An Eye Movement Analysis Algorithm for a Multielement Target Tracking Task: Maximum Transition-Based Agglomerative Hierarchical Clustering

Ziho Kang and Steven J. Landry

Abstract—An algorithm was developed to characterize, compare, and analyze eye movement sequences that occur during visual tracking of multiple moving targets. When individuals perform a task requiring interrogating multiple moving targets, complex and long eye movement sequences occur, making sequence comparisons difficult in whole and in part. The developed algorithm characterizes a sequence by hierarchically clustering the targets that an individual interrogated through an unordered transition matrix created from the frequencies of eye fixation transitions among the targets. Then, the resulting sets of clustered targets, which we define as multilevel visual groupings (VGs), can be compared with analyze performance. The algorithm was applied to an aircraft conflict detection task. Eye movement data were collected from 25 expert air traffic controllers and 40 novices. The task was to detect air traffic conflicts for easy, moderate, and hard difficulty scenarios on simulated radar display. Experts' and novices' multilevel (level one composed of pairs, and level two composed of three or four targets) VGs were aggregated and visualized. Chi-square tests confirmed that there were significant differences for easy (level one: $p < 0.001$, level two: $p = 0.004$), moderate (level two: $p = 0.047$), and hard (level two: $p < 0.001$) difficulty scenarios. The algorithm supported identifying different eye movement characteristics between experts and novices. Scans of the experts had multilevel VGs around the conflict pairs, whereas those of the novices included different aircraft. The results show promise for using the compact representation of eye movements for performance analysis.

Index Terms—Eye movement, eye tracking, scanpath, target tracking task.

I. INTRODUCTION

E YE movement analysis can be used to understand the underlying cognitive processes and to analyze performance of an individual or groups of individuals. Thus, eye movement
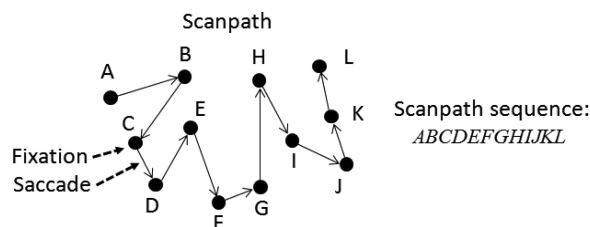
Fig. 1. Example of a scanpath and how it is characterized as a scanpath sequence.

analysis has drawn attention from diverse fields, including business marketing [1], target tracking [2], driving [3], [4], viewing of geographic information [5], drawing [6], and reading [7]–[9].

However, eye movement analysis has focused primarily on static visual tasks, meaning that the targets (e.g., elements such as a banner in a website, a side-view mirror, or a word in a sentence) or areas of interest (AOI) [10] are spatially fixed on a display. An eye movement analysis method, such as an attention map (or heatmap) [11], [12] can show where fixations occurred and their durations, but it is limited in that the time-ordered sequence of eye movements composed of fixations and saccades are not evaluated.

A collection of time-ordered eye movements composed of eye fixations and saccades is called a "scanpath." To determine a fixation, a threshold between 50 and 100 ms is typically used based on the characteristics of a task [5], [13], [14]. A saccade is a fast linear movement that shows the transition from one fixation to another. A scanpath can be analyzed by 1) drawing the obtained scanpath over an interrogated display or 2) transforming the scanpath into a string of enumerated types, as shown in Fig. 1. This string of enumerated types is also referred to as a "scanpath sequence."

Analyzing scanpaths for types of tasks that involve the interrogation of dynamically moving targets is a challenging issue, since the physical position of the targets changes over time. The issue may become more complicated if the scanpaths are complex and long. Therefore, for certain types of analyses, analyzing the scanpath sequences may be more helpful than analyzing the spatial layout of the scanpaths.

Scanpaths or scanpath sequences differ among individuals [15], meaning that large deviations can exist among them; therefore, it is challenging to compare them in whole and in part. Scanpaths have been compared using the string edit algorithm (SEA) [16]–[18], ScanMatch method [13], sequential pattern

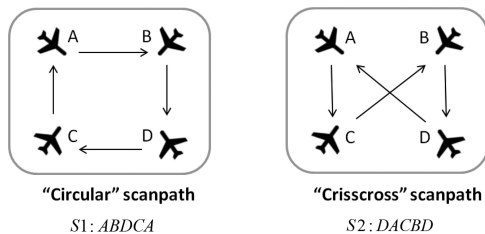"Circular" scanpath   "Crisscross" scanpath

S1: *ABDCA*   S2: *DACBD*

Fig. 2.   Scanpath examples.

mining (SPAM) algorithm [19], dot-plot method [20], binomial tests of Markov matrices [4], distance-based method [21], [22], and multimatch or vector-based approach [23], [24].

