Scanpaths of motion sequences: where people look when watching movies

Eli Peli, Robert B. Goldstein, and Russell L. Woods,
The Schepens Eye Research Institute & Harvard Medical School, Boston, MA, USA

***Abstract*— Magnification around the most important point of the scene (center of interest - COI) might be an effective aid for people with vision impairments that cause resolution loss. This requires that a COI exist for most video frames. Operationally, we defined the COI by recording the eye movements of normally-sighted subjects as they watched movies. Here we report the frequency that people looked at the same place during the movies and the spatial distribution of their COIs, and investigate age and gender differences.**

***Index Terms*— Eye Movements, Magnification, Scanpath, Video, Visual Aids, Low Vision, Television.**

I. Introduction

People who suffer loss of visual resolution due to eye diseases could benefit from modified information displays. The most common modification used today is magnification. Magnification inherently restricts the field of view and thus may impede the acquisition of information attained in normal vision by the use of scanning eye movements. This problem may be addressed by dynamic control of the displayed information. Dynamic control of large text presentation is helpful for people with low vision [1-4]. We propose a similar approach to improve access to movies and television.

Magnifying moving images using electronic zoom [5] would enable users to select and vary the desired level of magnification from time to time. However, only part of the magnified scene can be presented on the screen. Consequently, large parts of the scene become invisible. Manual zoom-and-roam devices are available in commercial television systems (e.g. DVD players). However, the rapid changing of scenes in most movies may not allow for effective manual control of the magnified section of the image. We proposed pre-selecting the point in the scene on which to center the magnified view (the center of interest - COI) and providing that position with each frame [6, 7]. This selection should maintain the most relevant details in view.

Together with DigiVision (San Diego, CA), we have developed a computer controlled zoom-and-roam device for playback of movies on a television. The computer plays a DVD and simultaneously reads the COI. These coordinates are sent to the zoom and roam device so that the magnified image is centered on the COI coordinates. We proposed using eye movement recordings from normally-sighted observers watching the movie to determine the desired COI. Although other methods of determining the COI can be envisioned, eye movement recording is automatic and objective.

Choosing the COI using eye movements is akin to finding the scanpath for a movie sequence. Much work has been done regarding the scanpath of still images [8-11], but little is known about viewing moving images. With the exception of a few studies [12-14] most development that depends on knowing where the gaze is directed (e.g. compression schemes [15] and transmission of images for limited screen space [16]) assume that most people look at the same place all the time while watching movies. To our knowledge this assumption has not been verified experimentally. Here we quantify the proportion of the time multiple people look at the same place while watching a movie, and begin to examine the effects of age and gender on this behavior.

Film editors have used assumed knowledge of viewer's eye movements - and even blinks - to assemble movies [17]. Stelmach et al. [12] recorded 24 observers viewing 15 forty-five second clips to determine if viewing behavior can be incorporated into video coding schemes. They found that there was a substantial degree of agreement among viewers in terms of where they looked. In a follow-up experiment related to gaze-contingent processing techniques [13], recorded eye movements of subjects were used to create a "predicted gaze position". Tosi et al. [14] recorded the scanpaths of 10 subjects watching a variety of clips totaling about 1 hour and reported that, qualitatively, individual differences in scanpaths were relatively small. Theoretical saliency models [18, 19] predict where people will look and have made no assumptions regarding individual differences in predictions of regions of interest.

Here we address three specific questions relevant to our proposed low-vision aid for viewing television. (1) To what extent do people look at the same place when watching a movie? (2) Does that vary with age and gender? (3) Does the position of the COI differ from the center of the screen?

II. Methods

Six movie clips were selected to span a broad range of scene activity, from stationary newscasters to athletes in motion and

Supported in part by NIH Grants EY05975 and EY12890.

Corresponding author: Eli Peli is at The Schepens Eye Research Institute, 20 Staniford Street, Boston, MA, 02114 USA. (E-mail: eli@vision.eri.harvard.edu).

to appeal to both younger and older audiences (Table 1). The movie clips from DVDs were presented in a 16×9 movie format on a 26.5-inch diagonal NTSC (4×3) monitor as interlaced video at 30 frames per sec. (60 fields per sec.).

Table 1. Category, names, and length of movie clips used.

