Scanpath comparisons for complex visual search in a naturalistic environment

Mary E. Frame¹ · Rik Warren² · Anna M. Maresca¹

Published online: 3 December 2018

© Psychonomic Society, Inc. (This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply) 2018

Abstract

Naturalistic surveillance tasks provide a rich source of eye-tracking data. It can be challenging to make meaningful comparisons using standard eye-tracking analysis techniques such as saccade frequency or blink rate in surveillance studies due to the temporal irregularity of events of interest. Naturalistic research environments present unique challenges, such as requiring specialized or expert analysts, small sample size, and long data collection sessions. These constraints demand rich data and sophisticated analyses, particularly in prescriptive naturalistic environments where problems must be thoroughly understood to implement effective and practical solutions. Using a small sample of expert surveillance analysts and an equal-sized sample of novices, we computed scanpath similarity on a variety of surveillance data using the ScanMatch Matlab tool. ScanMatch implements an algorithm initially developed for DNA protein sequence comparisons and provides a similarity score for two scanpaths based on their morphology and, optionally, duration in an area of interest. Both experts and novices showed equal dwell time on targets regardless of identification accuracy and both samples showed higher scanpath consistency across participants as a function of target type rather than individual subjects showing a particular scanpath preference. Our results show that scanpath analysis can be leveraged as a highly effective computer-based methodology to characterize surveillance identification errors and guide the implementation of solutions. Similarity scores can also provide insight into processes guiding visual search.

Keywords Visual search · Scanpath analysis · Applied research

Background

Variables and Data Collection Tools in Surveillance Research

Naturalistic research provides many opportunities to understand cognitive phenomena in real-life working environments. By examining cognition as it naturally unfolds, it becomes easier to develop a fuller understanding of applied research problems and implement reasonable solutions, but there are challenges that are not typical in laboratory studies. Naturalistic environments require laboratory tasks that are high fidelity to the environment where software and

technologies will be implemented, necessitating a sacrifice of some experimental control. In the real world, a person may engage with a task for hours and seldom experience a key event. For example, a baggage screener may work a full shift and only encounter a few instances of minor violations and never see an instance of a gun or bomb-making materials. Rarely, there may be multiple sequential or simultaneous violations. Furthermore, applied research may require a highly-specialized expert sample that cannot be represented with undergraduates, resulting in a low number of subjects. Real-world tasks might also not have as well-defined goals, such as explicit “correct” solutions. Finally, certain tools may not be permitted or practical to implement, such as scene-recording eye-tracking equipment in a classified research space. These limitations necessitate leveraging cutting-edge analyses techniques.

To promote effectiveness of surveillance screeners—termed *analysts*—behavioral, cognitive, and physiological metrics are used in both controlled-laboratory and real-world environments to assess analyst effectiveness. The goal of this research is to augment the performance of analysts

✉ Mary E. Frame
mary.frame@wright.edu

¹ Wright State University, Dayton Ohio, USA

² Air Force Research Laboratory, Wright-Patterson AFB Ohio, Dayton Ohio, USA

while simultaneously decreasing workload. This paper focuses on the tasks of Eyes-On (EO) analysts who engage in active monitoring of either still images or Full Motion Video (FMV). Their primary task is to identify specific Essential Elements of Information (EEIs) from surveillance FMV over an 8–12 hour shift. Due to the highly visual nature of eyes-on tasks, eye tracking metrics are important as measures of workload, attention, and fatigue. Analyzing eye-tracking data using a variety of methods allows for a deeper understanding of problems that analysts face and provides a means of determining optimal intervention methods and of eliminating less helpful solutions.

Eye tracking metrics such as blink rate and pupil dilation effectively provide information on workload and fatigue, and can subsequently trigger interventions to reduce workload, increase alertness, or do both (Siegle et al., 2008; Stern et al., 1994; Van Orden et al., 2001). Fixation locations and durations serve as markers of attention. Generally, where a person is fixating on a screen for extended periods is highly correlated to what they are attending to (Gaspelin et al., 2017). There are dueling theories as to whether visual attention is captured more by salience of the activity on the screen (Theeuwes et al., 1998; 2003), or if attention is driven by goal motivation (Folk et al., 1992), such that a person will concentrate on goal-pertinent features while searching (Leber & Egeth, 2006). Some theories also try to reconcile the various bottom-up and top-down processes involved in visual search, stating that top-down explanations can explain repetitive eye movements over repeated images, but that this can also be guided by bottom-up processes (Sawaki & Luck, 2010; Gaspelin et al., 2017; Foulsham & Underwood, 2008). This is an important debate, as the solutions implemented to improve the performance of surveillance analysts are dependent on which factors are causing attention-related performance decrements. Within a real-world surveillance setting, both feature-salience and motivational factors are likely relevant and contribute to errors. Visual occlusions, such as a sandstorm blowing by, reduce scene clarity, leading to more errors. Likewise, a highly salient EEI such as a brightly colored vehicle entering a compound might draw attention away from a simultaneously occurring but less salient EEI, such as a person in dark clothes digging on the other side of the road. Top-down errors might include failing to attend to and report non-EEI activity that is still highly relevant to overall mission objectives due to myopic concentration on a predefined EEI list.

Studies of visual search in still images have demonstrated that it may be difficult for even experts to identify task-irrelevant visual anomalies (Drew et al., 2013, 2016, 2017), which adds support to the idea that attention is motivation-driven. For example, inattention blindness studies, such as Drew et al. (2013), have found that expert radiologists

examining X-ray images fixated on and repeatedly backtracked to an embedded task-irrelevant gorilla, but the vast majority did not notice or report the anomaly. This and similar studies show the added value in characterizing the pattern of eye scanpaths above and beyond a simple count of presence/absence within areas of interest (AOIs) or average fixation duration. Scanpaths in inattention blindness tasks demonstrate that analysts may “see” the gorilla, but may not perceive and report it. This contradicts the notion that image features simply need improved salience to increase attention since fixation rate or duration may be similar to correctly classified information in an image. EO analysts may experience inattention blindness to important items whether they are EEIs or not. Knowing when this effect occurs is crucial for implementing aids to improve screeners’ performance on surveillance tasks. Scanpath metrics provide opportunities for prescriptive guidance to improve EO analyst performance and may help distinguish experts’ versus novices’ search strategies to improve training of novice analysts.

Scanpath analysis: ScanMatch

Among the multitude of rich eye-tracking metrics that can be leveraged in applied surveillance contexts is scanpath analysis. A scanpath is defined by the temporal sequence of point-by-point (x,y) screen coordinates of fixations. At minimum, scanpaths encompass one or more full fixation-saccade-fixation sequences (Poole & Ball, 2006). Scanpaths can capture fixation, re-fixation, and backtrack patterns. This in turn can provide useful metrics of analysts’ attention, conscious or otherwise. Comparisons can also be made between scanpaths, as in comparing scanning behavior in a search task between an expert and a novice (Kübler et al., 2015). Research has found that incorporating scanpaths can also greatly improve models’ predictions of fixation locations (Foulsham & Underwood, 2008). Figure 1 shows three plots one might wish to compare for morphological similarity. Assuming that Plot A is the “optimal” scanpath in a visual search, Plot B might be characteristic of an expert scanning the scene and Plot C might be an example of a novice. Although it is visibly clear that Plot B is more similar to Plot A than Plot C, visual inspection alone cannot quantify this difference. Further complicating this quantification are duration differences, for example comparing a 60-second segment of scanpath data to a 15-second segment. One might wish to compare different scanpath lengths based on morphology alone or account for temporal fixation duration differences between the two.