Each of these scanpath comparison methods has unique strengths, explained below; however, the methods are difficult to apply to a multielement target tracking task. Consider a scanning task of four aircraft (see Fig. 2). If the scanpaths were short (based on only a few targets (aircraft) on a display), we may obtain circular or crisscross types of scanpaths. However, imagine dozens of moving targets on the display. With more targets on the screen, the scanpaths can become much more complex making it difficult to effectively characterize and compare them. Barriers for such a task include: 1) no predefined starting or ending eye fixation points exist; and 2) large deviations across complex and long scanpaths from different individuals.

To address these issues, the maximum transition-based agglomerative hierarchical clustering (MTAHC) algorithm is introduced in order to characterize, compare, and analyze performance using the number of transitions among multiple moving targets from long and complex scanpath sequences obtained among individuals. MTAHC creates multilevel visual groupings (VGs), defined as sets of targets that have higher number transitions between or among them compared with other targets. Then, the multilevel VGs can be mapped to cognitive processes to support performance analysis.

The rest of the paper is constructed as follows. The background section provides details on 1) existing scanpath comparison methods, and 2) a multielement target tracking task. Then, the concepts of VG and multilevel VG are explained, followed by the MTAHC algorithm. The algorithm is then applied to a simulated aircraft conflict detection task. This is followed by the experimental results and a discussion.

II. Background

A. Scanpath Comparison Methods

The most widely appreciated scanpath comparison method is the use of SEA [16]–[18]. For scanpath sequences composed of characters, SEA calculates the minimum number of edits through insertion, deletion, and substitution in order to transform one sequence to another. Software has been developed to automate the pairwise comparisons [25].

Consider three sequences composed of four targets $\{A, B, C, D\}$.

S1: ABDCA
S2: ABDCB
S3: DACBD

Within the pair $\{S1, S2\}$, only one substitution is required (either substituting the last "B" in $S1$ with "A," or the last "A" in $S2$ with "B"), whereas pairs $\{S2, S3\}$ and $\{S3, S1\}$ require every character to be substituted to transform one sequence to the other. Using sequence similarity indexes, SEA can show which sequences are more similar.

The ScanMatch method [13] enhances the concept of SEA by comparing sequences based on the Needleman–Wunsch algorithm. Instead of counting the number of edits (by insertion, deletion, or substitution), a similarity score is computed through a "substitution matrix" that assigns positive values for a match and negative values for a mismatch. The benefit of this method is that the values within the substitution matrix can be adjusted to accommodate relationships between certain targets or their features, such as physical distances or color; however, the comparison results may change based on how the analyst adjusts the matrix.

An issue with SEA and ScanMatch is that the number of transitions from one state to another is not considered. If many transitions exist between or among certain targets compared with others, it may imply that an individual is highly attentive to certain relationships or characteristics among those targets based on the individual's cognitive process, as supported by the goal-oriented or top-down perspective [26].

The SPAM algorithm has been applied to extract the occurrences of the same sequences among long scanpath sequences [19]. The SPAM algorithm is an efficient pruning mechanism that creates a lexicographic sequence tree [27]. For example, with two elements "a" and "b," the searched sequence "ab" expands to searching "aba" and "abb," etc.

Another novel method, the dot-plot method [20], aligns two sequences to create a dot-plot for any matching eye fixation locations. If multiple matches exist, then the dots will form into different forms of lines that represent sequences, and linear regression can be carried out to find similar sequences. The dot-plot method has advantages, such as providing a visual representation of which sequences share similarities, and does not require a cost (or a substitution) matrix.

Both the SPAM and the dot-plot methods can be effective methods for comparing scanpaths; however, both methods do not consider the number of transitions among all the targets that may provide important eye movement information when pairs or groups of moving targets are interrogated from the set of multiple moving targets.

A viable method that can address the above issues is through using the Markov matrix that represents the probability of eye fixation transitions from one target to another [4]. Underwood and colleagues evaluated each transition probability value within the Markov matrix through binomial tests. Given a number of targets n, the method assumes that the expected transition probability from one state to another is $P_1 = 1/(n-1)$. Given the observed transition probability from state A to B as P_2, the method compares P_1 and P_2.

The binomial tests within the Markov matrix were able to effectively distinguish the eye movements of expert and novice drivers [4]; however, several limitations exist. The expected transition probability may not be true, and it is difficult to apply
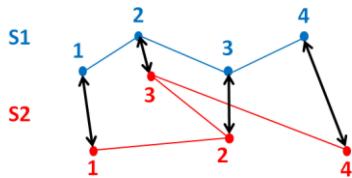
Fig. 3.   Distance-based scanpath comparison.

the method if there are no transitions from one state to another. Moreover, since each transition value within the matrix is evaluated, many binomial tests are required [28]. For example, if there are 20 targets, then a 20×20 transition matrix is created, which requires 400 binomial tests per matrix.

One may consider comparing the transition matrices instead of comparing each value in the transition matrix. A pilot study was performed to apply the Mantel test to compare whole matrices [29]. The method worked well for four to six targets, but as the number of targets increased, sensitivity suffered due to too many zero values (no transitions) in the matrices compared with the amount of transition data.

