. The process of the gaze map updates must be sufficiently smooth because abrupt updates distract the surgeon's attention and affect visual quality. We utilized a smooth updating method which is graphically explained in Fig. 4. The ROI is updated at equal time intervals T. We assume that, at $t = 0$, the ROI is in a circular shape. Inside the ROI, the cross represents the current focal point of attention. The dashed circle centered at the cross with radius r is a predefined effective range to be used in ROI updates. At $t = 0$, no update is necessary since the effective range is entirely inside the circular ROI (first panel). At $t = T$, suppose that the tip location changes and a portion of the effective range moves outside of the existing ROI. Then, we define a new ROI as an enclosed region defined by the existing ROI, the effective range, and the two tangential lines (second panel). At $t = 2T$, the ROI is similarly updated. However, we do not delete the previous update in favor of a gradual morphing of the shape (third panel). At $t = 3T$, the effective range moves back to the position inside the default ROI, we remove the update at $t = T$ (fourth panel). At $t = 4T$, the effective range is still within the default ROI, the update of $t = 2T$ is now removed. Fig. 5 shows the result of detected ROI with updating, where the video data was

obtained during a neurosurgery performed at the University of Pittsburgh Medical Center. It can be observed that the updated ROIs are cornerless and smooth. As the surgeon moves instruments within the surgical field, the gradually changing ROI with the memory effect just described allows a natural visualization of the surgical landscape. At the same time, the nonessential details outside the ROI is gracefully degraded.

3.2 Pre-Processing

In order to evaluate the performance of the content-based video coding technique, we utilized a video segment provided by Intuitive Surgical, Inc. This video segment has a screen size of 720×576 pixels and contains 200 frames. One of the original video frames is shown in the second panel of Fig. 6 where a scene of robotic surgery by using the *da VinCi* surgical system can be observed. We utilized a popular MPEG-4 video codec (Windows Media Series 9) with the quality parameter value set to 90 to compress the original video frames. The measured average bitrate for the 200 frames was 2707.08 Kbps.

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

Table 1: Pre-defined Parameters for Regions of Different Importance

Because the use of experimental eye tracker during surgery could affect surgical outcomes, we utilized simulated eye gazing maps in our experiments based on the theoretical model presented previously. The top panel in Fig. 6 shows one of our simulations where the fixation point of the eye is located at the center of the screen. The circle indicates the ROI within which the quality of the video is to be preserved. We implemented the adaptive bilateral filtering algorithm which adjusted the geometric spread parameter to control the degree of smoothing according to the gaze map, and the photometric spread parameter to allowing more (less) smoothing in the direction along (across) edges in the video image. In the experiment, we predefined three sets of parameters for regions with different importance levels, denoted as ROI, transit, and non-ROI, respectively, as shown in Table 1. The pre-processed 200-frame video segment was compressed, again using Windows Media Series 9 video codec with the same quality parameter of 90. The average bitrate for the pre-processed video seg-

