EVALUATION OF TOTAL WORKSTATION CT INTERPRETATION QUALITY: A SINGLE-SCREEN PILOT STUDY

Beard DV[1,3], Perry JR[1], Muller K[2], Misra R[3], Brown P[3], Hemminger BM[1], Johnston RE[1], Mauro M[1], Jaques P[1], and Schiebler M[1],

Departments of Radiology[1], BioStatistics[2], and Computer Science[3], University of North Carolina at Chapel Hill

ABSTRACT

An interpretation report, generated with an electronic viewbox, is affected by two factors: Image quality, which encompasses what can be seen on the display, and computer human interaction (CHI), which accounts for the cognitive load effect of locating, moving, and manipulating images with the workstation controls. While a number of subject experiments have considered image quality, only recently has the affect of CHI on total interpretation quality been measured.

This paper presents the results of a pilot study we conducted to evaluate the total interpretation quality of the FilmPlane2.2 radiology workstation for patient folders containing single forty-slice CT studies. First, radiologists interpreted cases and dictated reports using FilmPlane2.2. Requisition forms were provided. Film interpretation was provided by the original clinical report and interpretation forms generated from a previous experiment. Second, an evaluator developed a list of findings for each case based on those listed in all the reports for each case and then evaluated each report for its response on each finding. Third, the reports were compared to determine how well they agreed with one another. Interpretation speed and observation data was also gathered.

1. FILMPLANE2.2

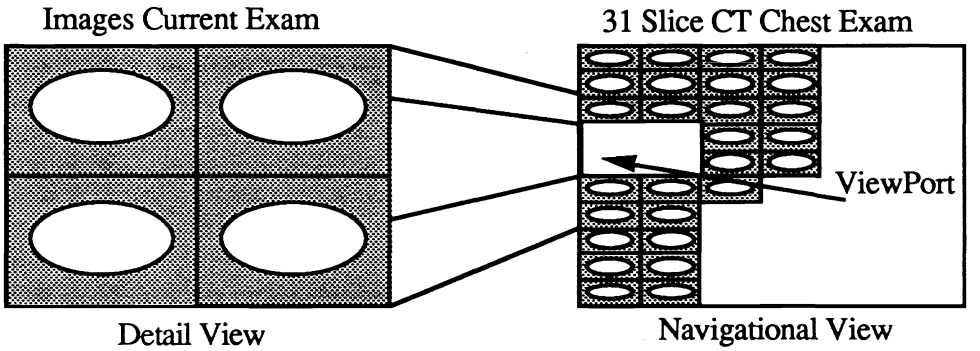

Figure 1
FilmPlane2.2 Viewport, Navigation View, and Detail Views.

FilmPlane2.2, the latest FilmPlane prototype, continues to provide the user with a image index mental model [Beard 1990B] of a large two dimensional surface upon which are arrayed all the images in the patient's folder (see Figure 1). By imagining themselves "flying" over this surface, users can quickly learn how to access any image and scroll through the various examinations. FilmPlane2.2, imple-

 0-8194-0541-8/91/$4.00

mented using X11, runs on UNIX workstations that support the X windowing system, such as DEC 3100s and Sun SPARCs.

Users can move the viewport and corresponding detail view through the navigation view in two ways. First, they can toggle to the navigation view, grab the viewport with the mouse, and move it to a new location in the navigation view, moving the corresponding detail view. Second, they can scroll [Beard 1989, Walker 1990] sequentially through the images in an examination. The user could toggle the system to scroll in increments of either 256 or 512 pixels. FilmPlane2.2 also allows smooth scrolling.

FilmPlane2.2 provides 12-bit dynamic intensity windowing with 8-bit framebuffers. Static intensity windowing is accomplished as follows: Every time an image is displayed, its 12-bit data (actually stored as 16-bit data) is window-width-and-leveled [Pizer 1989] into 8-bit pixels using a lookup table before being moved to the framebuffer. Presets can be used to change the width and levels reflected by the lookup table (Figure 2). The choice of presets automatically changes as the user moves the cursor among the various modality studies in the patient's folder.