Other scanpath comparison methods have focused on the comparing physical distance between the scanpaths [21], [22]. These methods use Euclidian distances to compare scanpaths; however, the order of fixations is not considered. Fig. 3 shows an example, in which the second fixation of S_1 is matched with the third fixation of S_2.

The Multimatch approach [23], [24] uses five dimensions, namely shape, length, direction, position, and duration, to compare whole scanpaths. It is challenging to make comparisons across different dimensions, and multimatch provides a weighting method. An interesting aspect of this method is that saccades, which show transitions from one state to another, are divided into vectors. These "saccade vectors" are compared to find differences between the scanpath shapes.

The multimatch approach enhances eye movement analysis through the use of multiple dimensions; however, the distance-based method and the multimatch method are challenging to apply if the targets are moving on the display. Specifically, the shapes of the scanpath would change according to the moving targets. Moreover, if the sequences are long, scanpath shapes may not be apparent (discussed next).

B. Multielement Target Tracking Task to Detect Aircraft Conflicts

A conflict detection task involves identifying near-future conflicts of pairs of aircraft among multiple moving targets. Therefore, a conflict detection task is at least partially a complex multielement target tracking task that provides an interesting environment to investigate human cognitive processes. The task is: 1) dynamic, meaning that targets are moving; 2) time critical, meaning that the detection must be performed prior to any actual conflict occurring in the near future; and 3) uncertain, meaning that a conflict may or may not actually occur. For aircraft conflict detection task, air traffic controllers need to select priorities and manage their cognitive resources to maintain a high level of performance [30].
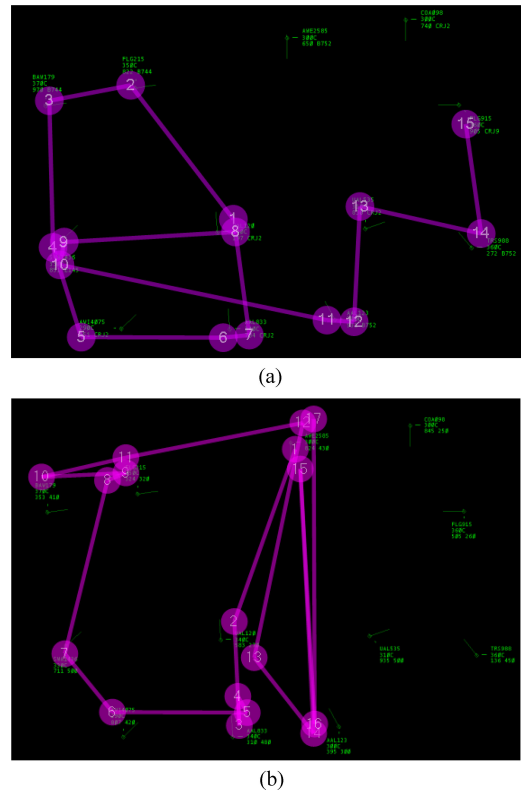

Fig. 4.   Scanpaths for two individuals for a 5-s period overlaid on a still image. (a) 5-s scanpath example. (b) 5-s scanpath example.

Fig. 4 provides examples to illustrate how scanpaths occur when multiple targets are being observed. In Fig. 4, there are multiple aircraft (moving targets): each target consists of the aircraft itself (small diamond shape), its direction (line that stems out from the aircraft), and its data block, which contains aircraft information. The circles indicate the eye fixations, the numbers within the circles indicate the order of eye fixations, and the lines connecting the circles indicate saccades. Note that in Fig. 4, the scanpaths were captured only for 5 s. The scanpaths were overlaid on a still image of the initial layout of the aircraft. Therefore, as time progressed, all aircraft moved in the defined directions.

Large deviations exist among scanpaths. When collecting data from a participant for a long period, there are no real starting or ending points, and there are no guidelines for which path to follow among the multiple targets. Issues become clearer when long sequences need to be compared. Fig. 5 shows a full scanpath of one individual with a duration of approximately 3 min. Note that all aircraft are moving and, thus, the fixations shift according to the current aircraft location.

