

1 Introduction

A mobile ultrasound system has been developed, which makes ultrasound examinations possible in harsh environments without reliable power sources, such as ambulances, helicopters, war zones, and disaster sites. Different wired/wireless communication technologies could be integrated into the ultrasound system for possible utilization in remote data applications where medical information may be transmitted from the mobile unit to some centralized base station, such as an emergency room or field hospital (Dickson, 2008). However, video/audio transmissions over various communication methods are lossy, and the loss in compression/decompression and transmission affects the diagnostic information conveyed to doctors and trained personnels, and thus the ensuing decision-making process. In addition to telecommunication methods, other factors such as time of the day, video codec/compression method, video resolution and the dynamics of transmitted video content all affect the quality of transmitted videos.

In this paper, we examine ways to quantify video quality degradation in tele-ultrasonography, and compare the amount of video degradation under various conditions.

The basic setup of transmission is as follows. In the source terminal, an ultrasound transducer is applied to the patient under examination and ultrasound images/data are recorded and displayed. The transducer and peripheral hardware then encodes the recorded video (and possibly also audio for real-time conversation between operator and doctor), streams it and feeds to the transmitter. Depending on the type of transmission, the multimedia stream is sent using one of real-time transmission protocols and gets degraded by network conditions such as package loss, latency, etc. On the receiver side, the stream is picked up, decoded and displayed for the doctors and personnel. It is easy to see that the encoder/decoder, transmission path and the video content itself are all attributable to quality degradation at the receiving end. We will concentrate on evaluation of video quality degradation of tele-ultrasonography in this study.

Possible video codecs used in our study include: Video Optimization SDK (VOSDK), Windows Media Version 3 (WMV3) pull mode (proprietary), WMV3 push mode (proprietary), and uncompressed raw data. It had been shown (Dantcheva, 2007 and Stuhlmuller et al., 2000) that the MPEG2/H.264 codecs are sensitive to the blocking effect of compression artefact due to its block-based DCT coding schema. Such artefact is usually undesirable in ultrasound video, so we avoid the MPEG2/H.264 codecs.

Possible transmission methods used include:

- (i) local area network (LAN) using wired connection
- (ii) IEEE 802.11 for short-range wireless communication (typically less than 100 meters)
- (iii) third-generation (3G), and
- (iv) fourth-generation (4G) wireless transmission

Their properties and effects on ultrasound imaging transmission have been studied. Ultrasound video formats include: Video Graphic Array (VGA) with resolution 640x480 pixel, and Quarter Video Graphic Array (QVGA) with resolution 320x240. The dynamics of the transmitted video is studied, as high dynamics and/or contrasts affects the compression rate of codecs, and thus the amount of data transmitted. The degradation will also be dependent on factors such as time of the day (morning/evening) and day of the week as traffic conditions of wireless transmission and to a less extent, wired transmission, are determined.

The study is divided into two steps. First, we developed software for evaluating video quality degradations, and use it on processed videos using iPresence-U streaming video software product. Such equipment is capable of importing and transmitting at different transmission speeds, and record the transmitted videos using VOSDK codec and Audio Video Interleaved (AVI) format. Second, we use ultrasound scanner and transceiver equipment and test under combination of aforesaid factors. We use

the developed analysis software package on these real videos and try to give suggestions on optimal video codec/resolution for a given transmission condition.

First we use simple methods to register frames of both videos under consideration, then quantify video degradation using several classes of methods. Complicated image registration techniques like (Lucas and Kanade, 1981) that utilize spatial intensity gradient information to direct the search for the position tend to be time-consuming and may only bring limited improvement for cyclic video clips typically encountered in telemedicine. We solely rely on the information header (meta-data) associated with the videos and frame replica for detecting frame drops. Assuming that frames are paired up perfectly between two videos, we then quantify video dynamics of a video using its p -norm difference of adjacent frames. The video degradation is then quantized by inter-frame difference, characterized by RMSE, histogram difference and 2D-DFT energy distribution.

2 Method of Analysis

2.1 General Method

The streaming video software package “iPresence-U” for network transmission and capture is developed using C++ under Windows OS environment. It is capable of streaming ultrasound video from the Terason Ultrasound capture over landlines connection or wireless connection for different coding schema. Fig 1 shows the GUI interface of iPresence-U at work.