Multiple algorithms have been proposed to characterize and compare fixation sequences between two or more scanpaths, including ScanMatch (Cristino et al., 2010) and

Fig. 1 Three plots of scanpath data. The leftmost image (Plot A) is highly similar to the scanpath of Plot B, but is highly dissimilar to the scanpath of Plot C. All 3 plots were made with an identical number of raw gaze points

MultiMatch (Dewhurst et al., 2012), among other algorithms (Foerster & Schneider, 2013). ScanMatch and MultiMatch are both MATLAB packages that take different approaches to parsing gaze data. ScanMatch utilizes a string-edit distance methodology similar to the Levenshtein distance (Levenshtein, 1966), but with important improvements. Cristino et al. (2010) utilize the Needleman-Wunsch algorithm, which has primarily been implemented in DNA sequence comparison, to align eye movement sequences spatially using either user-specified or gridded Areas of Interest (AOIs). ScanMatch can segment the screen into smaller rectangular areas, up to a 26x26 grid, and label each as an AOI. Although this is the most fine-grain grid resolution, it may not always be the most appropriate segmentation, which we will discuss later in this paper. However, the 26x26 grid allows for a much finer resolution segmentation than many predecessors of ScanMatch. Figure 2 shows how grid AOI is specified in the ScanMatch Graphical User Interface, as well as how other ScanMatch parameters are set. Additionally, ScanMatch is sensitive to temporal as well as spatial similarity. The output of ScanMatch consists

of normalized similarity ratings between paired string comparisons. Figure 3 illustrates the process of inserting substitutions and gaps to determine the similarity of two strings of gaze sequences from a 3x3 grid. Cells of the grid are labeled in arbitrary sequential order from A to I.

In contrast to ScanMatch's string-edit methodology, MultiMatch uses a vector-based approach to eye gaze segmentation. Scanpaths are aligned based on their shape, but the algorithm does not factor temporal similarity based on dwell duration into the overall similarity scoring (Jardzka et al., 2010; Dewhurst et al., 2012). Although MultiMatch does align the sequence based on temporal order, ScanMatch additionally can factor duration of each element within a sequence. Instead of outputting a single similarity score, MultiMatch outputs five scores: 1) Vector Similarity, 2) Length, 3) Direction, 4) Position, and 5) Duration. This method provides greater detail in spatial scanpath structure. This makes MultiMatch well-suited for analyses with specific predictions, but less well-suited for exploratory analyses. ScanMatch has distinct strengths for analyzing data from a naturalistic visual-search task in an

Fig. 2 The main GUI of ScanMatch, where one can specify the resolution of the screen the eye data was collected on as well as the AOI grid resolution, gap penalty, etc. If there are no pre-specified AOIs, such as might be expected in a scanning task, the screen can be segmented into an up to 26x26 grid

Original Sequence

Before Event	A	B	C	A	C	B	D	E	D	F	G	H	I	H	B	C	D	E
After Event	A	C	A	B	D	E	B	A	F	G	H	I	D	E				

Optimized Result: 6 gaps added, 1 substitution

Before Event	A	B	C	A	C	B	D	E	D	--	F	G	H	I	H	B	C	D	E
After Event	A	--	C	A	--	B	D	E	D	A	F	G	H	I	--	--	--	D	E

Fig. 3 Illustration of the logic of substitutions and gap insertion in the ScanMatch algorithm. Although there are many ways to transform the bottom sequence to match the top sequence, the algorithm attempts to make as few changes as possible. Gaps can be more or less penalized by varying the gap penalty. With a harsher penalty for gaps, the algorithm favors making substitutions. If adding gaps is not penalized, then inserting gaps will be incentivized

applied research environment. This methodology is particularly suited to exploratory analyses where one might not have a predicted direction of effect (Cristino et al., 2010). Both algorithms represent state-of-the-art parsing tools in their respective methodologies (i.e., string-edit comparison versus vector-based comparison). For the purposes of our experiment, only ScanMatch is used due to the exploratory nature of these applied analyses and due to the potential noise from mobile eye tracking.

Experiment 1: Scanpaths of expert surveillance analysts

The first experiment implemented scanpath analysis on a small data set of expert surveillance analysts in a high-fidelity overwatch task. Consistent with real-world mission execution, participants were tasked with identifying EEIs and reporting them as they were spotted by pressing a button and recording a brief message using speech-to-text software. The goal of our scanpath analyses was to diagnose what gaze patterns characterize analyst performance failures. We do this by testing some common hypotheses generated in the field: 1) Failures in classification are due to failing to see the EEI in time to identify it, 2) Failures in classification are due to changing search strategies to a less efficient path, 3) Emulating the search strategy of the highest performing expert should contribute to better performance, and 4) Search strategies change to adapt to differences in the EEIs.

Assumptions about the above hypotheses can directly lead to implemented solutions, sometimes without much testing of their validity. However, using scanpath analysis allows us to test all of these proposed hypotheses directly. Additionally, we tested ScanMatch under a variety of AOI grid resolutions and gap penalties to determine the robustness of our findings to differing parameterizations. Due to the screen resolution and relative size of the EEIs, we hypothesized that higher-resolution AOI grids would be more sensitive to meaningful differences in scanpaths than

a coarser grid resolution, which may not be sensitive to variations or inefficiencies in scanpaths.

Experiment 1 method

Tasks, software & scenarios

Analysts viewed FMV and identified EEIs. See Fig. 4 for an illustration of the interface. They were instructed to provide detailed call-outs, spoken descriptions of an EEI, while pressing a speech-to-text button when an EEI occurred on screen. Information recorded in call-outs included the time, description of activity, and a slant count consisting of the total number of men, women and children present. For example, a slant count of 1/2/3 indicates 1 male, 2 females, and 3 children present in the video. Analysts used the Real-Time Annotation and Dissemination (RTAD) tool, a software tool developed in-house to emulate a real-world security overwatch environment. RTAD possesses an additional suite of processing tools that allows for easy annotation and dissemination of video screenshots. RTAD permits a user to watch FMV, designate an important area of an image by clicking and dragging (known as “chipping” or annotating), edit the designated EEI on the chipped scene, commit the image and metadata to a MySQL database, and disseminate the resulting product via e-mail (as a Microsoft PowerPoint file). In Fig. 4 these tools are on the right hand side. RTAD is accessible from a Chrome web browser on desktops, laptops, tablets, and smartphones. Analysts used RTAD to create annotated images, screenshots with EEIs indicated within a red box created by clicking and dragging.