Fig. 5 illustrates that a scanpath is very complex, making it difficult to analyze or compare. It may be possible to divide the scanpaths based on short-time frames to minimize the movements of targets, but the time threshold to use is not clear. Furthermore, in a pairwise conflict detection task, fixation order within a pair may not matter. For example, for a pair composed of two targets $\{A, B\}$, it may not matter whether A is observed first followed by B or *vice versa*.
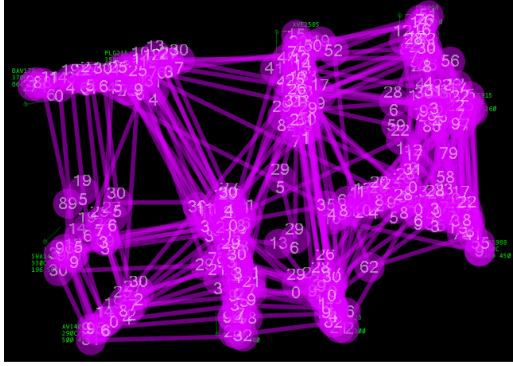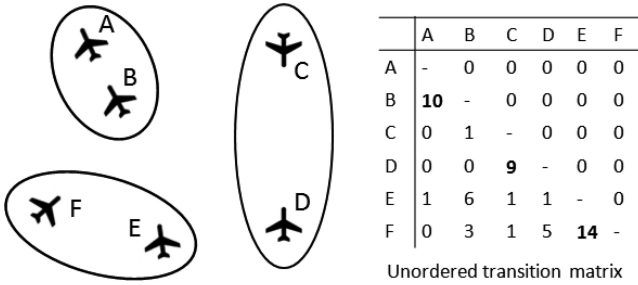
Fig. 5.   3-min scanpath example.

Fig. 6.   Example of VG sets.

To address the mentioned issues, a different approach is required to effectively characterize and analyze the scanpaths obtained from a conflict detection task.

III. CONCEPT OF VISUAL GROUPING SETS

This study introduces the concept of a VG. A VG set is a set of targets having higher transitions between or among the targets compared with other targets. The idea is that for a long scanpath, we want to understand which transitions occur more often among the targets. The ultimate goal is to support mapping the VG sets to cognitive processes.

Fig. 6 shows a simplified example. In Fig. 6, six moving aircraft are depicted. Letters (A, B, C, D, E, F) are assigned to aircraft for the purpose of expressing the transitions as a matrix, as shown on the right. This matrix is an unordered transition matrix, meaning that the transitions from state A to B and from state B to A are summed. One can see which aircraft pairs the individual interrogated. Higher numbers of transitions are highlighted in the matrix, and the respective VG sets ($\{A, B\}, \{C, D\}, \{E, F\}$) are represented as ovals surrounding pairs of aircraft.

As the task is to detect conflicts among multiple targets, an individual may look back and forth between aircraft in conflict [31]. Thus, an individual would likely form VG sets around the pairs that may have conflicts. Therefore, VG sets may be used to analyze the performance of the task.

However, there are two issues to address: How the VG sets are determined and how the remaining transitions are considered within the matrix? Only considering the few highlighted

values within the matrix shown in Fig. 6 would not be correctly evaluating the full-transition matrix. Therefore, we should be able to consider other lower transition values in order to characterize and evaluate the overall eye movements. These issues are addressed next.

A. Determining Visual Grouping Sets

Let $S = \{s_1, s_2, \ldots s_n\}$, where S represents the set of targets, and S_n represents the nth target. Then, the number of transitions from one target to another can be represented as $t(s_i, s_j)$, and the set of number of transitions between two targets can be represented as

$$T = \{t(s_i, s_j) | 1 \leq i < j \leq n, i \in \mathbb{N}, j \in \mathbb{N}\}. \qquad (1)$$

From T, find a VG set composed of two targets $\{s_a, s_b\}$, where $t(s_a, s_b) = \max T$.

Cluster the two targets: $\{s_a, s_b\} \rightarrow s_{ab}$.

Reconstruct S by inserting s_{ab} and deleting $\{s_a\}$ and $\{s_b\}$

$$S' = \{s_1, s_2, \ldots, s_{ab}, \ldots, s_n\} - \{s_a\} - \{s_b\}. \qquad (2)$$

Note that S' is composed of $n - 1$ targets. For simplicity, the targets in (2) can be expressed as

$$S' = \{s'_1, s'_2, \ldots, s'_{n-1}\}. \qquad (3)$$

Then, the next set of number of transitions between two targets (T') from S' is

$$T' = \{t(s_i, s_j) | 1 \leq i < j \leq n - 1, i \in \mathbb{N}, j \in \mathbb{N}\}. \qquad (4)$$

The process iterates for finding the next VG set.

How VG sets are determined is an adaptation of the agglomerative hierarchical clustering (AHC) algorithm. A detailed survey of various AHC algorithms is provided in [32]. The adaptations to the AHC are that: 1) maximum transition values are used to cluster the targets; 2) no geographically fixed coordinates are required to calculate the distance matrix that the original AHC algorithm requires; and 3) the hierarchical clusters are classified into multilevel VGs, which are shown next. If the same maximum values exist, the described process solves them sequentially.