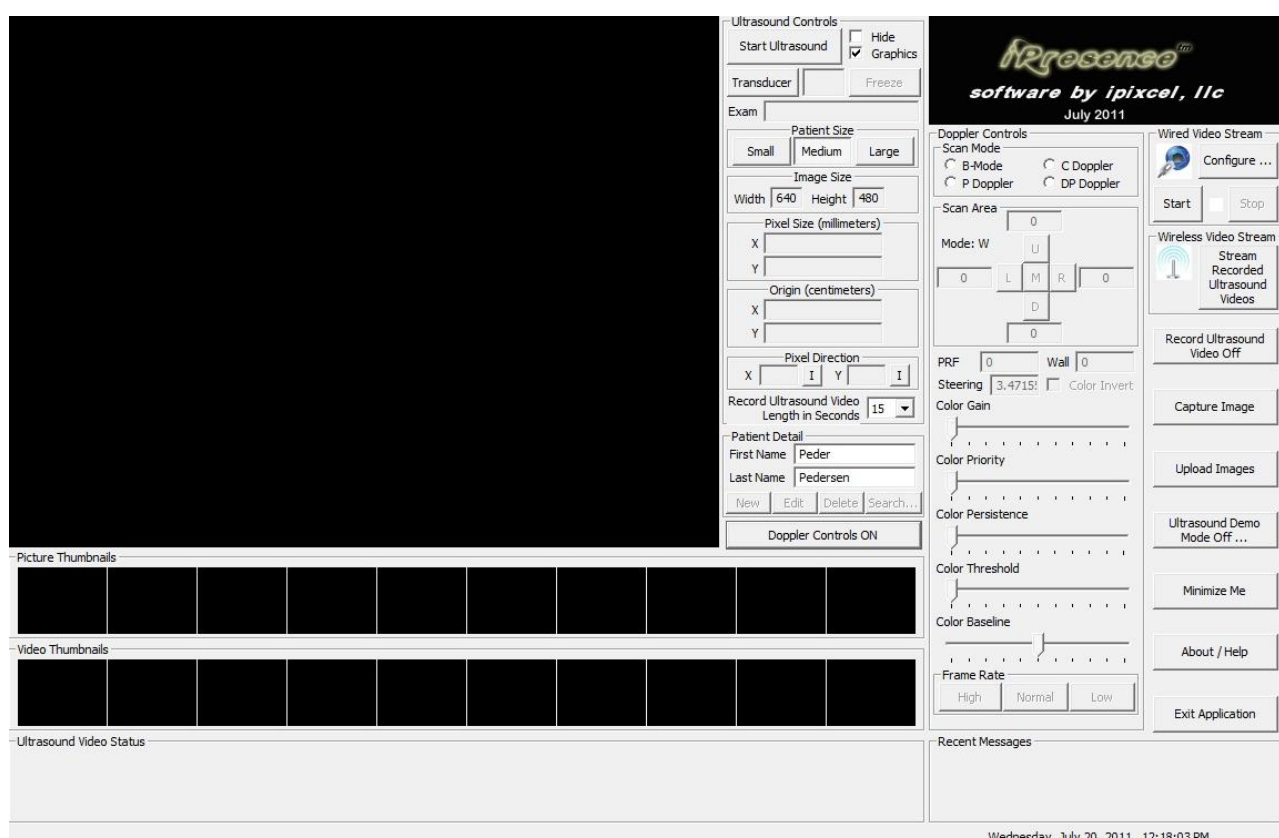


Figure 1 iPresence GUI

The core part of our software package is developed using C++ and OpenCV APIs under Linux OS environment with the GUI written using FLTK. It analyzes two sets of videos (original and transmitted) and gives some quantitative measure of individual and comparison of the quality degradation associated with the transmitted one.

To start with, a pair of original and transmitted videos are subject to our analysis. Usually, we have several transmitted videos from the same original video under various conditions. We try to match up both frames from the video pair when comparing as described later.

Fig 2 shows the GUI interface (preliminary stage) of video analysis package when loading paired videos for comparison. Upper left corner shows the prompt for video files, upper right shows the first frame of original video, and lower right the first frame of transmitted video. Fig 3 shows some 2D-DFT transform results of trivial patterns. The three figures in the first row are the ringings, the 3x3 tiled ringings, and gridding pattern; the three figures below are their corresponding 2D-DFTs.

Fig 4 shows the pair of video frames in the right (upper-right transmitted video, lower-right the transmitted one at 192 kbps; upper-left the 2D-DFT of the current frame of transmitted video, and lower-left the 2D-DFT of the current frame of transmitted video; upper middle shows the histograms for both frames)(Gonzalez and Woods, 2007). It also shows a comparison of unnormalized spatial energy distributions of two frames in the lower middle part. For example, the bars to the left correspond to low-frequency energy regardless of direction (angle), and bars to the right correspond to high-frequency energy. We can see that for the current frame pair, the original and transmitted have roughly equal energy at lower frequencies, whereas the original frame has higher frequency distribution than the transmitted. This may imply that certain details of the image are lost in compression/transmission. However, we observed that for some frames in this video pair, the high-frequency energy are higher in the transmitted one, and vice versa. It is consistent that the low-frequency energy distributions are roughly the same throughout the whole frame sequence.

First, when the two input videos have different resolution (VGA vs. QVGA), the one with smaller resolution is up-sampled (doubled in both width and height) and we use the resulting high-resolution to perform analysis. Note that up-sampling does not increase the information content of the original video.

The ultrasound videos of VGA or QVGA formats have a 3-channel, 8-bit in blue-green-red (BGR) color space. However, the content of the video is mainly in gray-scale, with the exception of color Doppler flow mappings (CFM) as shown in Fig 5. Here the annotations in CFM provide real-time information important for diagnosis, but are of little use in analyzing video qualities. We thus first converts the BGR video into gray-scale, and then developed region of interest (ROI) selection tool that allows interaction for specifying polygon ROI. The statistics inside ROI are calculated instead of the whole image if it is specified.

2.2 RMSE Difference

When considering frame loss due to transmission, we track the AVI meta data associated with each frame (avi, 1996) which contains total frame count, current frame ratio (between 0 and 1, which indicate beginning and end of the video), frame per second, etc. When a frame drop is detected from the meta data of AVI header, or when a frame in either (or both) video is missing, we discard the corresponding frame of its counterpart if applicable. The complicated frame registration techniques between video pair is time-consuming especially with cyclic patterns observed in cardiac motion observed in Fig 6, where it shows the dynamics of some original/transmitted video pairs. (Note that the trailing ‘frozen-state’ of the transmitted video gives zero-dynamics at the end and is not plotted, nor used in statistics.) We see that the pattern of dynamics in the transmitted videos follows the original video with some error, thus hypothesize that simple frame registration methods using pairwise frame-difference (described below) may not work well, as the two curves do not seem to have obvious delay between them. As a matter of fact, Fig 7 shows the naive RMSE-based frame registration with search bounds ± 20 frames. We see that the naive method cannot ‘correctly’ match frames, as the curves of registration are not line of identity. We could also work with histogram difference or in difference in frequency domain, but they are less intuitive and the results are less ideal.