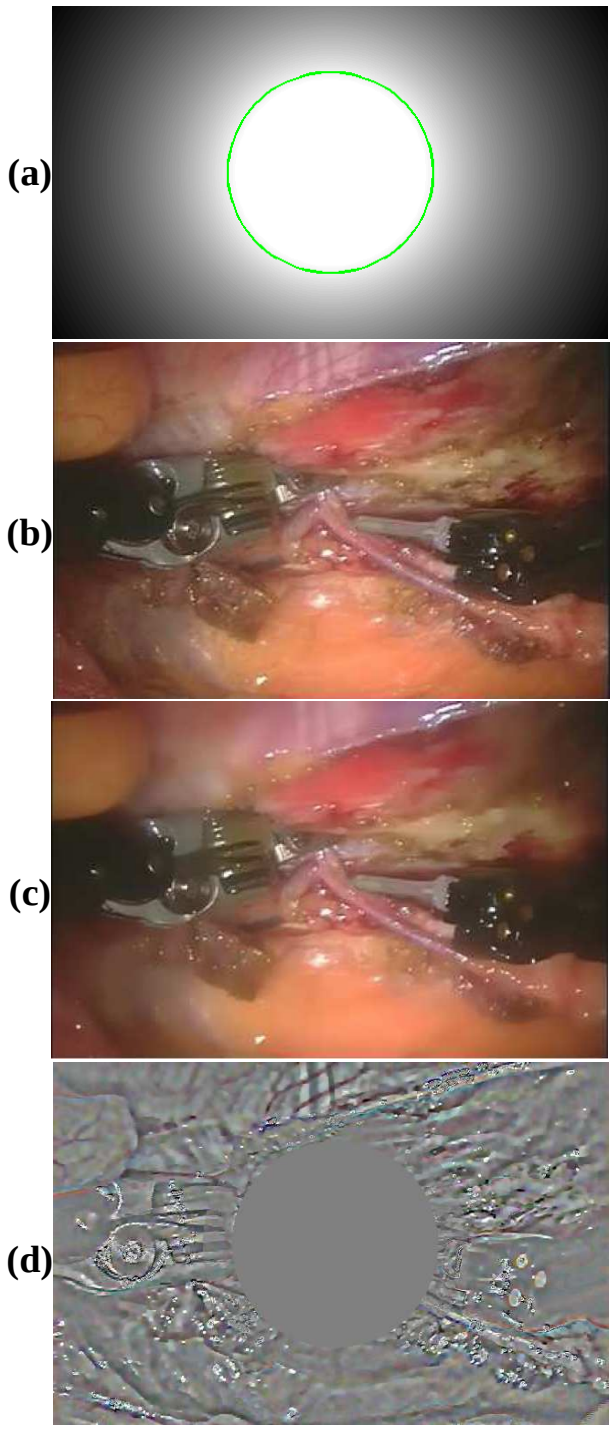

Figure 6: (a) Gaze map representing the visual acuity of the surgeon; (b) One of 200 original video frames (720x576 screen size) during a robotic prostatectomy using the da VinCi surgical system. The average bitrate is 2707.08 Kbps. (c) The same 200 frame after pre-processing by bilateral filter based on the gaze map. The average bitrate is reduced by over 50% (1372.83 Kbps). (d) Difference between (b) and (c). The error within the ROI is nearly zero.

ment was 1372.83 Kbps, a more than 50% reduction compared to the un-processed segment. Although the bitrate reduction is significant, the qualities of the original and pre-processed images (the second and third panels in Fig. 6) are visually indistinguishable when the viewer fixates at the center of each image. The difference between the original and filtered images is shown in the bottom of Fig. 6 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. It is rather surprising that, without pre-processing, these redundant "data" would have taken more than half of the bandwidth!

4 Conclusion

We have presented a content-based, special-purpose video coding method to support low-latency video data transmission in a number of military telemedical applications, especially in robotic telesurgery. Our method detects visual attention by tracking the eye movements of the remote surgeon. A gaze map is constructed in which each value indicates the importance of the corresponding pixel. Then, the original video frames are pre-processed based on the gaze map. A adaptive bilateral filter has been developed which removes unnecessary details in non-important regions of the video screen according to both the visual attention information in the gaze map and the network traffic condition. When the network performance deteriorates, this adaptive filtering is conducted more aggressively in the regions where the surgeon pays less attention during surgical manipulation. Our experimental results show that, using this attention based video coding approach, more than 50% reduction in datarate can be achieved while the quality of the video data necessary to perform a telesurgery is nearly invariant.

Acknowledgments

This work was supported in part by National Institutes of Health grants No. NS38494 and EB002309; Telemedicine and Advanced Technology Research Center (TATRC), US Army Medical Research & Materiel Command (USAMRMC); and Computational Diagnostics, Inc.

Opinions, interpretations, conclusions, and recommendations expressed in this work are those of the authors and are not necessarily endorsed by the U.S. Army and other funding agencies.

References

[1] L. Chiariglione, "Impact of MPEG standards on multimedia industry," *Proceedings of the IEEE,* 86(6):1222-1227, June 1998.