B. Multilevel Visual Grouping sets

When following the process of obtaining the VG sets, all numbers of transitions within the matrix are used to form larger clusters resulting in multilevel VG sets (see Fig. 7). The numbers of targets do not necessarily group from two to four or from four to eight. For example, VG set $\{A, B\}$ from Fig. 7 can cluster with $\{C\}$ into $\{A, B, C\}$. Therefore, level 1 VGs are always composed of two targets. However, level 2 VGs can be composed of three or four targets, and level 3 VGs can be composed of targets between five and eight, etc.
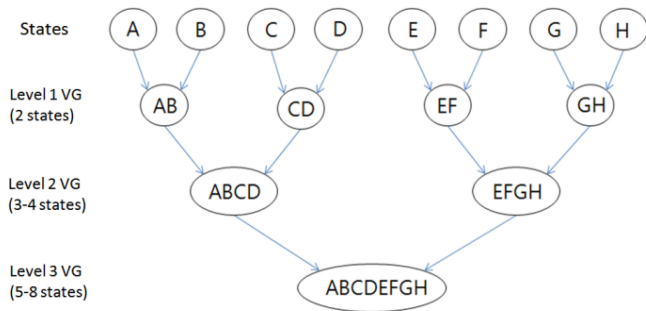
Fig. 7.   Multilevel VG sets.

	A	B	C	D	E	F	G	H	I	J	K	L
A	-	-	-	-	-	-	-	-	-	-	-	-
B	8	-	-	-	-	-	-	-	-	-	-	-
C	0	5	-	-	-	-	-	-	-	-	-	-
D	0	0	4	-	-	-	-	-	-	-	-	-
E	0	0	1	0	-	-	-	-	-	-	-	-
F	0	0	0	1	15	-	-	-	-	-	-	-
G	0	0	0	0	6	2	-	-	-	-	-	-
H	0	0	2	0	4	5	1	-	-	-	-	-
I	0	0	0	1	8	7	3	9	-	-	-	-
J	2	1	6	13	0	0	0	0	0	-	-	-
K	2	3	10	5	0	2	0	0	2	4	-	-
L	6	9	1	0	0	0	0	0	0	0	2	-

Unordered transition matrix

VG level 1: {E,F}, {D,J}, {C,K}, {H,I}, {B,L}
VG level 2: {A,B,L}, {E,F,H,I}, {C,D,J,K}
VG level 3: {E,F,G,H,I}, {A,B,C,D,J,K,L}
VG level 4: {A,B,C,D,E,F,G,H,I,J,K,L}

Fig. 8.   MTAHC transition matrix example.

IV. MAXIMUM TRANSITION-BASED AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

The MTAHC algorithm is explained followed by an example of how multilevel VGs are obtained.

MTAHC Algorithm

Step 1: Obtain an unordered transition matrix from a scanpath.
Step 2: Iteratively cluster the targets using the single linkage function

$$T(C_i, C_j) = \max_{x \in C_i, y \in C_j} t(x, y) \qquad (5)$$

where C_i and C_j are clusters, and $t(x,y)$ is the number of transitions between the targets in the clusters.
Step 3: Assign levels based on the size of the clusters (multilevel VG sets).
Step 4: (For comparing between groups) Compare proportions of obtained VG sets from the total obtained VG sets at each level of VG.

Fig. 8 shows an example of how an observed long scanpath is transformed into an unordered transition matrix and how multilevel VGs are obtained. In the matrix, the higher transitions are highlighted in red, forming level 1 VG sets. The next higher transitions are highlighted in blue, forming level 2 VG sets. Evaluating the transition matrix becomes challenging as the number of targets increases, and the MTAHC method effectively shows the sets of targets that have a higher number of transitions.

V. METHODS

A proof of concept experiment was conducted in part to find out whether the MTAHC algorithm can be used to analyze performance by characterizing scanpaths with multilevel VGs, and to characterize the possible differences in the cognitive processes among groups of individuals with different expertise for a conflict detection task. The MTAHC algorithm was applied to a task of interrogating multiple aircraft on a radar screen in order to detect conflicts that would occur in the near future. Multilevel VGs were evaluated based on the expertise (experts and novices) and the difficulty of the task (easy, moderate, and hard). Oculomotor statistics were calculated. Conflict detection rates for the novices were investigated for relationships to their multilevel VGs.

A. Participants

Twenty-five air traffic controllers certified by the Federal Aviation Administration with an average of 20.7 years (*SD* 7.1) of experience were recruited from the Indianapolis ARTCC. Ages were not recorded. Forty novice graduate and undergraduate students were recruited. The average age of the novices was 23.5 years (*SD* 2.9).

B. Apparatus

A Tobii X60 eye tracker collected data at a rate of 60 Hz. The threshold for defining a fixation was set at 100 ms. A 19-inch monitor (1280 × 1024 pixels) was used. The eye tracker was positioned below the monitor and was not in direct contact with the participant. The accuracy of the eye tracker was 0.5° of visual angle. Each participant's eyes were approximately 1 m away from the radar display; therefore, the fixation error could be up to 1 cm on the display.

Simscope and Simtarget [33] were used to simulate the air traffic. The simulation settings were based on a digital surveillance radar mode, and the aircraft movement refresh rate was set to 5 s. An aircraft was represented as a diamond shape (0.2 cm × 0.2 cm) with a direction indicator (1.5 cm in length) and a data block (1.1 cm in height × 1.5 cm in length).