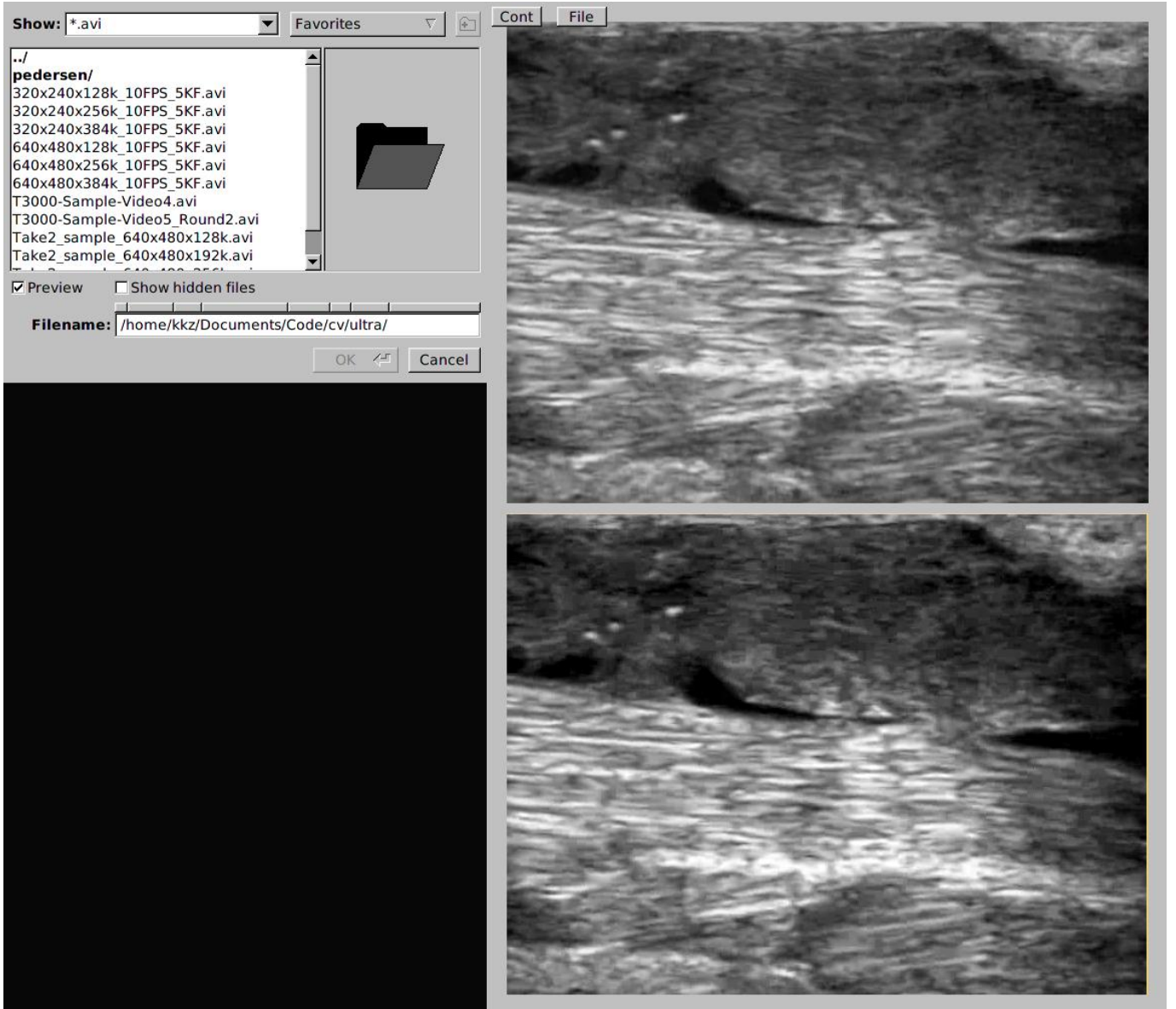


Figure 2 Video Analyzer GUI. Upper left window opens video files, upper right and lower right figures show the first frame of original and transmitted video.

To quantify the dynamics of a video, pair-wise frame difference is used. This simple method is usually sensitive to image transformations encountered in computer vision such as zooming, shearing, rotating and contrast change, where more advanced methods in feature space and object registration are often employed. However, without proper prior knowledge of the content of video used, it is hard to apply these techniques into our study.

We use the p -norm difference between adjacent frames $k - 1$ and k of $m \times n$ resolution to quantify instantaneous dynamic of the video at pixel $[i, j]$, and its mean to quantify the video dynamic, as shown in (1a). It raises the pixel-wise difference to p -th power, sums up the total difference across all pixels, and then normalizes its p -th root, so the p -norm difference has a unit same as pixel brightness. When the video record is still or has little motion, the dynamic is quite small, and vice versa. The measurement becomes mean absolute value (MAV) when $p = 1$ and becomes root-mean square error (RMSE) when $p = 2$. We use RMSE to quantify the video dynamics.