Analysts viewed two simulated security overwatch FMVs created using Meta-VR. Meta-VR is visual simulation software for creating 3D, high fidelity, and geographic-specific scenarios using high-quality gaming graphics. The 30-minute scenarios simulated an overhead surveillance view of a compound where people were frequently gathering, entering, and exiting. Both videos were rendered in 1080p. In each video, 27 total EEIs needed to be identified. Three of these EEIs involved simultaneously occurring events, such as two to three people entering or exiting the compound together. However, these were reported as single events by participants and are therefore classified as single EEIs, leaving a total of 21 differentiated EEIs for each scenario. All non-overlapping events occurred between 11 and 174 seconds apart. The high variability in EEI occurrence is typical of real-world observation environments. A still image from one of the scenarios is included in Fig. 4. The EEIs specified for analysts to identify in both scenarios were:

- 1) People entering or exiting the compound

Fig. 4 An example screenshot of an EEI from one of the two scenarios developed in-house using Meta-VR (Boydston et al., 2018). In this image two people enter the compound, which is chipped by clicking and dragging

- 2) Vehicles stopping and dropping off or picking up people near the compound
- 3) Weapon retrieval or weapon exchanges between people in or around the compound

Analysts used Speech-to Text for Enhanced PED (STEP) to transcribe the verbal call outs of EEIs they had identified in the FMV. STEP is a suite of tools developed by the US Air Force Research Laboratory (AFRL) and Ball Aerospace and Technologies Corp to aid in Processing, Exploitation, and Dissemination (PED) of Intelligence, Surveillance, and Reconnaissance (ISR) FMV. This tool recognizes, records, and transcribes utterances spoken by an analyst. Analysts were instructed to choose a push-to-talk (PTT) key on the keyboard prior to beginning the experiment. To make a verbal call-out, analysts held down the PTT key while speaking and released the key when finished. After release, STEP creates a text transcription and logs the call-out, the time stamp of the PTT key press, and the response time.

All eye-tracking data were collected using Tobii Glasses 2, sampled at 50 Hz. All eye tracking data were collected in a consistently well-lit environment that simulated a standard workspace for a surveillance task. Each analyst viewed two screens. The left screen displayed an FMV in RTAD and the right screen contained an Internet Relay Chat window and either a visualization window of the speech-to-text software STEP, or a Powerpoint slide with a reminder of the EEIs for the task.

Participants/analysts and experimental procedure

Data was collected from 9 expert ISR analysts with surveillance experience. All were previously trained in making verbal call-outs (e.g., making slant counts, reporting Zulu time, etc.) and were comfortable with the task procedure. One

expert analyst's data could not be analyzed due to recording errors in the speech-to-text and behavioral metrics.

All analysts received a short training including a Power-Point presentation describing the task and user interface, then engaged in self-paced practice for 5 to 10 minutes. The practice video allowed analysts to become familiar with the RTAD chipping tool and STEP. After training, analysts donned a set of Tobii Glasses 2 and underwent a short calibration procedure. Analysts then sat at separate stations to watch the first surveillance video at an average distance of 58.64 cm from the screen. Each screen in a 2-monitor setup was 54x31 cm with a pixel resolution of 1920x1080. Analysts were instructed to either identify listed EEIs using only verbal call-outs with STEP (single-task condition), or to make both call-outs and chip images by dragging the cursor to make a box around EEIs on screen (dual-task condition). Instructions were counterbalanced across the two scenarios. After completing the first surveillance task, analysts filled out the Standard Usability Scale (SUS) (Brooke et al., 1996) and the NASA-TLX (Hart & Staveland, 1988; Hart, 2006) to measure subjective workload. After survey completion, analysts began the other surveillance task with the opposing instructions to the first task.

Eye-tracking data-cleaning procedure

Prior to analysis, eye tracking data was plotted on a common coordinate system. The Tobii Glasses 2 projects gaze points in a three-dimensional coordinate space by default and, naturally, the head position of each analyst relative to the screen differed. Although it is optimal to position the participant directly and squarely in front of a monitor, for this experiment, data was collected using a dual-screen setup. Since analysts were positioned between these two screens, there was a slanted visual angle for both screens,

making the coordinates in two-dimensions project to a trapezoid rather than rectangular screen.

To analyze two-dimensional scanpaths projected onto the screen in a typical Cartesian coordinate plane, the data was standardized via a cleaning procedure. Tobii Analyzer's automated gaze mapping uses pattern analysis of the ongoing video and still image of the scene, ascribing fixations to a snapshot image corresponding to locations on the video screen. These mappings were vetted afterwards by an experimenter. After mapping gaze projections, the coordinates for the corners of the screen on the snapshot were computed and coordinates within those bounds were transformed to a common coordinate framework. Following coordinate standardization, the data were segmented based on EEI events. Three segments from each scenario involved simultaneously occurring EEIs (e.g., groups of people exiting compound with little spatial dispersion). An AOI was generated for each segment based on where the event occurred on the screen with a visual angle of 9.5 degrees. Segments began at the start of an EEI to 10 seconds after the EEI appeared, which is typical for military surveillance tasks.

Comparisons of interest

Though ScanMatch is a valuable tool for making scanpath comparisons, the normalized similarity-score results can be influenced by fluctuations in user-adjustable parameters. Although these changes might not affect relative between-group differences, they certainly influence absolute similarity scores. To illustrate the influence of grid size using extreme values, imagine two scanpaths like those in Fig. 5. With an extremely coarse grid resolution, such as a 2x2 grid, these would be given a similarity score of 1, indicating that they are identical. However, even with a cursory look at these scanpaths it is clear that they are not identical and should not be quantified as such. Likewise, a 26x26,

maximally granular resolution, might be excessively punitive to small ocular movements or variability due to noise. Not all minor differences may be cognitively meaningful within a particular task environment. As such, the resolution of the grid should be granular enough to detect meaningful differences, but not so granular as to lead to spurious differentiation.

To test the effect of grid resolution, we tested four granularity levels. Each level maintains the relative proportion of the 1920x1080 screen resolution such that each cell of the substitution matrix is relatively square. All similarity analyses were conducted at a resolution of: 6x3, 10x6, 20x11, and 25x14 AOI segments.

In addition to testing these resolution levels, the gap penalty was also varied to either penalize or not penalize for sequential timing differences. In ScanMatch, a Gap Penalty (GP) of 0 indicates that adding gaps will lead to lower similarity scores. Smaller GP values inflict a higher penalty for gaps, whereas higher numbers are more lenient in regard to adding gaps to unequal-length strings. By contrast, GP equal to 1 indicates that there is virtually no gap penalty and thus adding gaps will not strongly impact similarity scores. As such, we expect similarity scores to be higher in the GP = 1 parameterizations compared to the GP = 0 parameterizations.

Experiment 1 results

Behavioral results

Correct identification of EEIs was defined as providing a call-out (or annotation in the secondary task condition) within 10 seconds of the EEI appearing on screen. An EEI is classified as incorrect if analysts took longer than 10 seconds to respond, or did not respond at all. Analysts were highly accurate

Fig. 5 The left panel illustrates a potential problem when scanpaths are plotted on a grid with too coarse a resolution. Despite the vast differences in morphology of these two scanpaths, their sequence is identical by quadrant over time. Consequently, they would be characterized as being identical. The right panel illustrates a potential problem from a grid resolution that is too granular. In this segmentation, portions of the scanpath trace that only deviate from one another by 50–60 pixels

will be characterized as different and require a substitution. Although this distinction may be appropriate on certain tasks with fine details, on a larger surveillance task the segment from both lines might represent tracking of the same object in the FMV and thus would more accurately be classified as the same. The most appropriate grid resolution will be heavily influenced by the task EEIs or AOIs and as such grid resolution should be planned accordingly

at making call-outs, with ($M = 80.89\%$, $SD = 5.76\%$) accuracy in the single-task condition and ($M = 82.47\%$, $SD = 7.93\%$) when they simultaneously made annotations, with no significant primary task accuracy differences. There were greater but non-significant performance differences between scenarios regarding annotations, with ($M = 77.93\%$, $SD = 18.38\%$) on Scenario 1 and a lower, more variable score of ($M = 33.27\%$, $SD = 23.36\%$) on Scenario 2. There was no significant difference in mean response times between Scenario 1 ($M = 5.99$ sec, $SD = 2.92$) and Scenario 2 ($M = 6.41$ sec, $SD = 2.47$). However, response times for making call outs were significantly higher when also making annotations ($M = 7.45$ sec, $SD = 2.58$) versus only call-outs ($M = 4.95$ sec, $SD = 2.12$), $t(8) = 4.19$, $p < .01$.