C. Task

The task was to detect all aircraft pairs expected to have conflict in the near future. A conflicting pair is defined as a pair that has lost separation standards by not maintaining a separation of 1000 ft vertically and 5 nmi horizontally.
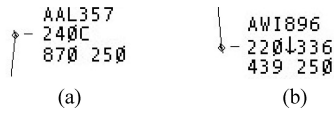
Fig. 9. Aircraft and its data block. (a) Constant altitude. (b) Change of altitude.
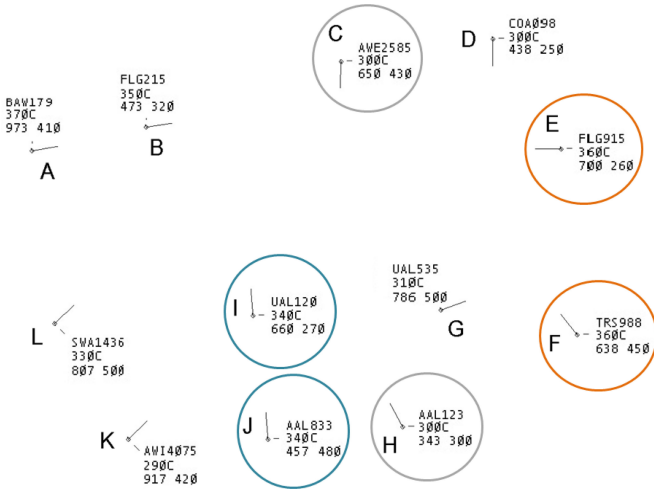
Fig. 10. Easy difficulty scenario.

Fig. 11. Moderate difficulty scenario.

Fig. 12. Hard difficulty scenario.

D. Scenarios

Two training scenarios and three experimental scenarios (easy, moderate, and hard levels of difficulty) were created. Each training scenario showed eight aircrafts, and each experimental scenario showed 12 aircrafts. Each sample scenario had two conflict pairs, and each experimental scenario had three conflict pairs.

Conflicts were designed to occur in at least 8 min after the start of each scenario. The lead time of conflicts were designed based on a pilot study, where air traffic controllers identified all near-future conflicts within 3 min or less. Thus, participants were expected to complete the conflict detections well before they would or would not occur.

An aircraft has a direction indicator and a data block that contains a call sign, altitude, and speed (see Fig. 9). The diamond shape represents the actual aircraft, which is headed south in the example. The length of the direction indicator represents 5 nmi. The data block in Fig. 9(a) indicates that the aircraft's call sign is AAL357, altitude is constant at 24 000 ft (240C), computer ID is 870, and speed is 250 kts. The computer ID is not needed for detecting conflicts.

If the aircraft climbs or descends, the altitude change information is shown on the second line of the data block [see Fig. 9(b)]. The aircraft AWI896 is aiming for an altitude of 22 000 ft (220), currently descending (↓), and currently at an altitude of 33 600 ft (336). The altitude changes of aircraft were indicated immediately at the start of each scenario, and the climb or descend rate was set at a constant rate of 1800 ft./min.

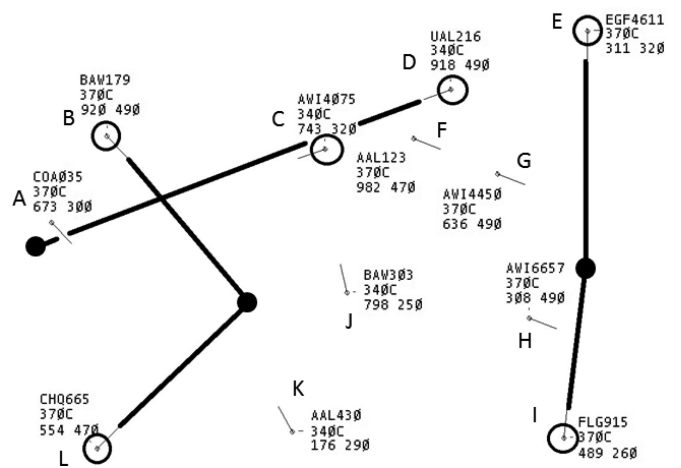Fig. 10 shows the initial layout of the scenario with an easy level of difficulty. All aircrafts move in every 5 s. All aircrafts in the easy scenario maintain constan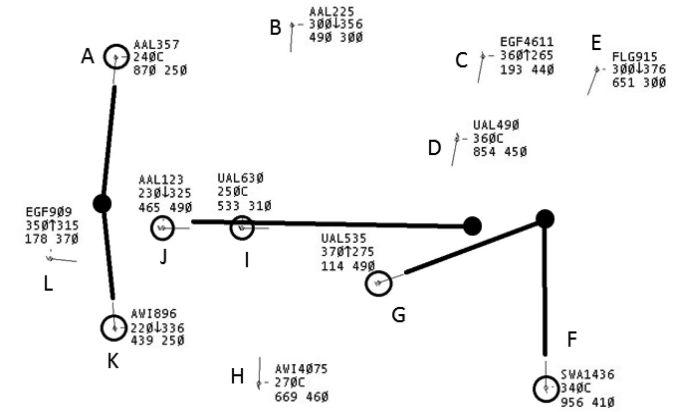t altitudes. The conflict pairs are $\{I, J\}$, $\{C, H\}$, and $\{E, F\}$ are easier to find, since all nonconflicting pairs have different altitudes. In the case of $\{I, J\}$, a conflict occurs because J will overtake I due to speed differences.