[2] Thomas Sikora, "MPEG digital video-coding standards," *Signal Processing Magazine,IEEE,* 14(5):82-100 Sept. 1997.

[3] Thomas Sikora, "The MPEG-4 Video Standard Verification Model," *IEEE Trans. on Cir. and Sys for video Tech.,* 7(1):19-31, Freb. 1997.

[4] S. Battista, F. Casalino and C. Lande, "MPEG-4: a multimedia standard for the third millennium," *Multimedia, IEEE,* 6(4):74-83, 1999.

[5] "Special Issue on H264 Video Coding Standard," *IEEE Trans. Circuits Syst. Video Technol.,* 13(7), Jul. 2003.

[6] P.Lambert, W. De Neve, P. De Neve, I. Moerman, P. Demeester and R. Van de Walle, "Rate-distortion performance of H.264/AVC compared to state-of-the-art video codecs," *IEEE Trans. on Cir. and Sys. for Video Tech.,* 16(1):134-140, Jan. 2006.

[7] H. Lee, T. Chiang and Y. Zhang, "Scalable rate control for MPEG-4 video," *IEEE Trans on Cir. and Sys. for Video Tech.,* 10(6):878-894, Sept. 2000.

[8] R.Steinmetz, "Human perception of jitter and media synchronization," *IEEE Journal on Selected Areas in Communications,* 14(1):61-72, Jan. 1996.

[9] M. D. Fabrizio et al., "Effect of time delay on surgical performance during telesurgical manipulation," *J. Endourology,* 14(2), Mar. 2000.

[10] J.M.Thompson, M.P.Ottensmeyer and T.B.Sheridan, "Huma Factors in Telesurgery: Effects of Time Delay and Asynchrony in Video and Control Feedback with Local Manipulative Assistance," *J Telemedicine,* 5(2):129-137, 1999.

[11] G. Darren, R. E. Kraut and S. R. Fussell, "The Impact of Delayed Visual Feedback on Collaborative Performance," *CHI 2006,* Montral, Qubec, Canada, April 2006.

[12] R. L. Cruz, "A Calculus for Network Delay, Part II: Network Analysis," *IEEE Trans. on Information Theory,* 37(1):132-141, January 1991.

[13] J. Xu, R.J. Sclabassi, Q. Liu, L. Chaparro and M. Sun, "A Content-Based Video Coding Method for Remote Monitgoring of Neurosurgery," *Proc. IEEE Multimedia Signal Processing,* Shanghai, China, Oct., 2005.

[14] "Understanding Delay in Packet Voice Networks," *Cisco white paper,* document ID 5125.

[15] "Implementing QoS Solutions for H.323 Video Conferencing over IP," *Cisco white paper,* Document ID: 21662.

[16] W. O. Clyde, *The Human Eye: Structure and Function,*Sinauer Associates, Inc., Sunderland, Massachusetts, pp.664, 1999.
[17] H. Liou and N. A. Brennan, “Anatomically accurate, finite model eye for optical modeling,” *J. Opt. Soc. Am. A*, 14(8), August 1997.
[18] D. D. Garcia, B. A. Barsky and S. A. Klein, “CWhatUC: a visual acuity simulator,” *Proc. SPIE*, 3246:290-298, June 1998.
[19] Daly, J.Scott, Matthews E.Kristine, Ribas-Corbera and Jordi,“Visual eccentricity models in face-based video compressio,” *SPIE in Human Vision and Electronic Imaging IV,* 3644:152-166, 1999.
[20] V. Virsu and J. Rovamo, “Visual resolution, contrast sensitivity, and the cortical magnification factor.” *in Experimental Brain Research*, V. 37, 1979.
[21] A. Johnston, “Patial scaling of central and peripheral contrast sensitivity functions,” *JOSAA* V.4(8), 1987.
[22] C. Tomasi and R. Manduchi,“Bilateral Filtering for Gray and Color Images,” *Proc. of the 1998 IEEE International Conference on Computer Vision*, Bombay, India, 1998.