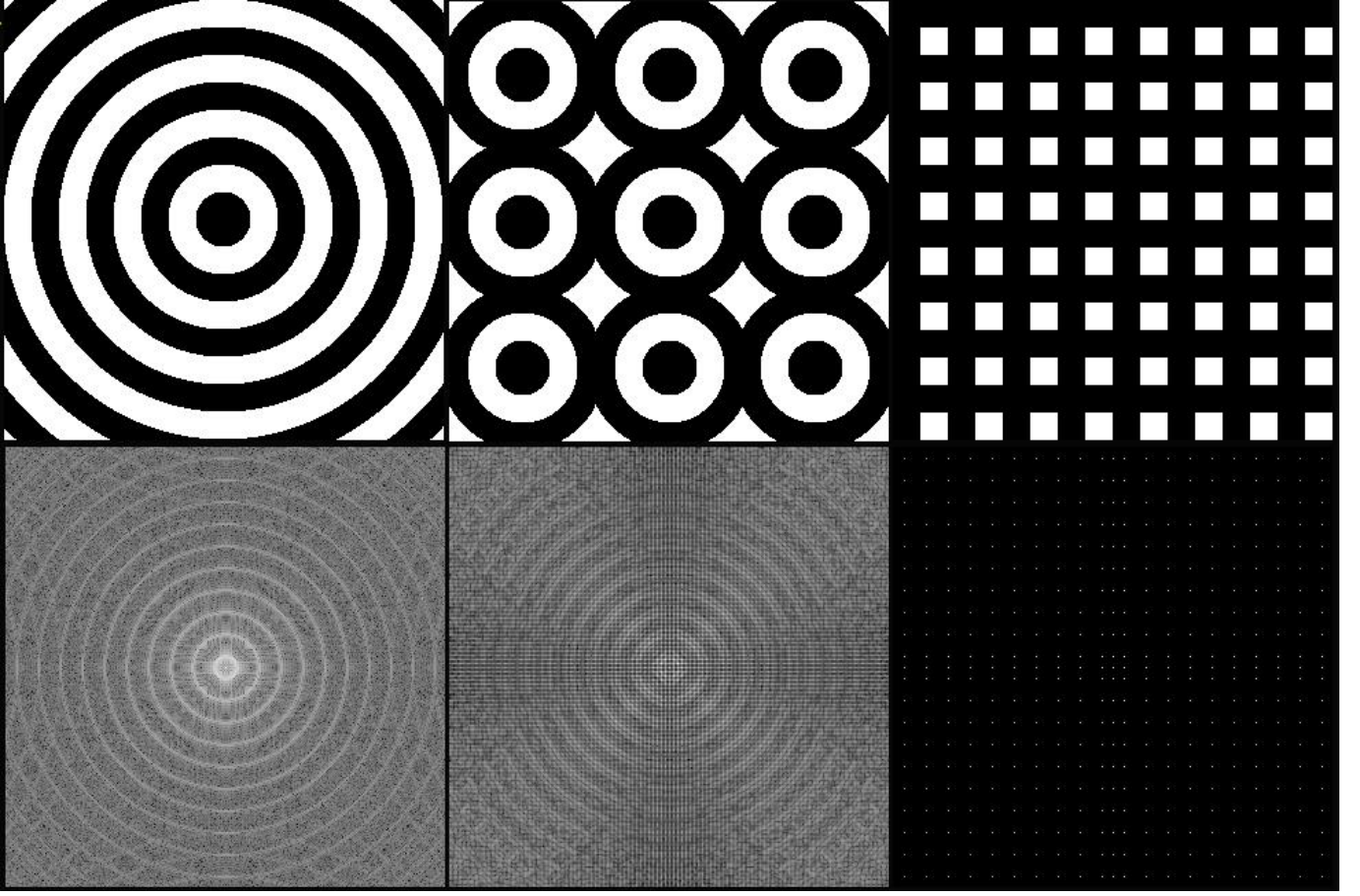


Figure 3 2D-DFT Results. Top row shows some spatial patterns with their DFT transform in the bottom.

$$|F_k - F_{k-1}|_p = \frac{1}{mn} \left| \sum_{i=1}^m \sum_{j=1}^n \left(|F_k^{i,j}|^p - |F_{k-1}^{i,j}|^p \right) \right|^{\frac{1}{p}} \quad (1a)$$

$$D_F = \frac{1}{K} \sum_{k=2}^K (|F_k - F_{k-1}|_2) \quad (1b)$$

The final statistics for the RMSE dynamics is averaged through all K frames of a video as given in (1b).

$$|F_k - G_k|_p = \frac{1}{mn} \left| \sum_{i=1}^m \sum_{j=1}^n \left(|F_k^{i,j}|^p - |G_k^{i,j}|^p \right) \right|^{\frac{1}{p}} \quad (2a)$$

$$D_{F,G} = \frac{1}{\min\{K_F, K_G\}} \sum_{k=1}^{\min\{K_F, K_G\}} |F_k - G_k|_2 \quad (2b)$$

It is assumed that for an original video, the transmitted one with higher bandwidth (or better network condition) will be a more accurate representation of the original video than the transmitted one with lower bandwidth, thus similar dynamics. This does not imply lower dynamics with poorer conditions, as we have observed that lower bandwidth sometimes gives higher dynamics. A manual inspection suggests that this could be due to contrast differences of the two videos, i.e. the histogram transmitted video frames have broader distribution, which can be seen from the two histograms in Fig 4, where the gray bars corresponding to the transmitted video frame have lower amplitude in the middle and higher