AOI analysis results

Determining where analysts were looking while identifying EEIs allows for a richer understanding of both when and why categorization errors occur. Failing to correctly identify an EEI can be due to focusing on the wrong area of the screen, thus never having the opportunity to see the EEI. Alternatively, analysts may look directly at the EEI for a similar duration as during correct trials, indicating inattention blindness akin to results of Drew et al. (2013). If a stimulus is ambiguous or unexpected, there may be a signal detection error. The AOI around each EEI was defined with a diameter of 350 pixels, accounting for approximately 9.5 deg of visual angle. For both scenarios, there were no significant differences in AOI duration on incorrect versus correct responses and there were no significant differences in time to first fixation on correct versus incorrect responses (Fig. 6). These results indicate that expert analysts visually attended to incorrectly identified EEIs as quickly, and for the same duration, as to correctly identified EEIs, but failed to accurately categorize them, congruous to inattention blindness experiments. Performance errors reflect a failure in the decision making process rather than basic sensory processes. This is useful for the design of intervention strategies to augment analyst performance. One way to improve EO analyst performance might be to reduce

the ambiguity of on-screen events' importance via training in controlled task environments. Another possibility is to develop AI software that can learn patterns associated with and alert analysts about potential EEIs.

Scanpath similarity within subject

By accuracy

Before comparing scanpaths between subjects, analyses were performed to determine the degree of scanpath consistency within-subject throughout the full duration of the surveillance task. Analyses were performed to determine the degree of internal scanpath consistency when comparing two correct trials (CC), two incorrect trials (II), and pairs of trials where one EEI was correctly identified and the other was not (CI). This was done to determine if differences in scanpath morphology led to meaningful differences in accuracy. If analysts' strategy changed on incorrect trials in a way that was suboptimal, we would expect to see a high degree of similarity on CC and on II trial comparisons but a significantly lower similarity score in the CI condition. However, if the degree of similarity is relatively invariant across comparisons, then search strategies most likely do not differ as a function of behavioral accuracy.

The left panel of Fig. 7 shows that regardless of parameterization, there were no significant differences in Scenario 1 scanpath similarity based on accuracy. This indicates that on Scenario 1 there is no evidence of visual search strategies varying as a function of behavioral accuracy. There were also no significant differences, regardless of ScanMatch parameterization, for Scenario 2.

By EEI characteristics

Next, scanpaths were compared within subject based on EEI content, either a vehicle or a human. Due to the differences in characteristics of vehicles and people, such as visual size, we wanted to compare scanpaths when analysts looked at humans versus vehicles. Since there were fewer EEIs involving vehicles in both scenarios, the Vehicle-Vehicle

Fig. 6 The left plot shows no significant difference in cumulative fixation duration within AOIs for correct and incorrect trials. The right plot illustrates no significant difference in the time to first fixation on

correct versus incorrect trials. These results indicate that on incorrect trials, analysts were visually attending to EEIs, but were not able to correctly identify them

Fig. 7 Similarity of scanpaths between Correct-Correct (CC) pairings, Incorrect-Incorrect (II), and Correct-Incorrect (CI) for Scenario 1 (left panel) and Scenario 2 (right panel) under all eight ScanMatch parameterizations. There were no significant differences

and Human-Human comparisons were aggregated into one category (Congruous) and compared to Vehicle-Human pairings (Incongruous). If there is no difference in mean scanpath similarity between the congruous and incongruous conditions, this indicates that analyst scanpaths are consistent regardless of the EEI content. If there are significant differences when comparing EEI type such that congruous pairs have a higher similarity score, this indicates that there are scanpaths characteristic of search based on the content of the EEI, and these scanpaths are distinct from one another.

The left panel of Table 1 shows that for all grid resolutions and gap penalty parameterizations, the Congruous stimuli between EEIs had a higher scanpath similarity score than Incongruous stimuli for Scenario 1. For Scenario 2, all parameterizations with the exception of the two most granular resolutions with a gap penalty imposed, were significant (right panel of Table 1). Effect sizes, measured using Cohen's *d*, are much larger for Scenario 1. However, both Scenarios show a robust and large effect, meaning the difference between groups is not likely to be spurious. Even for the non-significant Scenario 2 parameterizations, the effect size is still moderately large. Figure 8 provides

an additional summary. This pattern of results denotes that scanpaths were consistent based on EEI type, but that the pattern of ocular movement can be distinguished for scanning a human versus scanning a vehicle, and this pattern is robust across analysts.

Scanpath similarity between subjects

ScanMatch was first run on all between-subjects pairings, matched on trial (within scenario) for each of the ScanMatch parameterizations. The first descriptive-level comparison was whether analysts were more consistent with their own scanpath strategies, or if there was greater consistency between expert analysts matched on trials. If within-subject similarity scores are higher than between-subjects similarity parings, this would suggest that analysts have a more consistent search strategy in a surveillance task. However, if the between-subjects similarity scores paired on the same trials are higher, this would suggest that the analysts examine specific EEIs in a similar manner to other analysts, but unique to EEI. For Scenario 1, there was a significant difference between the within- and between-subjects similarity scores for each parameterization. The top panel of Table 2 shows the mean similarity scores for each condition as well as the *t* and *p* values. For each parameterization, the between-subjects similarity scores were higher than the within-subjects similarity scores. Likewise with Scenario 2 (see bottom panel of Table 2), between-subjects similarity scores were significantly higher on four of the eight parameterizations. For the other parameterizations there were no significant differences, but all means were in the same direction. The implication is that there is more regularity for how experts scan particular EEIs than there is for how an analyst scans an FMV feed overall. Figure 9 illustrates similarity when comparing pairs of trials within-subject versus matched trials between subjects.