Category	Title	Time (min)
Sports	Any Given Sunday (1999)	4:12
Comedy	Big (1988)	6:29
Documentary	Blue Planet (2001)	8:14
News	Network (1976)	4:02
Game Show	Quiz Show (1994)	6:40
Drama	Shakespeare in Love (1998)	7:06
	Total per subject	**37:29:00**

26 normally-sighted subjects were seated 46 inches from the screen which spanned a 26.3°×14.8° visual angle. Subjects viewed movie clips while eye movements were recorded with an ISCAN model RK726PCI eye tracking system. The ISCAN had a nominal accuracy of 0.3° over a ±20° range and a sampling rate of 60Hz. Thus we could acquire two eye samples per video frame. The ISCAN compensated for head movements, permitting gaze monitoring without head restraint, and thus allowing a comfortable viewing situation. The ISCAN was calibrated using a 5-point calibration scheme. To optimize tracking [20, 21], we performed a pre-clip calibration (external to the ISCAN) which was repeated before each movie clip was viewed. Recording of the next clip did not proceed until all 5 calibration points had satisfactory (45 samples) data yield. A post-clip calibration was also recorded and the analysis program averaged the pre and post-clip results for the calibration equations.

During the recording phase, immediate feedback was available regarding the amount of valid data available. If less than 80% was valid, the subject's data were not considered for inclusion, and recording stopped. We did not repeat movie clips, as we wanted to record the subject's eye movements during their first viewing of the clip. This happened only with one subject. Of the remaining 25 subjects, the 5 subjects with the best eye movement data yields in each of 4 groups were selected. The 20 subjects were grouped by age and gender: Younger Female 18-29y, Male 16-36y; and Older Female 51-62y, Male 42-66y.

III. ANALYSIS

A. Preprocessing of individual records

The individual subjects' recordings were processed to apply the calibration to the data and remove recording artifacts caused by blinks and other failures. Recording could fail if the head moved too fast or when specular reflections, such as tear film menisci, were erroneously detected by the ISCAN as a cornea reflection. Blinks and loss of tracking were filtered from the file by removal of records containing zero value data or frames where the pupil diameter was out of a set range.

The DirectX 8.1 (Microsoft, Redmond, WA) DVD interface only interrupts the processor every 0.4 to 1.0 second with timing information. Frame numbers between such interrupts were calculated from the elapsed time, assuming a 30 frame per second rate. This procedure resulted, on occasion, in non-monotonic or duplicate frame numbers. Non-monotonic frames were discarded. Because the video and ISCAN data were recorded asynchronously, each assigned frame could be associated with one, two or three eye records. We have designated these multiple records per frame as "subframes" (note – these subframes are not video fields).

Table 2. Yield of acceptable eye samples data from each movie clip did not vary significantly between clips. Yield was greater for male subjects ($F_{1,16}$=7.8, p=0.01) and slightly greater for older subjects ($F_{1,16}$=3.2, p=0.09).

	OM	YM	OF	YF
Sunday	97.2%	95.2%	93.4%	93.0%
Big	96.9%	94.6%	94.8%	92.6%
Blue	97.7%	94.2%	94.6%	93.0%
Network	95.7%	95.6%	93.3%	92.6%
Quiz	96.3%	94.4%	93.6%	91.4%
Shakes	96.0%	95.1%	93.9%	91.8%

B. Merging of eye recordings of multiple subjects' records to find extent of overlap

The 120 subject data files (20 subjects × 6 clips) were processed to count how many of those subjects had valid data for each subframe (Fig. 1).

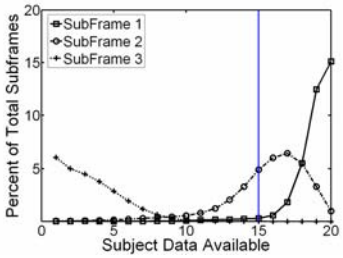

Fig. 1. The percentage of the total number of subframes for which the number of subjects had valid data. The percent of subframes where 15 or more subjects contributed data (vertical line) was 63%, which increased to 85% when data from subframe 3 were discarded.

For each subframe the calibrated (x, y) coordinates of individual subjects gaze points are distributed across the screen. Various methods have been applied to compute the level of coincidence between the gaze points of multiple subjects [11, 12, 22-25]. We chose to calculate the area of the best-fit bivariate contour ellipse (BVCEA) to quantify the degree of spatial coincidence of the eye positions of all the subjects with valid data points. This measure has been used in the past to quantify fixation eye movement stability [2, 26]. The *k* parameter of the BVCEA determines the degree of enclosure of the ellipse. We set *k*=1, for which 63% of the points would have been enclosed by the ellipse.