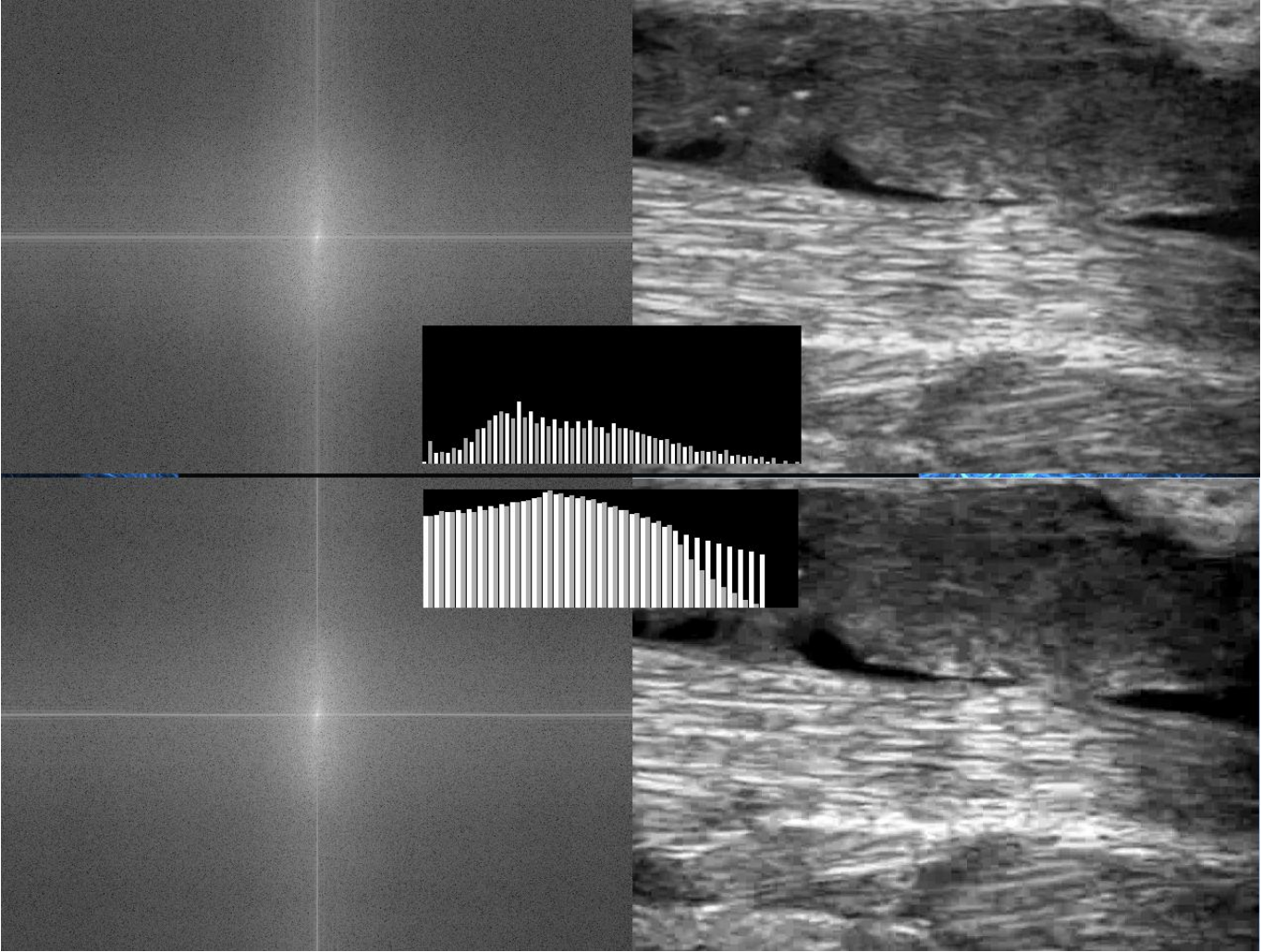


Figure 4 Video Analyzer GUI at work. The upper right image shows a frame from the original video whose 2D-DFT transform is shown in the upper left corner; lower right the matched frame of transmitted video with its 2D-DFT in the lower left corner. Both histograms are shown in the upper middle: bright bars are from the original video and gray bars from transmitted video; the energy distribution of spatial frequencies of the 2D-DFTs are also shown in the bar form in the lower middle: bright bars the energy distribution for original video frame and gray for the transmitted.

at both ends, and a comparison of the two frames at same position suggests that it could be attributed to blurring.

To quantify the differences of a pair of videos (or a pair of images), similar method can be applied to calculate the p -norm difference between such pair. Ideally, the difference would be 0 if the transmission and compression/decompression are lossless. High p -norm difference suggests large degradation of video quality as given in (2a); for K_F matched-up frames in video F and K_G matched-up frames in video G, the averaged paired difference is given in (2b).

2.3 Histogram and Frequency Measurements

Histogram gives a contrast measure of a static image. It estimates the gray-level distribution of a static image, and reduces the two-dimensional representation to a one-dimension representation by dropping its spatial distribution. As stated before, we perform analysis on the gray-scale images converted from BGR color images. With a specified number of uniform bins n , we calculate the histograms $H_{i,m} \geq 0$ ($i = 1, 2, 0 < m \leq M_i$) of original/transmitted pair for two videos under study, where the first sub-index indicates the original/transmitted video, and the second the frame index of the i -th video.

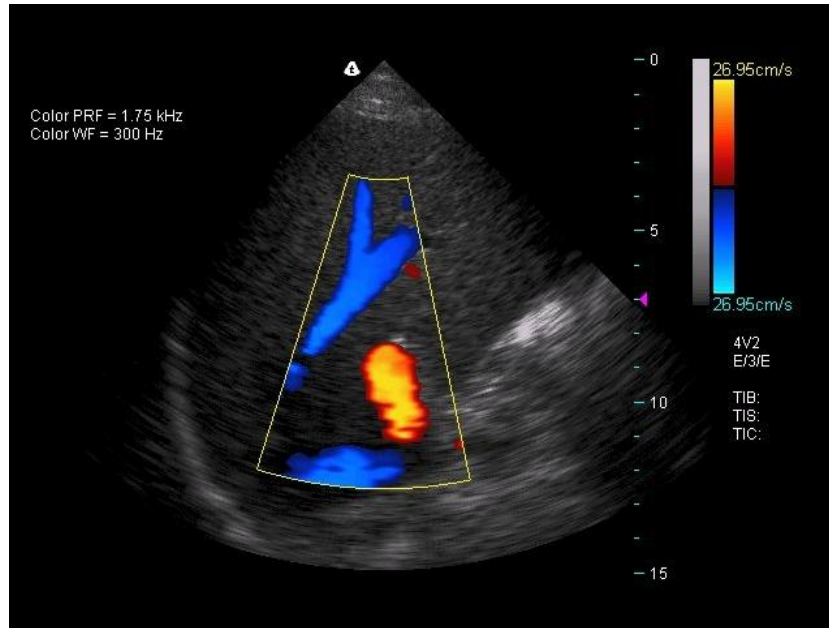


Figure 5 Ultrasound image frame with color Doppler flow mapping and fan-shaped ROI.