After comparing scanpath similarity scores within and across analysts, a series of between-subjects comparisons was conducted. First, similarity scores was separated based on whether both analysts correctly identified the EEI, both incorrectly identified the EEI, or if one analyst correctly classified the EEI while the other did not. It was hypothesized that analysts who were both correct or both incorrect would have higher similarity scores than analysts who were incongruous in respect to accuracy. This pattern would indicate dissimilarities in search strategies that could lead to disparities in correctly identifying EEIs. However, for each of the ScanMatch parameterizations in Scenario 1, there were no significant differences between similarity scores on congruous (both correct or incorrect) versus incongruous trials (all *p* > .05). This seems to indicate that similarity scores do not vary as a function of analyst accuracy, which may mean that visual search strategies are

Table 1 Result of paired comparisons between Congruous and Incongruous EEIs and scanpath similarity scores for Scenario 1 and Scenario 2

Scenario 1 Similarity by EEI Type					Scenario 2 Similarity by EEI Type				
Condition	Mean Diff	T	p value	Cohen's d	Condition	Mean Diff	T	p value	Cohen's d
GP0Grid25x14	0.0733	5.402	0.001*	2.067	GP0Grid25x14	0.0483	1.406	0.202	0.479
GP1Grid25x14	0.1015	10.476	0.000*	4.924	GP1Grid25x14	0.0634	3.411	0.011*	1.483
GP0Grid20x11	0.0664	6.455	0.000*	2.235	GP0Grid20x11	0.0716	2.196	0.064	0.681
GP1Grid20x11	0.0908	11.697	0.000*	5.503	GP1Grid20x11	0.0723	3.771	0.007*	1.462
GP0Grid10x6	0.0995	7.900	0.000*	2.779	GP0Grid10x6	0.0772	2.384	0.049*	1.065
GP1Grid10x6	0.0936	12.192	0.000*	4.922	GP1Grid10x6	0.0718	3.819	0.007*	1.580
GP0Grid6x3	0.1530	11.072	0.000*	4.325	GP0Grid6x3	0.0889	3.014	0.020*	1.841
GP1Grid6x3	0.1051	11.117	0.000*	4.832	GP1Grid6x3	0.0748	4.467	0.003*	2.143

similar when analysts are accurate and inaccurate. Although none of the results were significant for Scenario 1, they were generally in the predicted direction, and similarity scores were higher between congruous accuracy pairings than incongruous pairings (see top panel of Table 3 for mean differences, *t*-statistics, *p*-values, and effect sizes for each ScanMatch parameterization). For Scenario 2 however,

there were significant differences between congruous and incongruous similarity scores. The congruous similarity scores ($M = .426$, $SD = .191$) were significantly higher in the 25x14 resolution condition with gap penalty than incongruous similarity scores ($M = .381$, $SD = .198$), $t(533) = -2.062$, $p = .04$. The other 25x14 resolution condition with no gap penalty also saw congruous scores ($M = .771$, $SD = .087$) significantly higher than incongruous similarity scores ($M = .745$, $SD = .111$), $t(533) = -2.434$, $p = .015$. There was also a significant difference in congruous ($M = .529$, $SD = .191$) and incongruous similarity scores ($M = .504$, $SD = .186$) in the 10x6 condition with no gap penalty $t(533) = -2.065$, $p = .039$. All of the non-significant parameterizations mean differences were in the predicted direction (see bottom panel of Table 3).

Prescriptive strategies in applied environments often take the form of recommending behavioral strategies that emulate a high performing individual, which may or may not generalize to overall performance improvements for other analysts. It may be assumed that a higher performing analyst is using a more adaptive search strategy. Examining behavioral data alone is insufficient to inform whether this solution strategy will work in practice, but scanpath analysis can provide additional insight. Indeed, if the best performing analyst is using the most adaptive search strategy, one would expect that there is at least a moderate correlation between similarity of scanpath to the highest performer and behavioral performance. After gleaning pertinent overall between-subjects comparisons, similarity scores were analyzed between the analyst with the highest behavioral accuracy, Analyst 4, and all other analysts. Analyst 4 had a combined accuracy score of 84% on the call-out task across both scenarios and a score of 100% on the annotation task, for a combined task accuracy of 92%. First, the degree of similarity was calculated using ScanMatch between Analyst 4 and all other analysts on each trial, to determine the degree of similarity across experts.

In both scenarios, there was a significant difference in scanpath similarity to Analyst 4 between analysts, indicating

Fig. 8 Similarity of scanpaths between EEIs where EEIs were of the same type (vehicles or humans) versus incongruous pairings for Scenario 1 (left panel) and Scenario 2 (right panel) under all eight ScanMatch parameterizations. There are robust significant differences between Congruous and Incongruous EEIs, indicating that scanpath varied as a function of EEI characteristics

Table 2 Comparison of within-subject similarity scores and between-subject similarity scores for each parameterization of Scenario 1 and Scenario 2

Scenario 1 Degree of Similarity Within and Across Subjects					
Parameterization	Within	Between	t	p	Cohen's d
GP0Grid25x14	0.3071	0.4221	5.251	0.001*	2.763
GP1Grid25x14	0.6837	0.7664	12.238	0.000*	5.540
GP0Grid20x11	0.3064	0.4147	4.048	0.004*	2.540
GP1Grid20x11	0.6680	0.7816	12.928	0.000*	7.644
GP0Grid10x6	0.4187	0.5533	5.858	0.001*	3.885
GP1Grid10x6	0.6879	0.7879	10.701	0.000*	6.060
GP0Grid6x3	0.5890	0.7522	11.927	0.000*	7.358
GP1Grid6x3	0.7249	0.8409	11.899	0.000*	7.165

Scenario 2 Degree of Similarity Within and Across Subjects					
Parameterization	Within	Between	t	p	Cohen's d
GP0Grid25x14	0.3944	0.4206	0.437	0.675	0.283
GP1Grid25x14	0.7372	0.7719	1.663	0.140	0.983
GP0Grid20x11	0.3740	0.4352	0.962	0.368	0.641
GP1Grid20x11	0.7143	0.7954	3.067	0.018*	1.980
GP0Grid10x6	0.4736	0.5252	0.971	0.363	0.672
GP1Grid10x6	0.7325	0.7969	2.674	0.032*	1.683
GP0Grid6x3	0.6372	0.7230	2.771	0.028*	1.918
GP1Grid6x3	0.7655	0.8355	3.251	0.014*	2.137

that the sample is not homogeneous in use of search strategy. Table 4 summarizes ANOVAs run on all ScanMatch parameterizations. Nearly all of the ANOVAs are significant, indicating that the scanpaths of at least one analyst significantly differs from the scanpath strategy implemented by the highest performing analyst. These differences are significant when there is a gap penalty present (nearly all GP = 0 are significant) indicating that temporal dynamics and scanpath morphology differ.

It was hypothesized that if there was a significant correlation between analyst scanpaths, then analysts who have a search strategy more similar to Analyst 4 will have higher behavioral accuracy. By extension, a useful intervention might be to train analysts to take a similar search strategy to Analyst 4 to improve performance. However, there were no significant correlations between scanpath similarity to the highest scoring analyst and behavioral accuracy on corresponding scenarios. In an applied environment, non-significant or null results can be useful, by suggesting that potential intervention strategies are unlikely to work, saving time and resources. In this case, the results indicate that a simple intervention strategy of training analysts to emulate the highest performing expert would be insufficient to improve performance. For this surveillance task, there is likely no single prescriptive optimized search strategy, but rather a combination of individualized

interventions should be implemented. The goal of these methods is to determine useful predictions and correlates of performance that can eventually be parsed in real-time to improve analyst performance.

Experiment 2: Comparison with a novice sample

We were interested in comparing the results of the expert analysts with a group of similarly-aged novices. When searching a visual scene, novices most likely implement a fairly entropic scanning strategy. Scanpath analysis can allow us to determine if experts implement a more consistent search strategy than non-experts. We conducted similar analyses to those performed in Experiment 1.