The cumulative distributions of the BCVEA found for each movie clip (Fig. 2) were fit with a logistic function,

$$y = c + (1-c)/\left(1+\exp\left(-(x-a)/b\right)\right) \qquad (1)$$

where *a* is the mid-point of the function, *b* is related to the rate of rise, and *c* is the lower asymptote. Similarly, to quantify effects of age and gender, the BCVEA was calculated for every subframe for which there were eye position data for at least 4 subjects from each group of 5 subjects. In subsequent

data analyses we used the ½ point and *b*. Data were evaluated using analysis of variance, with movie clip treated as a within-subject factor.

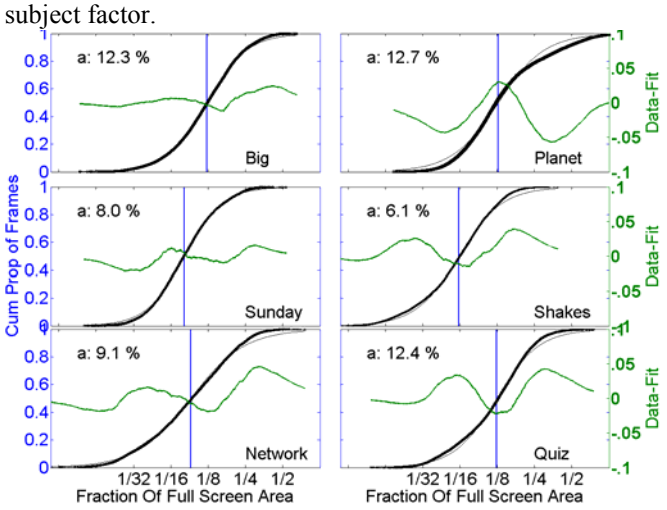

Fig. 2. For each movie-clip subframe, eye position coordinates were used to calculate the BVCEA as a fraction of the full screen area. Only those subframes where 15 or more subjects had data were used. The cumulative curves show the proportion of the total subframes for which the BVCEA was less than a given fraction of movie screen area. Logarithmic transforms of the distributions were fitted to a logistic function (with *c*=0) that were then used to calculate the screen fraction for which 1/2 of the samples had a smaller BVCEA (vertical line and the value indicated by the inset). The residuals of the fits are shown.

IV. RESULTS

As shown in Fig. 2, for all six movie clips, more than 1/2 of the time most of the subjects (15 to 20) looked within an area that was less than 13% of the movie scene. This represents an area equivalent to a circle with a diameter of about 8 deg.

To examine the effects of age and gender on COI we performed analyses of variance on the 1/2-point and *b* of the fits to the data shown in Fig. 3. As seen in Fig. 4, male and older subjects were more likely to look in the same direction (smaller 1/2-point) than female ($F_{1,20}$=6.3, p=0.02) and younger ($F_{1,20}$=22, p<0.001) subjects, respectively. Older subjects were slightly more variable (slower rise – larger *b*) than younger subjects ($F_{1,20}$=3.4, p=0.08). Between the movie clips, there were significant differences in *b* ($F_{5,18}$=4.4, p=0.009) but not of the 1/2 point, indicating that the COI was more variable for some movies. Subjects were more variable in *Network* than *Planet* (p=0.04), Big (p=0.03) and *Sunday* (p=0.004). This shows that, as might be expected, movies with high level of motion more tightly control the observer COI than movies with relatively static scenes.

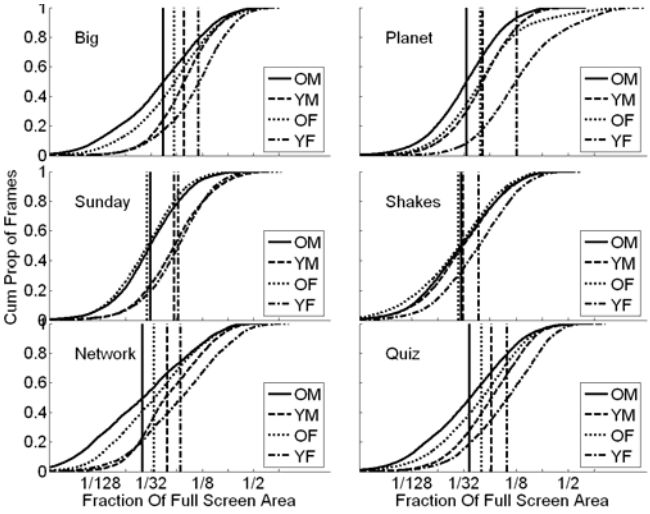

Fig. 3. The cumulative area distributions and positions of the 1/2-point of each clip for each age-gender group (where at least 4 out 5 subjects had useable data). Older and male groups had tighter positions of gaze points.