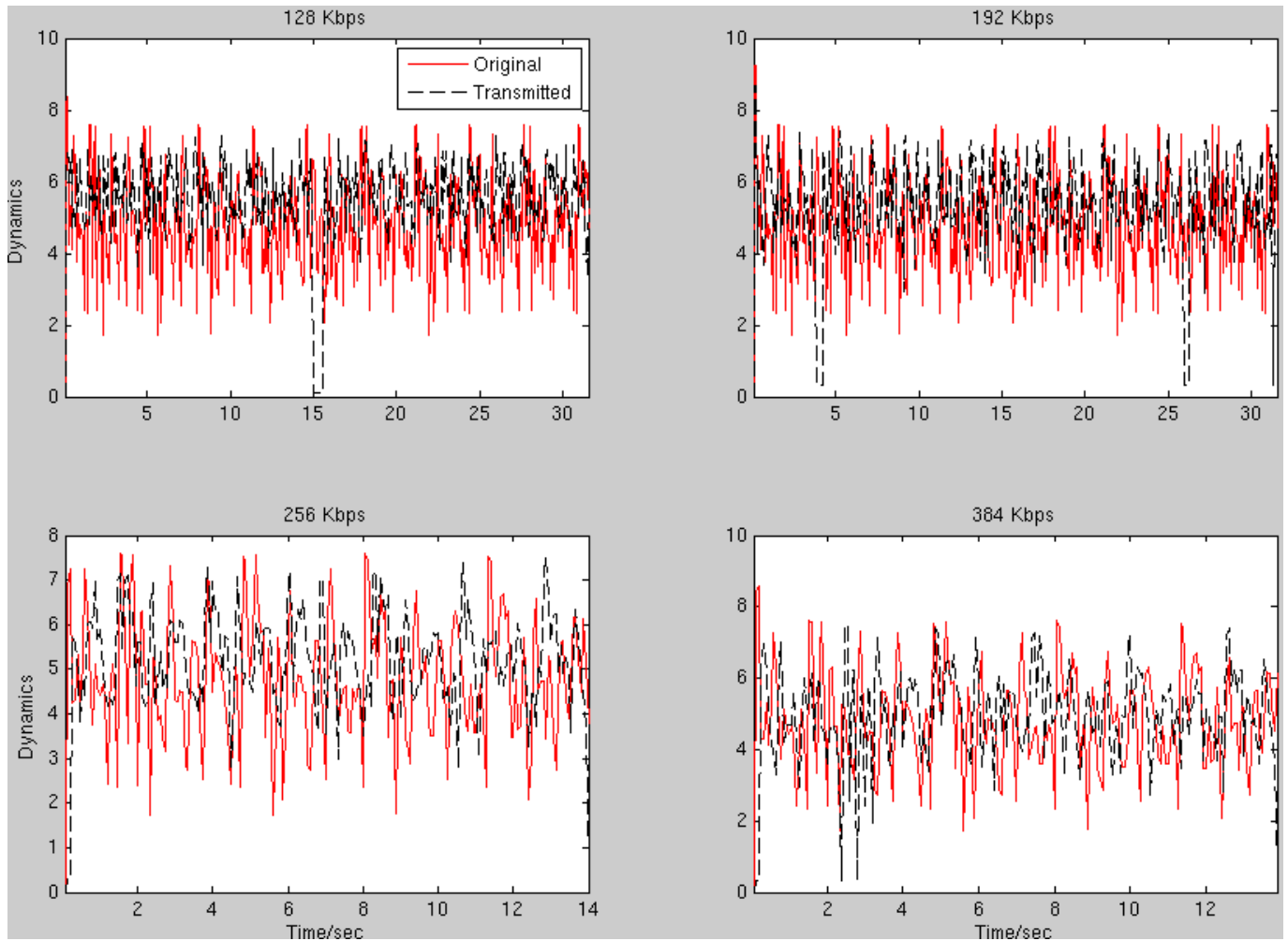


Figure 6 RMSE Dynamics of original and transmitted videos. The four figures show the RMSE difference between adjacent frames when transmitted using 128/192/256/384 Kbps. Red solid lines show the original dynamics and black dash lines the transmitted.

The commonly used criteria (correlation, Chi-square, Intersection and Bhattacharyya) for calculating histogram differences are given below from (3a) to (3d), where N is the total number of pixels, or the