Disentangling the top-down versus bottom-up nature of the task should become clearer with a comparison to non-expert visual search in the same task structure. If top-down (listed EEI characteristics) are driving visual search more strongly than bottom-up saliency features (e.g., the larger size of stimuli such as a vehicle), we might expect more consistency within subject and between trials for experts. This would be indicative of a particular expertise search strategy. However, we would expect to see, by contrast, more inconsistency of search for novices. If novices also

Fig. 9 These plots compare the consistency of scanpaths across all parameterizations both within-analysts (i.e., comparing an individual's scanpaths during different EEIs to one another) and across-analysts (i.e., comparing scanpaths during the same EEIs for different analysts). Most of these comparisons show that similarity scores were significantly higher during the same EEIs than they were within analysts. Significant differences are marked with an asterisk

rely on more bottom-up processing, we might expect to see higher similarity scores based on congruous stimuli type (vehicles or humans). To provide a richer and more thorough comparison, we replicated our initial study on a novice sample.

Experiment 2 method

Data was collected from 8 novice participants (equal to the number of expert analysts in the previous study) with no experience in surveillance. Participants performed the identical experimental procedure as in the Experts' Method section. Eye-tracking cleaning procedures were identical as well.

Experiment 2 behavioral results

Behavioral accuracy on the primary call-out task was not significantly different between experts and novices. Examining novices' data alone, there was no significant difference between call-out accuracy when it was a single task ($M = 83.33\%$, $SD = 8.44\%$) or one of two concurrent

tasks ($M = 80.95\%$, $SD = 8.44\%$). As with the expert sample, there were greater performance differences between scenarios regarding annotations, with ($M = 70.83\%$, $SD = 14.60\%$) on Scenario 1 but a lower and more variable score of ($M = 61.48\%$, $SD = 41.08\%$) on Scenario 2. Again, due to low sample size this difference was not significant.

There was no significant difference in mean response times for novices between Scenario 1 ($M = 3.10$ sec, $SD = 1.63$) and Scenario 2 ($M = 3.08$ sec, $SD = 1.88$). There were also no significant differences in response time in the single task condition ($M = 2.83$ sec, $SD = 1.20$) versus the dual task condition ($M = 3.35$ sec, $SD = 2.14$). Response-time scores indicated that novices responded more quickly than experts robustly. Experts responded significantly slower for both Scenario 1, Mean Difference (Expert - Novice) = 2.889, $t(14) = 2.444$ sec, $p < .05$, and Scenario 2, Mean Difference = 3.329 sec, $t(14) = 3.026$, $p < .05$. Experts were also significantly slower to respond when in a single task, Mean Difference = 2.118 sec, $t(14) = 2.583$, $p < .05$, and when managing dual tasks, $t(14) = 4.103$, $p < .01$. Although somewhat counterintuitive, this delay in responding could be due to greater deliberation by experts prior to identifying an EEI. This increased deliberation period did not seem to correspond to higher behavioral accuracy scores however.

Experiment 2 novice AOI results

AOI analyses were conducted to determine if novices spent a similar duration focusing on EEIs when they were correctly versus incorrectly identified and if the time to first fixation was significantly different based on accuracy. For Scenario 1, observers spent significantly more time fixating on the EEI when they correctly responded ($M = 17.7\%$, $SD = 11.3\%$) than when they failed to identify the EEI ($M = 11.6\%$, $SD = 11.0\%$), $t(7) = 1.89$, $p = .03$. However, for Scenario 2 there was no significant difference in time spent fixating on the EEI for correct or incorrect responses. For both scenarios, there was no difference in time to first fixation. With the exception of the first scenario, this pattern of fixation duration results is consistent with the AOI results for the expert analysts, indicating that AOI measures are not robustly diagnostic of identification accuracy (Fig. 10). The results overall lend additional credence that most errors involve classification rather than visually missing the event, but there may be a greater impact of fixation duration on accuracy for novices.

There were a few interesting differences between expert analysts and novices. One difference is that the novices show somewhat higher variance for AOI fixation duration and time to first fixation. Additionally, on Scenario 1, novices spent a considerably shorter duration in the AOI than experts, both when correct, Mean Difference (Expert -

Table 3 Top table: Scenario 1 similarity between analysts by accuracy congruence Bottom table: Scenario 2 similarity between analysts by accuracy congruence

Scenario 1 Congruous vs. Incongruous Accuracy Pairings				
Condition	Mean Diff	t	p	Cohen's d
GP0Grid25x14	-0.0024	-0.127	0.899	0.013
GP1Grid25x14	-0.0269	-1.754	0.082	0.213
GP0Grid20x11	-0.0073	-0.344	0.731	0.034
GP1Grid20x11	-0.0202	-1.900	0.058	0.170
GP0Grid10x6	-0.005	-0.298	0.765	0.029
GP1Grid10x6	-0.0031	-0.441	0.659	0.045
GP0Grid6x3	0.0055	0.493	0.622	-0.050
GP1Grid6x3	-0.0012	-0.152	0.88	0.015

Scenario 2 Congruous vs. Incongruous Accuracy Pairings				
Condition	Mean Diff	t	p	Cohen's d
GP0Grid25x14	-0.0448	-2.062	0.04*	0.231
GP1Grid25x14	-0.0254	-2.434	0.015*	0.255
GP0Grid20x11	-0.0442	-1.77	0.077	0.202
GP1Grid20x11	-0.0166	-1.574	0.116	0.181
GP0Grid10x6	-0.0247	-1.15	0.251	0.131
GP1Grid10x6	-0.0202	-2.065	0.039*	0.233
GP0Grid6x3	-0.0034	-0.221	0.825	0.025
GP1Grid6x3	-0.0158	-1.614	0.107	0.178

Novice) = 16.4%, $t(14) = 9.645$, $p < .001$ and incorrect, Difference = 18.1%, $t(14) = 10.662$, $p < .001$. However, for Scenario 2 novices spent significantly more time in AOI when correct, Mean Difference (Expert - Novice) =

−10.5%, $t(14) = -4.024$, $p < .001$, but time in AOI was much more comparable regardless of expertise for Scenario 2, Difference = −2.5%, $t(14) = -1.018$, $p = .326$. These results are somewhat inconclusive, but the largest and most

Table 4 ANOVA results of comparisons between the highest scoring analyst and all other analysts for Scenario 1 and Scenario 2

Scenario 1 Comparison with Subject 4			Scenario 2 Comparison with Subject 4		
Condition	F	p value	Condition	F	p value
GP0Grid25x14	6.927	0.000*	GP0Grid25x14	10.057	0.000*
GP1Grid25x14	5.565	0.000*	GP1Grid25x14	2.02	0.067
GP0Grid20x11	5.432	0.000*	GP0Grid20x11	6.825	0.000*
GP1Grid20x11	4.796	0.000*	GP1Grid20x11	2.163	0.051
GP0Grid10x6	2.818	0.012*	GP0Grid10x6	9.927	0.000*
GP1Grid10x6	4.094	0.001*	GP1Grid10x6	2.79	0.014*
GP0Grid6x3	1.415	0.212	GP0Grid6x3	3.699	0.002*
GP1Grid6x3	2.536	0.023*	GP1Grid6x3	2.884	0.011*

robust differences are in Scenario 1 demonstrate that novices seem to spend less time fixating on the AOI than experts.