Fig. 11 shows the initial layout of the scenario with a moderate level of difficulty. All aircraft in this scenario maintain constant altitudes. The conflicting aircraft $\{B, L\}$, $\{C, D\}$, and $\{E, I\}$ are circled in black. Their trajectories are shown as thick lines, and their collision locations in the near future are highlighted as black dots. Half of the aircraft were at altitudes of 34 000 ft, and the other half were at 37 000 ft. In this scenario, finding conflict pairs is more difficult because many aircraft share similar altitudes but not all are involved in conflicts.

Fig. 12 shows the initial layout of the scenario with the hard level of difficulty. The conflict pairs are $\{A, K\}$, $\{I, J\}$, and $\{F, G\}$. Finding conflict pairs is more difficult as aircraft are changing altitudes.

E. Procedure

The participants (25 experts and 40 novices) were provided with basic training on how to observe and detect conflicts from

the simulated enroute radar screen. The participants were trained to quickly and verbally answer (into a microphone) the aircraft names that would have conflict in the near future as soon as they detected them. After the participants felt comfortable identifying all possible conflict pairs, they were told to say "done," which concluded each scenario. The answers (conflicting aircraft) to the training scenarios were provided to the participants. The participants' eye movements were recorded per scenario. Each participant finished each scenario in less than 3 min. Three scenarios (easy, medium, and hard difficulty level scenario) were presented to the participants in random order.

F. Independent and Dependent Variables

The independent variables were the expertise (experts and novices) and the difficulty of the task (easy, moderate, and hard). The dependent variables were multilevel VG sets of both experts and novices, as well as the correct conflict detection rates of the novices. As part of the agreement with respect to obtaining eye movement data from these controllers, no performance data were recorded.

G. Processing Eye Movement Data

Eye movement data were collected using the Tobii Studio software. For defining a spatial fixation, velocity threshold identification (or I-VT) algorithm [34], [35] with the threshold of 0.42 pixels/ms (default setting of the Tobii Studio software) was applied.

Eye tracking tools provide pixel coordinates of where an eye fixation occurred, but do not provide information on whether the eye fixation occurred on the target. Therefore, researchers design spatially fixed AOIs [10] around the targets. As opposed to using dynamic AOIs created from interpolating sizes and shapes of polygons between two time points [36], the concept of target of interests (TOIs) [29], [37] was used so that the AOIs would move along with the target. A TOI for each target includes the aircraft, its direction indicator, and its data block. The TOIs were used to map the locations of the moving targets. Then, the scanpath sequences were obtained using the TOIs in order to apply the MTAHC algorithm. Custom software was developed to implement the MTAHC algorithm. A total of 195 matrices (65 participants × 3 scenarios) were created and processed to obtain multilevel VGs.

H. Data Analysis

Oculomotor statistics were calculated for eye fixation numbers (FNs) and durations.

The multilevel VGs were visualized by overlaying them onto the initial display of the scenarios. Two-sided Chi-square tests were used to determine the association between the multilevel VGs that included the conflict pairs and expertise.

For the novices, their overall correct conflict detection rates (obtained through participants' verbal answers) for each scenario were compared with the proportions of number of multilevel VG sets for each scenario.
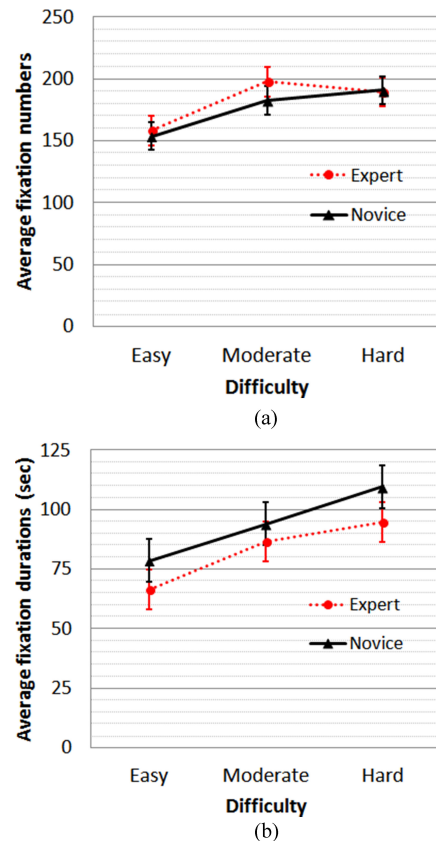

Fig. 13. Average eye FNs and durations. (a) Average FNs. (b) Average FDs.