Frame Registration

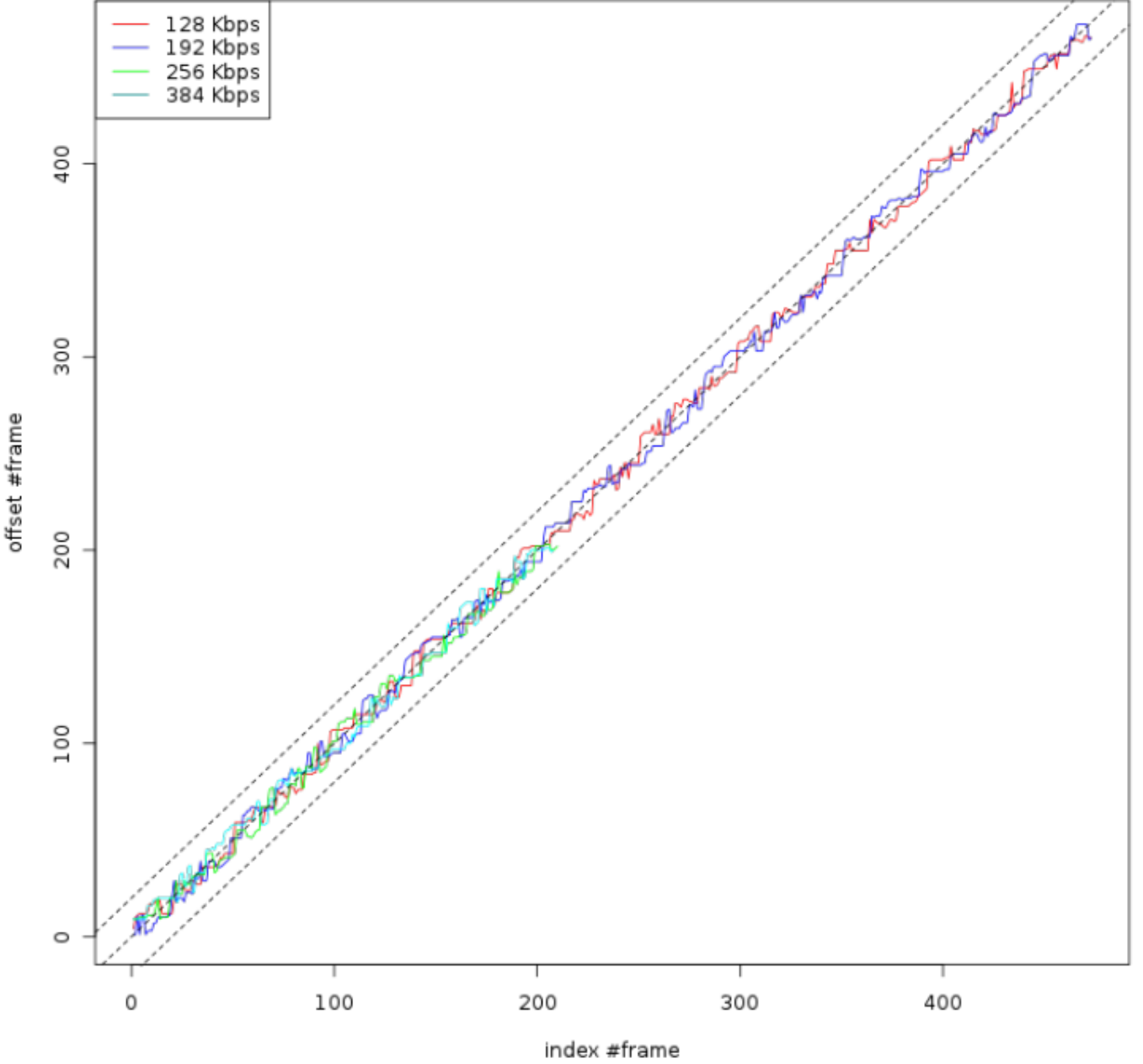


Figure 7 RMSE-based frame registration results for transmitted videos with different bandwidth. The x-axis is the frame number of original video, and the y-axis is the registered frame with minimum RMSE difference from the original. The perfect match and the search ranges that yield lines of slope=1 with offsets equal to 0 and ± 20 (search range) are shown in dashed lines.

sum of all bins in a histogram, and \bar{H}_i is the mean of histogram $H_{i,m}$ for all m . When the frame pair of the two videos are matched up, we have $m = n$. (3d) gives the averaged histogram difference of original/transmitted videos using correlation criterion when frames are matched up.

$$d_{Corr}(H_{1,m}, H_{2,n}) = \frac{\sum_{i=1}^n (H_{1,m}(i) - \bar{H}_1)(H_{2,n}(i) - \bar{H}_2)}{\sqrt{\sum_{i=1}^n (H_{1,m}(i) - \bar{H}_1)(H_{2,n}(i) - \bar{H}_2)}} \quad (3a)$$

$$d_{Chi}(H_{1,m}, H_{2,n}) = \sum_{i=1}^n \frac{[H_{1,m}(i) - H_{2,n}(i)]^2}{H_{1,m}(i) + H_{2,n}(i)} \quad (3b)$$

$$d_{Inter}(H_{1,m}, H_{2,n}) = \frac{\sum_{i=1}^n \min [H_{1,m}(i), H_{2,n}(i)]}{N} \quad (3c)$$

$$d_{Bhat}(H_{1,m}, H_{2,n}) = \sqrt{1 - \frac{\sum_{i=1}^n \sqrt{H_{1,m}(i)H_{2,n}(i)}}{N\sqrt{\bar{H}_1\bar{H}_2}}} \quad (3d)$$

$$\bar{d}_{Corr}(H_1, H_2) = \frac{1}{\min\{M_1, M_2\}} \sum_{k=1}^{\min\{M_1, M_2\}} d_{Corr}(H_{1,k}, H_{2,k}) \quad (3d)$$

We can see that if we pose constraint that $\sum_{i=1}^n H_{j,k}(i) = 1$ where n is the number of bins, then (3a) becomes the coefficient of correlation.

The Pearson's Chi-square statistics for goodness of fit is given in (4)(Baker and Cousins, 1984), where O_i and E_i are the observed and expected (Null-hypothesis) frequency respectively. The test statistic asymptotically approaches χ^2 distribution with large n . By comparing (3b) with (4), we can see that the numerator in (3b) is twice the fitting error in (4) if we let $E_i = \frac{H_{1,m}(i)+H_{2,n}(i)}{2}$, and the denominator in (3b) is also twice that in (4). In this sense, this histogram difference metric is the same as Pearson's Chi-square statistics. The intersection metric of histogram difference in (3c) is more intuitive in that the numerator is the "commonality" of both distributions.

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (4)$$

It is easy to see that $d_{Corr} \in [-1, 1]$ with $d_{Corr} = 1$ when the two histograms are same, and $d_{Corr} = -1$ when they are opposite; that d_{Chi} is non-negative and is zero when the two histograms equal; $d_{Inter} \in [0, 1]$ with $d_{Inter} = 1$; and $d_{Bhat} \in [0, 1]$ and $d_{Bhat} = 0$ when they equal. We can also normalize the Chi-square difference as we do with the intersection criterion when the bin number gets large or the two histograms are quite different. It is easy to formulate two images with same histogram distribution regardless of the number of bins but non-zero RMSE difference: a permutation of pixels would suffice; and to formulate two pairs of images with same RMSE difference but different histogram differences: increasing/decreasing the intensities of all pixels of an image pair by same amount would work. Therefore, the two methods measure different aspects of image content. In the same way, the quantized histogram difference can also be used to measure the dynamics of the same video.

A third measure of image differences is in the spatial frequency domain (5a). We assume that the size of image in spatial $f(k, l)$ and frequency domain $F(i, j)$ are $L \times K$. We see that it essentially multiplies a rectangle window of image size to the time domain before the Fourier transform is applied (zero-padding). The resulting transform has DC component at the four corners and high-frequency content in the center. We usually rearrange the four quadrants as in (5b) so that the lower frequency component is positioned in the middle.

$$F(i, j) = \sum_{l=0}^{L-1} \sum_{k=0}^{K-1} f(l, k) \exp \left\{ -2\pi \left(\frac{il}{L} + \frac{jk}{K} \right) \right\} \quad (5a)$$

$$F_c(i, j) = F \left(\text{sgn} \left(\frac{L}{2} - i \right) \left(\frac{L}{2} - l \right), \text{sgn} \left(\frac{K}{2} - k \right) \left(\frac{K}{2} - k \right) \right) \quad (5b)$$

Unlike histogram measure, it takes into consideration the spatial distribution of the images. For each frame, we take its 2-D DFT to convert the real image into complex planes. To quantify the image differences with a single number, we need to reduce the amount of information represented in frequency domain. First we discard the phase plane. Although human vision system is very sensitive to phase information of an image, we do so to eliminate the information associated with image shifting. Second,

we convert the magnitude plane from Cartesian coordinate into Polar coordinate, and integrate over the distance on magnitude axis, which is equivalent to integrating rings with specified width of the magnitude plane of Cartesian coordinate. By doing this, we further half the information by neglecting the orientations of the frequency pattern of original image, concentrating only on the energy distribution of frequency component in all directions. In this way, we can similarly get a histogram-like representation of the frequency domain content and fall back on the techniques for comparison described above. In getting the array-representation, we calculate the mean-brightness of each pixel representation in frequency domain by dividing the number of pixels in the ring. This ensures that the final array have elements less than 1. Note that DFT is not used in most mainstream image compression techniques as round-off errors in phase/magnitude and the presence of Gibbs phenomena indicates easily observed quality loss.