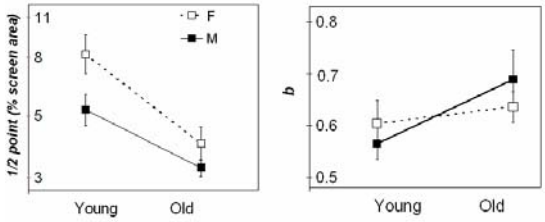

Fig. 4. There were significant effects of gender and age on the likelihood that the subjects in a group looked in the same direction (the 1/2-point); and there was a small, non-significant age effect on the variability of direction of gaze (*b*). Note that the scale labels for the 1/2-point are non-linear, since the fit was done in the logarithmic transform of the area. Error bars indicate SEM.

For those subframes for which there was eye position data for more than 15 of the 20 subjects, the position of the COI was determined as the mean x and y coordinates of the group. The COI distributions (32×24 bins for the letterbox area) are shown in Fig. 5. In general, the peak of the COI distributions were approximately in the center of the movie scene, though they varied by as much as 1/4 of the width or height from center, and the distributions varied between movie clips.

V. DISCUSSION

Measuring and providing the coordinates of the COI along with each frame may allow magnification to be used to its full potential as a low vision aid for watching movies (and other television programs). The eye movement method presented here is a natural and efficient way of determining these COIs. We envision that, just as programs are now being provided in "closed captioned" and described video formats, movies can be provided with these COIs encoded.

We have demonstrated that it is possible to determine the COI in a movie scene by recording the eye movements of normally-sighted observers while they watch a movie. Over

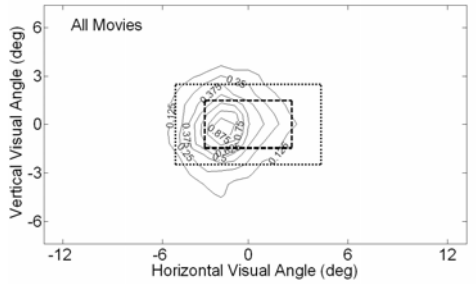

Fig. 5. COI distribution for all movie clips for all subjects (for subframes with eye position data for at least 15 subjects). Although the distribution peaks near the center of the screen, the spread indicates that a large proportion of the time, people did not look at the center of the movie scene. The dotted and dashed regions represent 1/9th (11%) (3X magnification) and 1/25th (4%) (5X magnification) of the screen area, respectively. 73% of COIs lay outside the 4% screen area.

1/2 of the time, the gaze of more than 15 subjects was contained within an area that was less than about 13%) of the movie scene (Fig. 2) or about 5% of the screen for 4 subjects (Fig.4). This is crucial for our application. We rarely expect to need to magnify the image by greater than a factor of 4 (showing 1/16th (6%) of the frame). Higher magnification might cause too much loss of context. The distribution of COIs (Fig. 5) illustrates that magnification centered on the COI would provide more information than magnification simply centered on the center of the movie scene. Also, we found that there are some significant differences in the observation behaviors between gender and age groups. The current analysis only found that the older and male observers' COIs were more tightly grouped than the younger and female observers (Fig. 4). We still need to determine if the COI locations varied with gender and age. Also, conditions or scenes that did not result in a tight COI (i.e. large BCVEA) might be just as interesting as the condition of tight COI.

In addition to our interest in the application of this technique to our movie (or television) magnification device, we see this work as a beginning of an interesting examination of the nature and characteristics of the motion scanpath of dynamic environments — the movie environment being one that is simpler to study — perhaps followed by the dynamic real world of a mobile observer.

ACKNOWLEDGMENT

Assistance was provided by Gang Luo and Shabtai Lerner.

REFERENCES

[1] G. Legge, G. Rubin, D. Pelli, and M. Schleske, "Psychophysics of reading II. Low vision," *Vision Research*, vol. 25, pp. 253-66, 1985.

[2] G. Rubin and K. Turano, "Reading without saccadic eye movements," *Vision Research*, vol. 32, pp. 895-902, 1992.

[3] E. Fine and E. Peli, "Computer display of dynamic text," *Vision 96: International Conference on Low Vision, Madrid, Spain*, pp. 259-67, 1996.

[4] E. Fine and E. Peli, "Benefits of rapid serial visual presentation (RSVP) over scrolled text vary with letter size," *Optometry and Vision Science*, vol. 75, pp. 191-6, 1998.