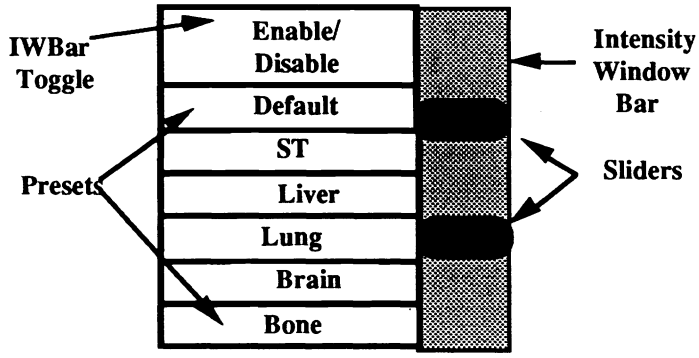

Figure 2
Intensity-Windowing Bar and Presets

Dynamic intensity windowing is a bit more complex(Figure 2). When the intensity-window-width-and-level slider-bar (IWbar) is "enabled," the IWBar visually changes, the 12-bit image being displayed is linearly reduced into an 8-bit image, and the current 12-bit window-width-and-level settings are reflected as corresponding 8-bit values realized in the workstation's color map. Thus, while the IWbar is enabled, moving the bar actually changes the machine's color map in real time, showing an image that is very similar to the actual 12-bit windowed image. When the IWbar is disabled, the normal linear color map is restored, and the 12-bit intensity window values corresponding to the last 8-bit settings are used to re-display the correctly windowed image. Thus, this intensity window method allows the use of low-cost 8-bit framebuffers [Beard 1990A], while still facilitating dynamic windowing.

2. PILOT EXPERIMENT

This study evaluated the relative interpretation accuracy of a single-screen FilmPlane2.2 radiology workstation to film for interpretation of forty-slice CT Chest studies. Our objectives were as follows: to determine whether any design changes were necessary to the FilmPlane2.2 computer-human interaction, to train a pool of radiologists to use FilmPlane, to evolve the experimental protocol, to provide

data for power analysis estimates for the final experiment, and to obtain preliminary results on FilmPlane's effectiveness, relative to Film, with respect to image quality and interpretation task time.

The workstation in question only used a single 1Kx1K monitor. While this is clearly inadequate, it did allow us to conduct a pilot study while the multiple screen version was begin completed.

Since we were interested in total interpretation accuracy, rather than just image quality, we could not use a traditional "structured response form" [Johnston 1990] to gather accuracy responses, for the form's structure would affect the radiologist's work patterns, working memory load, etc., resulting in a non-clinical situation. Therefore, workstation interpretation results were gathered using dictated interpretation reports similar to [Foley 1990]. These reports from the workstation interpretations were compared against the interpretation report generated with film in the clinic, and against two detailed findings lists from [Johnston 1990] to compare the various findings noted by the radiologists using film and workstation. Workstation task times were compared to typical task times for interpreting similar cases in the clinic.

2.1 Method

Subjects:Three board-certified Radiologists participated. Two had extensive CT chest experience, and the third was a senior chest radiologist with limited CT experience. In addition, three other radiologists read several cases allowing us to gather additional observations. All subjects are on the faculty of the University of North Carolina Department of Radiology.

Cases: Nine patient image folders were used, each containing a 12-bit-per-pixel CT chest study consisting of from 36 to 47 slices. The attending physician's requisition forms were provided on paper.

Apparatus: Images were interpreted on a one-screen version of the FilmPlane2.2 radiology workstation using a Sun 4/370 with 32 Megabytes of main memory, an 8-bit framebuffer and color monitor. A conventional dictation device was provided and a videotape system with a stop watch function was used to record times, hand and head motions, and verbal protocol.

Procedure: A session started with the subject seated in front of the one-screen Sun workstation in the subject laboratory. Since one purpose of this pilot was to train subjects, only a very brief training period was provided. During training, the experimenter described the workstation metaphor, and demonstrated the basic interactions, taking about two minutes. Then the subjects, with coaching, navigated through test patient folders and adjusted contrast with presets and with the double-slider bar until they indicated they were ready to proceed. The entire training period was typically under five minutes. Case interpretation order was randomized, and each subject read each case. For each trial, a patient image folder containing a single CT case was presented with the navigation view initially displayed. The subject was instructed to interpret the study and dictate a report with the same professional standards of accuracy as use for clinical film interpretations.