Table 1 shows the statistics of some available video pairs. With a 30-second original video of 15 frame per second (fps) and a total of 537 frame without annotations and ROI taking the full image. The statistics of an irrelevant video is also given to give a sense of closeness of these measures. The starred entry indicates closest match to transmitted video among candidates. The χ^2 entry of frequency measure is not given because all the results are very close to 0. We can see that measures of similarity using RMSE, histogram and frequency have four stars for 256 kbps transmission, three stars for 384 kbps transmission and one star for 192 kbps. The video dynamic of the 384 kbps transmitted video is closest to that of transmitted one. These indicate that under such measurements, the 256 kbps and 384 kbps are preferred.

Video		128 K	192 K	256 K	384 K	Irrelavent
#Frames	Transmitted	474	474	211	209	414
RMSE(2b)		6.8610	6.8652	6.8485	6.8377*	9.0974
Dynamics(1b)	Original	4.8289	4.8331	4.7925	4.8097	4.8070
	Transmitted	5.4776	5.3150	5.1914	4.9512*	2.2476
Histogram	Corr(3a)	0.9218	0.9197	0.9242*	0.9231	-0.2173
	χ^2 (3b)	0.05420	0.06055	0.05317*	0.05389	1.2777
	Inter(3c)	0.1073	0.1094*	0.1058	0.1061	0.7451
	Bhatt(3d)	0.1278	0.1296	0.1271*	0.1281	0.7061
Frequency	Corr	0.9742	0.9736	0.9765*	0.9762	0.9201
	Inter($\times 10^{-3}$)	7.622	10.92	9.669	11.15*	0.76
	Bhatt($\times 10^{-2}$)	1.680	1.605	1.553	1.530*	2.454

Table 1 Statistics of ultrasound video pairs of different transmission methods. The original video has a total of 537 frames, and all the missing frames of transmitted ones (except the “Irrelavent” column) happen at the end of the video, i.e. no frame drop is detected in the middle.

2.4 Frame Loss

To deal with the frame loss, the software is also (yet to be) designed to be able to interpolate frames when such events are detected, and can be set causal (offline or with a few frames buffered ahead of displaying) or non-causal (real-time sending transmitted frame to display). Two categories of methods are employed in frame interpolation.

The easier method is pixel-wise interpolation. For each pixel of the missing frame(s), linear/quadratic interpolation is performed on its adjacent frames to estimate its gray-level. We found that the quadratic interpolation tends to bring artificial edges, so we also introduced constrained quadratic interpolation. We found that empirically, the quadratic or constrained quadratic interpolation method do not give a boost on the quality of the interpolated frame compared with the linear method, or makes the interpolated video any smoother. It is easy to see that the draw back associated with pixel-wise interpolation is phantom (?) in presence of motion.

The complicated and more time-consuming technique is to borrow optical flow calculation in machine vision. This method tries to match each block (or object) of a series of images and gets its motion trajectory (Mallat, 2003). With such information, frame interpolation is simply inserting locations of a block. Common methods of optical flow include the Lucas-Kanades method, Horn-Schunck method, block-matching method and pyramidal implementation of Lucas-Kanades method (Bruhn et al., 2004).

3 Discussion of Preliminary Tests

We have quantized dynamics of a video using averaged frame-wise RMSE. We hypothesize that videos with higher dynamics are harder to compress, take more network bandwidth, and are more prone to degradation.

We then try to quantize the video degradation between original/transmitted videos using three classes of methods. First, we used averaged RMSE difference (or more generally, the p -norm difference) between pre-registered frames. Then we tried to characterize the gray-level distribution by means of histogram. Several formulas for comparing the histogram-pair of a pre-specified bin number are used. Third, we compared the energy distribution at different frequencies of frame pair under study using 2D-DFT. Here we ignore the directional by integrating rings with fixed width to project two-dimensional distribution into one-dimensional array. We have observed that the low frequency energy distributions are roughly the same, whereas the high frequency distribution can be quite different.

We avoided the more complicated frame registration techniques and showed registration results using simple technique based on frame-wise RMSE difference. While it is yet to be known if complicated frame registration algorithms using optical-flow method yield substantial improvements, we see that our registration results give random fluctuations. For actual application, we used AVI meta-data associated with each frame and RMSE differences of adjacent frames to detect a frame drop event, and found no drops in the transmitted videos except at the end. When such events are detected, they can be used to quantize video quality degradation due to transmission.

We have shown in table the quantitative measures that generally, higher bandwidth gives lower video degradation, in agreement with our surmise that more network resources diminishes quality loss.

4 Experimental Results

5 References

- Dickson, B. W. (2008). Wireless communication options for a mobile ultrasound system. .
- Dantcheva, A. (2007). Video quality evaluation. , pp. 14-16.
- Stuhlmuller, K., Farber, N., Link, M. and Girob, B. (2000). Analysis of video transmission over lossy channels. *IEEE Trans. Selected Areas in Communication*, 18, 1012-1032.
- Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *Proc. Imaging Understanding Workshop*, pp. 121-130.
- Gonzalez, R. C. and Woods, R. E. (2007). *Digital Image Processing*. Third edition Prentice Hall.
- (1996). Opendml avi file format extensions. *OpenDML AVI MJPEG File Format Subcommittee*, pp. 1-10.

- Baker, S. and Cousins, R. D. (1984). Clarification of the use of chi square and likelihood functions in fits to histograms. *Nuclear Instruments and Methods in Physics Research*, 221, 437-442.
- Mallat, S. (2003). *A Wavelet Tour of Signal Processing*. Second edition Academic Press.
- Bruhn, A., Weickert, J. and Schnorr, C. (2004). Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61, 211-231.