[5] S. Kazuo and J. Shimizu, "Image scaling at rational ratios for high-resolution LCD monitors," *Society for Information Display Digest of Technical Papers*, vol. 31, pp. 50-53, 2000.

[6] E. Peli, "Vision multiplexing - an engineering approach to vision rehabilitation device development," *Optometry and Vision Science*, vol. 78, pp. 304-315, 2001.

[7] R. B. Goldstein, H. Apfelbaum, G. Luo, and E. Peli, "Dynamic magnification of video for people with visual impairment," *Society for Information Display, Digest of Technical Papers*, pp. 1152-1155, 2003.

[8] L. Stark, "Abnormal patterns of normal eye movements in schizophrenia," *Schizophrenia Bulletin*, vol. 9, pp. 55-72, 1983.

[9] L. Stark, "New quantitative evidence for the scanpath theory: Top-down vision in humans and robotics," presented at First Meeting of the International Society of Theoretical Neurobiology, 1993.

[10] L. W. Stark, K. Ezumi, T. Nguyen, R. Paul, G. Tharp, and H. I. Yamashita, "Visual search in virtual environments," *Human Vision, Visual Processing, and Digital Display III/Human Perception, Performance, and Presence in Virtual Environments*, pp. 577-589, 1992.

[11] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 970-982, 2000.

[12] L. Stelmach, W. J. Tam, and P. Hearty, "Static and dynamic spatial resolution in image coding: An investigation of eye movements," *Human Vision, Visual Processing and Digital Display II*, vol. 1453, pp. 147-152, 1991.

[13] L. Stelmach and W. J. Tam, "Processing image sequences based on eye movements," *Human Vision, Visual Processing, and Digital Display V*, vol. 2179, pp. 90-98, 1994.

[14] V. Tosi, L. Mecacci, and E. Pasquali, "Scanning eye movements made when viewing film: preliminary observations," *International Journal of Neuroscience*, vol. 92, pp. 47-52, 1997.

[15] W. S. Geisler and H. L. Webb, "A foveated imaging system to reduce transmission bandwidth of video images from remote camera systems," NASA, Austin, TX 19990025482; AD-A358811; AFRL-SR-BL-TR-98-0858, 1998.

[16] U. Rauschenbach and H. Schumann, "Demand-driven image transmission with levels of detail and regions of interest," *Computers and Graphics*, vol. 23, pp. 857-866, 1999.

[17] E. Dmytryk, *On Film Editing*. Boston, London: Focal Press, 1984.

[18] X. Marichal, T. Delmot, C.DeVleeschouwer, V. Warscotte, and B. Macq, "Automatic detection of interest area of an image or a sequence of images," *International Conference on Image Processing (ICIP '96)*, vol. 3, pp. 371-374, 1996.

[19] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254-1259, 1998.

[20] A. J. Hornof and T. Halverson, "Cleaning up systematic error in eye-tracking data by using required fixation locations," *Behavior Research Methods, Instruments & Computers*, vol. 34, pp. 592-604, 2002.

[21] D. M. Stampe, "Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems," *Behavior Research Methods, Instruments & Computers*, vol. 25, pp. 137-142, 1993.

[22] D. Salvucci and J. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Eye Tracking Research & Applications Symposium*. Palm Beach Gardens, FL: Association for Computing Machinery, 2000, pp. 71-78.

[23] J. Goldberg and J. Schryver, "Eye-gaze-contingent control of the computer interface: Methodology and example for zoom detection," *Behavior Research Methods, Instruments & Computers*, vol. 27, pp. 338-350, 1995.

[24] L. Stark and Y. Choi, "Experimental metaphysics: the scanpath as an epistemological mechanism," in *Visual Attention and Cognition*, W. H. Zangemeister, H. S. Stiehl, and C. Freska, Eds. Amsterdam; New York: Elsevier, 1996, pp. 3-69.

[25] W. Osberger and A. M. Rohaly, "Automatic detection of regions of interest in complex video sequences," *Proceedings of SPIE. Human Vision and Electronic Imaging VI*, vol. 4299, pp. 361-372, 2001.

[26] G. T. Timberlake, M. K. Sharma, S. A. Grose, D. V. Gobert, J. M. Gauch, and J. H. Maino, "Retinal location of the preferred retinal locus (PRL) relative to the fovea in scanning laser ophthalmoscope (SLO) images," *Optometry and Vision Science*, vol. 82, pp. 177-185, 2005.