Initial fixations were later for novices than analysts on Scenario 1 when correct, Mean Difference = -988.77 ms, $t(14) = -6.385$, $p < .001$, no significant difference when incorrect, Difference = 128.46 ms, $t(14) = 0.547$, $p = .593$. However, there do seem to be differences by scenario, since for Scenario 2, the opposite pattern occurred. There were no significant differences in time to first fixation when subjects were correct, Mean Difference = 311.2 ms, $t(14) = 1.794$, $p = .098$, but novices were slightly faster to look at the EEI when they did not identify the stimulus, Mean Difference = 457.0 ms, $t(14) = 2.627$, $p < .05$. Even when mean values were comparable, there was higher variability for novices. This may be indicative that experts are more adept at attending to visual features consistent with EEIs when classifications are accurate, but novices attend to EEI features more quickly when EEIs are not correctly classified. However, results were not entirely conclusive and demonstrate that there may be characteristics of the visual scene that influence these patterns, even when those visual scenes are extremely similar.

Experiment 2 scanpath results

For the novice observer results, we chose to focus on a single ScanMatch parameterization. The 20×11 grid resolution was chosen as a sufficiently high resolution grid based on visual angle and the $GP = 0$ was chosen to penalize somewhat for temporal differences.

Analyses were replicated with a novice sample to elucidate whether visual search was more guided by bottom-up visual features or top-down goal motivation. This provided a further opportunity to compare these factors to expert scanpaths. As with the expert sample, there were no significant differences in ScanMatch similarity scores based on

accuracy. Figure 11 shows a sample scanpath from an EEI correctly identified by one participant and incorrectly by another, each superimposed over a still image from the trial.

Again, as with the experts, there was a significant difference in similarity score between Congruous EEIs ($M = .43$, $SD = .07$) and Incongruous EEIs ($M = .28$, $SD = .06$), $t(7) = 12.544$, $p < .001$ in Scenario 1 and ($M = .37$, $SD = .03$) as well as for Incongruous EEIs ($M = .34$, $SD = .02$), $t(7) = 2.805$, $p = .013$ in Scenario 2. Figure 12 illustrates the results of both of these analyses for the novices.

Between-subjects analyses were conducted to determine the degree of similarity of scanpaths on alike EEIs across participants. Participants had significantly more consistent scanpaths to one another in Scenario 2 ($M = .45$, $SD = .06$) than in Scenario 1 ($M = .50$, $SD = .05$), $t(7) = 5.85$, $p < .001$. As with the experts, novices had higher similarity scores to one another during the same EEIs than they did during different EEIs within participant. This indicates that search strategies were more likely a function of scene features rather than a particular subject's consistently preferred strategy (Fig. 13). In Scenario 1, similarity between subjects was .048 higher than within subjects, $t(7) = 4.34$, $p = .001$, Cohen's $d = .780$, and in Scenario 2, the difference between subjects was even greater, .141 points higher than within subjects, $t(7) = 4.84$, $p < .001$, Cohen's $d = 3.389$.

Finally, scanpath similarity comparisons were made between the novice and expert samples for the $GP = 0$, 20×11 grid parameterization. There were no significant differences between experts and novices for each pairing of correct/incorrect and congruous vs incongruous EEIs for both scenarios. This indicates that regardless of expertise, participants seemed to adapt scanning behavior to be consistent based on stimuli characteristics and with little variability as a function of accuracy and that these

Fig. 10 On Scenario 1, observers spent significantly more time fixating on correctly identified EEIs. For Scenario 2 there were no significant differences in fixation duration based on accuracy. For both scenarios, there was no significant difference in time to first fixation

based on accuracy. On incorrect trials, even observers were looking at the EEI, they failed to identify them. The large error bars indicate that there is high variability in time to first fixation and duration in AOI across novice observers

Fig. 11 Sample scanpath overlaid on a screenshot of the scenario (Boydston et al., 2018) when an EEI occurred, from an analyst who correctly (top) and incorrectly identified this specific EEI (bottom). In this example, the EEI is two people exiting the compound. As can be seen from this case, both analysts spent some time looking at the entrance of the compound, but the path for the incorrect trial appears more erratic

characteristics are not due to expertise. There were, however, interesting albeit difficult to interpret differences between experts and novices in regard to within and between subject similarity. For Scenario 1, there were no significant differences in between-subject similarity scores between experts and novices. However, novices did show higher scanpath consistency within subject ($M = .364$, $SD = .029$) than experts did ($M = .306$, $SD = .028$), $t(13) =$

4.035, $p < .01$. For Scenario 2 this pattern was the opposite, with no significant differences for within subject scanpath similarity, but novices had significantly more consistent scanpaths between subject on matched EEIs ($M = .505$, $SD = .031$) than experts ($M = .435$, $SD = .053$), $t(13) = 3.219$, $p < .01$. This pattern of results seems to indicate that novices took a more consistent individual strategy for Scenario 1 than experts, but relied on EEI

Fig. 12 Overall results by accuracy (left plot) and by stimulus type (right plot) for the novices. As with the experts, there were no significant differences in ScanMatch similarity scores based on whether observers correctly reported the EEI or not. There were however,

significant differences on both scenarios, with higher consistency in scanpaths when comparing two EEIs with the same type of stimulus (person or vehicle)

Fig. 13 These plots compare the consistency of scanpaths within-analysts (i.e., comparing an individual's scanpaths during different EEIs to one another) and across-analysts (i.e., comparing scanpaths during the same EEIs for different analysts). For both scenarios, similarity scores for novices were significantly higher during the same EEIs than they were within-analyst. This is consistent with the results for the expert sample

features to guide search more consistently than experts in Scenario 2.

Discussion

Taken together with behavioral data, eye tracking provides a rich data source in a surveillance environment. Scanpath analyses allow for more detailed understanding of this data. This can provide a catalyst for tailoring follow-up metrics and interventions to provide the greatest improvement using minimal resources. In this experiment, basic AOI analyses demonstrate that analysts and novices fixated at areas of the screen where unidentified EEIs were located. This indicates that errors were a result of a failure to categorize the event as an EEI, which is more likely to be a failure of behavioral pattern recognition rather than an issue of image salience such as insufficient brightness.

Above and beyond simple aggregated eye-tracking metrics, scanpath analysis using ScanMatch provided a richer analysis of the data both within and between subjects. Eye-scan strategy did not significantly change as a function of accuracy within subject, for example. This indicates that errors did not likely occur due to sudden changes to a more inefficient scanpath strategy. Along with the AOI results, it is clear that analysts saw the appropriate EEIs both when correct and incorrect. Scanpaths varied significantly as a function of EEI type, indicating that observers followed vehicles and humans in a consistent and distinguishable manner. There may be something adaptive about changing search method when a different type of EEI is present, using bottom-up perceptual features to guide search. These results

elucidate that there seem to be both bottom-up and top-down factors leveraged differentially.