Data Analysis: An evaluator examined the six reports for each case, that is, the original clinical report, the two findings forms from [Johnston 1990], and the three dictated reports from this pilot study. For each case, he compiled a list consisting of the union of all findings from all reports on that case. If one report contained a more detailed version of the same finding, e.g. if one report stated "lung tumor" and a second "tumor in right lung," the evaluator listed all finding permutations, that is, one for a lung tumor, and the second for a tumor in the right lung. Finally, for each of the six reports and for each finding, the evaluator determined whether the report stated that a particular finding was present. Only findings which were available on the findings form used in [Johnston 1990], were considered in the data analysis.

2.2 Results

Subject A completed the training case and all nine cases in a single 3 hour session. It was clear that the last few cases were rushed, but the radiologist felt the results were still of clinically acceptable quality. Subjects B completed the training and the first four cases in one session, and the remaining five cases in a second session. Subject C completed the training and the first three cases in an initial session, and three cases each in two subsequent sessions.

2.2.1 Data

Accuracy: Table 1 shows the amount of agreement between the findings of each pair of reports. Each entry shows the sum across all findings of the absolute value of the difference between each report's value for that finding, with "finding-present" valued as 1, and a "finding not present" valued as 0. This is equivalent to the "percent agreement" between these two reports. Average agreement for a given report, is the average of all five agreement-pair values in which that report participated. The average of all the workstation entries (SA, SB, and SC) is 0.32, which is the same (within two digits) as the average for all the film entries (CR, JFA, JFB).

	CR	SA	SB	SC	JFA	JFB
Original Clinical Rpt (CR)	0					
FilmPlane Subject A(OA)	.34	0				
FilmPlane Subject B(OB)	.25	.33	0			
FilmPlane Subject C(OC)	.26	.32	.27	0		
Johnston90 Form A (JFA)	.35	.36	.29	.38	0	
Johnston90 Form B (JAB)	.41	.38	.4	.34	.25	0
Average Agreement	.32	.35	.31	.33	.33	.32

Table 1
Agreement for each report pair, and average for each report.

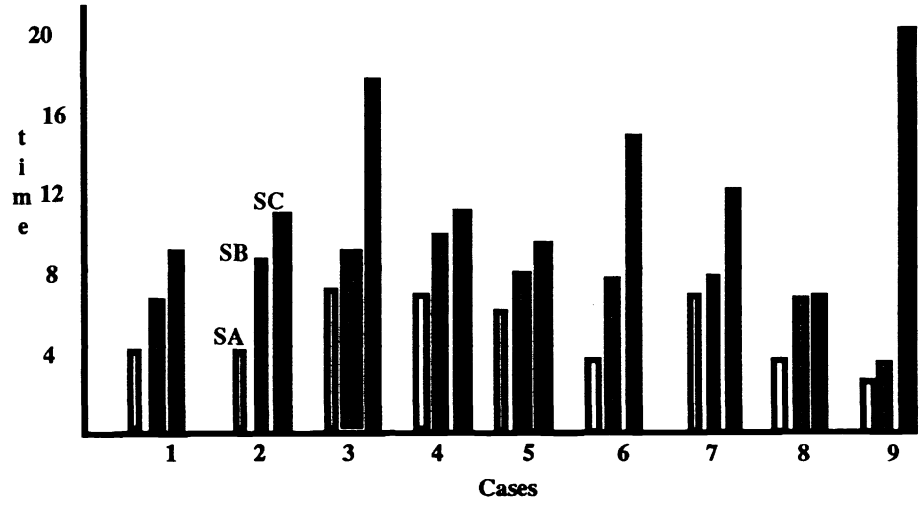

Figure 3
Task time for each subject and each case. Time in minutes.

Interpretation Speed: Figures 3 and 4 show the task times for the FilmPlane subjects interpreting the various CT cases. Figure 3 shows the interpretation of each separate case. Subject C, not being a frequent interpreter of CT, was considerably slower. Figure 4 shows the task times for the cases in the order in which they were read. While subject C does not show any learning, subjects A and B - those who frequently interpret Chest CT exams - seem to reduce their task times over the 9 cases in this pilot study. These data, along with observations made by the experimenter, indicate that these two subjects were able to use their considerable CT experience to develop more efficient methods of interpretation.