For both experts and novices, there was more consistency between analysts matched on EEI compared to the similarity within subject. Although this seemingly counterintuitive, these results indicate that EEI features, especially taken with the within-subjects analyses, elicit a similar pattern of responses regardless of expertise. This also illustrates that analysts are not just persisting with a single strategy across all EEIs, but are rather adapting their search strategies based on the specific events they are monitoring. Direct comparisons between experts and novices were inconclusive in regard to reliance on scenario characteristics between Scenario 1 and 2. Taken together though with AOI results, it seems that Scenario 1 may differ fundamentally from Scenario 2. For Scenario 1, experts showed a faster time to first fixation, lending credence to learned strategy guiding visual search. However, for Scenario 2, novices showed a faster time to first fixation as well as more consistent strategies based on stimulus characteristics than experts. The opposite pattern of results on each Scenario demonstrates that specific content may need to be probed, even when scenarios are designed to be highly similar. There may be bottom-up background characteristics that influence visual search, indicating that in real working environments, mission characteristics should be well-understood to inform findings.

In a real-world environment, one proposed solution to improve performance is to train people to behave more similarly to a better performing expert. Interestingly, the comparisons with the best performing expert show that there is no correlation between behavioral performance and degree of similarity with the scanning behavior of the most expert analyst. An intervention or augmentation strategy that seeks to improve scanning efficiency by emulating the best performer is unlikely to improve performance across analysts.

This assessment of ScanMatch probed the effect of parameterization on results. Using AOI grids that are too coarse or too granular may lead to an over or under-inflation of similarity scores, underscoring the importance of testing the robustness of results under different grid resolutions. Manipulating the gap penalty allowed for comparisons between scanpaths based exclusively on scanpath morphology versus differences of morphology and temporal components. As expected, similarity scores were consistently higher for all of the non-gap penalty-imposing conditions compared to conditions where both morphology and temporal dynamics were taken into account. Likewise, coarser grid resolutions yielded overall higher similarity scores. For most analyses within scenario, results tended to be significant or non-significant across parameterizations. The only set of results that deviates from this pattern are the within-

versus between-subjects similarity score comparisons. The division appears to be based on the grid resolution. Finer grid resolutions yielded non-significant results compared to coarser grids. Cohen's d scores were moderately high for all conditions even when there were no significant differences between similarity scores (as seen in Table 2). This illustrates a potentially robust effect that is penalized by the combination of the increased grid resolution, possibly under-inflating scores and simply having insufficient power due to low sample size.

Naturally, no single method of analysis provides a one-size-fits-all solution and scanpath analysis alone is insufficient for developing an augmentation aid to improve analyst performance. However, used in combination with other eye-tracking metrics, it is a powerful and robust tool for better classifying the possible cognitive hurdles analysts face during surveillance search tasking.

Future directions

This initial effort provided an opportunity to utilize scanpath analysis in a real-world complex task, as well as vet the parameterization of ScanMatch. ScanMatch has demonstrated value for use in naturalistic research and is adaptable to the specific challenges of applied research. Real-surveillance research has the limitation of small sample sizes due to requiring specific expertise. Despite the challenge of low statistical power, scanpath analysis using a tool like ScanMatch allows for richer analysis of a limited data set. In addition to being appropriate for small- n analyses, it can also be easily adapted for larger data sets from applied research environments, such as eye tracking data from full shifts, via batch processing in a supercomputer.

Furthermore, we are interested in using information from these experiments to develop algorithms that can diagnose potential problems and inefficiencies in search strategies. Interventions that can diagnose in real-time if an analyst is searching in a novice manner, or in a manner that indicates fatigue or overwork, would be extremely helpful for analyst augmentation. Another challenge of real-world environments that is difficult to capture inside the laboratory is the ambiguity of “truthing” real-world unfolding events. When observing in real-time, analysts and their supervisors don't know the “correct” or “incorrect” responses. However, by characterizing eye movements under correct versus incorrect conditions, perhaps we can get information that transfers to a more ambiguous environment such as indicators of inattention, inefficiency, or cognitive overwork. Further work will involve developing real-time scanpath analysis algorithms that can help characterize potential problems when accuracy cannot be determined.

As this task was more exploratory, we determined that ScanMatch would be the most appropriate scanpath comparison package to employ. However, there are many other algorithms that might be able to provide richer information about spatial scanpath similarity, such as MultiMatch. Although the present study is a fairly small-scale analysis, the rich data produced by one or more methods of scanpath comparison holds tremendous value in applied research environments.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Boydston, A. S., Maresca, A. M., Saunders, E., & Stanfill, C. (2018). *Real-time annotation and dissemination tool (RTAD) demo*. Dayton: Air Force Research Laboratory.
- Brooke, J. et al. (1996). SUS A quick and dirty usability scale. *Usability Evaluation in Industry*, 189(194), 4–7.
- Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). Scanmatch: a novel method for comparing fixation sequences. *Behavior Research Methods*, 42(3), 692–700.
- Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior Research Methods*, 44(4), 1079–1100.
- Drew, T., Vö, M. L. H., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattention blindness in expert observers. *Psychological Science*, 24(9), 1848–1853.
- Drew, T., & Stothart, C. (2016). Clarifying the role of target similarity, task relevance and feature-based suppression during sustained inattention blindness. *Journal of Vision*, 16(15), 1–9.
- Drew, T., Boettcher, S. E., & Wolfe, J. M. (2017). One visual search, many memory searches: an eye-tracking investigation of hybrid search. *Journal of Vision*, 17(11), 1–10.
- Foerster, R. M., & Schneider, W. X. (2013). Functionally sequenced scanpath similarity method (funcsim): Comparing and evaluating scanpath similarity based on a task's inherent sequence of functional (action) units. *Journal of Eye Movement Research*, 6(5), 1–22.
- Folk, C. L., Remington, R. W., Johnston, J. C., et al. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 1030–1044.
- Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, 8(2), 6–6.
- Gaspelin, N., Leonard, C. J., & Luck, S. J. (2017). Suppression of overt attentional capture by salient-but-irrelevant color singletons. *Attention, Perception, and Psychophysics*, 79(1), 45–62.
- Hart, S. G. (2006). NASA-Task load index (NASA-TLX); 20 years later. *In Proceedings of the human factors and ergonomics society annual meeting*, 50, 904–908.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, 52, 139–183.
- Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 211–218).

- Kübler, T., Eivazi, S., & Kasneci, E. (2015). Automated visual scan-path analysis reveals the expertise level of micro-neurosurgeons. In *Miccai workshop on interventional microscopy*.
- Leber, A. B., & Egeth, H. E. (2006). It's under control: Top-down search strategies can override attentional capture. *Psychonomic Bulletin & Review*, 13(1), 132–138.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707–710.
- Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. *Encyclopedia of Human Computer Interaction*, 1, 211–219.
- Sawaki, R., & Luck, S. J. (2010). Capture versus suppression of attention by salient singletons: Electrophysiological evidence for an automatic attend-to-me signal. *Attention, Perception, and Psychophysics*, 72(6), 1455–1470.
- Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology*, 45(5), 679–687.
- Stern, J. A., Boyer, D., & Schroeder, D. (1994). Blink rate: a possible measure of fatigue. *Human Factors*, 36(2), 285–297.
- Theeuwes, J., Kramer, A. F., Hahn, S., & Irwin, D. E. (1998). Our eyes do not always go where we want them to go: Capture of the eyes by new objects. *Psychological Science*, 9(5), 379–385.
- Theeuwes, J., De Vries, G. J., & Godijn, R. (2003). Attentional and oculomotor capture with static singletons. *Attention, Perception, and Psychophysics*, 65(5), 735–746.
- Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors*, 43(1), 111–121.