None of the subjects interpreted these CT cases as fast as is done in the clinic where a case is typically read in 2 or 3 minutes. The average interpretation time was 8 minutes and 13 seconds. The two frequent CT interpreters averaged 6 minutes 12 seconds overall, and 4 minutes and 50 seconds on the last three cases, which is still from 50% to 100% greater than typical clinical times. Subject C averaged 12 minutes 15 seconds over all the cases.

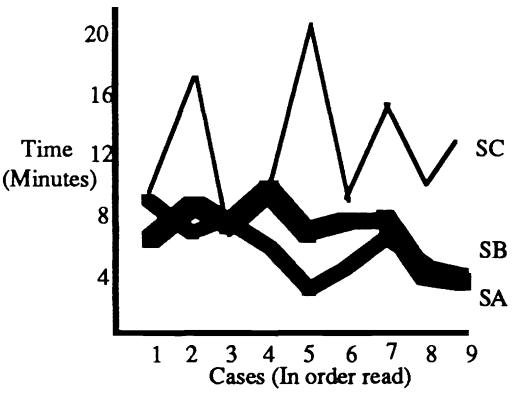

Figure 4
Task time for each subject and each case in the order read. Time in minutes.

2.2.2 Observations and Subject Comments

Learning: All subjects were able to use the system after only two minutes of instruction, followed by three minutes of interaction with assistance. Radiologists expressed confidence in their understanding of the interaction. The FilmPlane metaphor again proved useful, and the navigation view often was used to determine the radiologist's position in the examination, and for rapid access to remote anatomy. Slice ordering seemed critical to maintaining the metaphor. In one case the slices were in reverse order with the abdomen slices to the upper left and lung slices to the lower right. Radiologists had difficulty remembering which direction they needed to move in order to navigate to a desired anatomy.

Scanning approaches: Subjects viewed the CT cases by anatomy, typically starting with an overview of the navigation view and scout view to see the extent of the examination, followed by a lung-preset scan of the lung, a soft-tissue-preset scan of the entire study, and a liver scan using the liver window. Then radiologists would typically go back and review critical anatomy with one or more intensity window settings. This took a long time, apparently because of the limited screen space.

Intensity windowing: In general, the subjects used the default presets for intensity windowing. However, there were a number of occasions in which they used the IWbar to choose an intensity window setting not available with the presets. It was noted that when a non-preset window was used, it was often repeatedly reacquired during that interpretation; it would be useful to have a mechanism allowing the subject to temporarily reset a preset to a new setting.

In general, the dynamic intensity bar was successful, allowing 12-bit intensity windowing with only an 8-bit framebuffer. However, a few problems were encountered. First, when the bar was enabled, a very narrow intensity-window setting - such as a liver window - would produce a image that was far more "pixelly" than desirable. This was because an intensity window width of 100 intensity levels, which would produce an acceptable image, was reduced to a width of only about eight intensity levels. The image was, of course, correctly displayed when the bar was disabled and all 12-bits used. Second, the subjects sometimes forgot to disable the intensity window bar after selecting a new window setting. This could leave them with a less-than-optimal image.

Navigation: Subjects were able to successfully navigate over the images in the patient folder. Random images were accessed using the navigation view by moving the viewports. Minor problems were encountered with grabbing the viewport.

Mouse: There were problems with the optical mouse used in the experiment. When the mouse pad is square to the table, the arm of a right-handed radiologist is tilted slightly to the left, so that while the radiologist thought he was moving the mouse up and down, he was actually moving the mouse from 11 O'clock to 5 O'clock. Further, while most subjects dictated with the left hand while mousing with the right, one right-handed subject would keep the dictation microphone in the left hand when scrolling, and then switch it to the right for dictation. We observed that this had no significant impact on response time, and the radiologist did not comment on this as a problem.

Screen Space: The single 1Kx1K screen was acceptable when merely scanning through a series of images. But when trying to understanding an anatomical abnormality that required more than four slices, the overhead required to scroll back and forth over a large number images was bothersome, time consuming, and invited errors caused by either not seeing something, not being able to correctly understand the 3D anatomical relationships, or simply forgetting to dictate critical abnormalities. Comments such as "I cannot see enough of the heart at the same time," were heard frequently.

Response time: The typical response time of 1.5 seconds (for a scroll operation replacing 1/2 of the screen), seemed acceptable and radiologists often said it seemed fast. However one radiologist, when doing a screening scan through a sequence of images, would start the scroll operation before he had completely viewed the slices in question, thus anticipating the scroll. While it was good to see that the radiologist had developed an optimal strategy for using the workstation, it also indicated that a faster response time might be beneficial.

Strained working memory : Several radiologists noted that they could not mentally keep track of items for dictation at the end of the interpretation as they normally did with film. All the radiologists started dictating midway through the interpretation, and two asked for pen and paper in order to take notes. Several noted that the typical CT interpretation method taught to residents - in which each organ group is individually viewed with the appropriate intensity windowing - while appropriate for film with its rapid sequential access, did not work with the slower sequential rates of a workstation. Several radiologists took so long on a case that they lost track of what had been completed. One noted "Did I look at the pancreas?", and another said "Did I see the kidneys?"

Lack of Confidence: Despite the apparent acceptable accuracy of the dictated results, all the radiologists expressed a lack of confidence for the first few cases, and subject C, with the least amount of

CT experience, expressed a lack of confidence throughout the experiment. One subject jokingly noted "My degree of confidence is low - I ain't got them in my hand." All subjects were unhappy with the lack of screen space. One noted, toward the end: "You do get into a pattern, but I don't enjoy it"

Other Observations: First, one subject constantly touched the screen with his finger, leaving finger prints. We will either have to keep a lot of screen cleaner near by, or train him not to do this. Second, the scout view was critical, for counting ribs and locating findings. Third, several radiologists suggested the need for a tool to determine the intensity value of a particular pixel. Forth, several also requested a tool to measure the size of a mass. Fifth, the radiologists stated that they would like to have the postage-stamp images in the navigation view reflect the intensity window settings of the actual images: FilmPlane currently uses contrast-limited adaptive histogram equalization (CLAHE) [Zimmerman 1988, Pizer 1989] for the navigation view images.

3. CONCLUSIONS

The FilmPlane metaphor again proved effective in providing a viable interaction that can be learned in under five minutes. Nevertheless, as we expected, it is clear that a single 1Kx1K screen is not enough for a primary diagnosis of even a single CT examination. Most importantly, it was clear that none of the subjects would choose to use this single screen version of FilmPlane for typical clinical work. Since starting this pilot experiment, FilmPlane has been modified to operate with any number of screens.

Dynamic intensity windowing with 12 bit images and an 8 bit framebuffer seems viable, and the basic interaction implemented with FilmPlane2.2 is effective, through it needs additional developmental work. A workstation allowing true 12-bit intensity windowing is clearly superior, but given the cost advantage of 8-bit framebuffers, it is encouraging to see indications that commodity priced 8-bit units are viable.

Based on the results of this pilot study, we plan to conduct a similar large scale study evaluating a multiple-screen version of FilmPlane with cases containing multiple examinations.

4. REFERENCES

Beard, DV, Parrish, D, Stevenson, D, A Cost Analysis of Film Image Management and Four PACS systems based on Four Network Protocols, *Journal of Digital Imaging,* May 1990 (A).

Beard DV, Designing a Radiology Workstation: A ocus on Navigation During the Interpretation Task, *Journal of Digital Imaging* Aug 1990 (B).

Foley WD, Jacobson, DR, Taylor AJ, et. al, Display of CT Studies on a Two-Screen Electronic Workstation Versus a Film Panel Alternator: Sensitivity and Efficiency Among Radiologists, *Radiology,* 174, 769-773, March 1990.

Johnston RE, Yankaskas BC, Perry JR, Pizer SM, Delany DJ, Parker LA, Agreement Experiments: A Method for Quantitatively Testing New Medical Image Display Approaches, SPIE vol. 1234 Medical Imaging IV, 1990.

Pizer, SM, and Beard, DV, "Medical Image Workstations: State of Science & Technology". *Journal of Digital Imaging*, Nov.1989 2(4) 185-193.

Walker J and Beard DV FilmPlane2: Design and Implementation, *SPIE medical Image* '90 1990.

Zimmerman JB, Pizer SM, Staab E, Perry J, McCartney W, and Brenton B, "An Evaluation of the Effectiveness of Adaptive Histogram Equalization for Contrast Enhancement," *IEEE Transactions on Medical Imaging*, 7(4): 304-312, 1988.