VI. RESULTS

A significance level of $\alpha = 0.05$ was used for statistical tests.

A. Fixation data

Fig. 13 provides the average eye FNs and durations of the experts and the novices for each scenario. The average eye FNs were higher for the experts, and the average eye fixation durations (FDs) were higher for the novices. A Mann–Whitney test indicated a significant difference for expertise on FN ($U = 3646, p = 0.026$). Kruskal–Wallis tests on scenarios (easy, moderate, and hard) showed significant differences on both FN ($H = 13.9, p = 0.001$) and FD ($H = 26.4, p < 0.001$). Posthoc analyses were performed using the Mann–Whitney tests with $\alpha = 0.017$ (or 0.05/3). Significant FN differences were found for easy versus moderate difficulties ($U = 1483.5, p = 0.003$) and easy versus hard difficulties ($U = 1362.5, p < 0.001$). Significant FD differences were found for easy versus moderate difficulties ($U = 1300, p < 0.001$) and easy versus hard difficulties ($U = 959, p < 0.001$).

B. Multilevel Visual Grouping Sets for the Easy Difficulty Scenario

Chi-square tests for the easy scenario showed significant differences between the experts and the novices in terms of
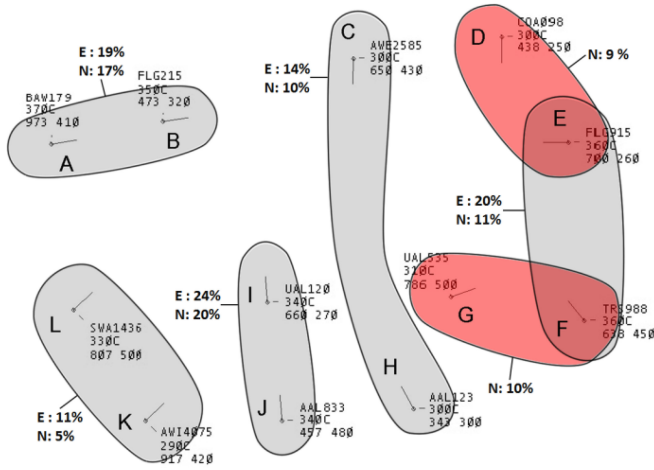
Fig. 14. Level 1 VG sets from easy difficulty scenario. "E" stands for experts and "N" for novices.

(a)

(b)

Fig. 15. Level 2 VG sets from the easy difficulty scenario, (a) Experts. (b) Novices.

the proportion of multilevel VGs that contained conflict pairs. Level 1 VG, $\chi^2 \, 1, 245 = 20.666$, $p < 0.001$, and level 2 VG χ^2 $1, 118 = 8.504$, $p = 0.004$.

Fig. 14 provides the level 1 VG sets that showed proportions of 5% or higher. The proportions of the VG sets are represented for each VG set. Experts and novices created similar VG sets, which are highlighted in gray; however, novices created additional VG sets, which are highlighted in red. Both experts and the novices created high proportions of VG sets from $\{C, H\}$, $\{E, F\}$, and $\{I, J\}$, which accord with the conflict pairs. However, novices created additional high percentages of VG sets from $\{D, E\}$ and $\{F, G\}$, which were not in conflict, but in close proximity and converging.

Fig. 15 shows the level 2 VG sets. The conflict pairs are highlighted in small black circles along with their trajectories until conflict occurs. The experts correctly formed VG sets around the conflict pairs [see Fig. 15(a)], whereas the novices formed many other VG sets around nonconflict pairs [see Fig. 15(b)]. The correct detection rates for novices were 60%, 82.5%, and 67.5% for the conflict pairs $\{C, H\}$, $\{E, F\}$, and $\{I, J\}$, respectively.

### C. Multilevel Visual Groupings Sets for the Moderate Difficulty Scenario

Chi-square tests for the moderate difficulty scenario showed no significant differences for the level 1 VG sets, but showed significant differences for the level 2 VG sets between the experts and the novices ($\chi^2 \, 1, 165 = 3.936$, $p = 0.047$).

For level 1 VG sets, the novices created additional VG sets around the nonconflicting pairs that were converging and in close proximity, but were safely separated by altitude; however, the experts also formed many VG sets around nonconflicting pairs that were converging under same altitudes.

For level 2 VG sets, Fig. 16 shows how experts formed correct VG sets, whereas novices did not. Fig. 16(a) shows that experts correctly created VGs around the conflict pairs $\{B, L\}$, $\{C, D\}$, and $\{E, I\}$; however, Fig. 16(b) shows that novices created
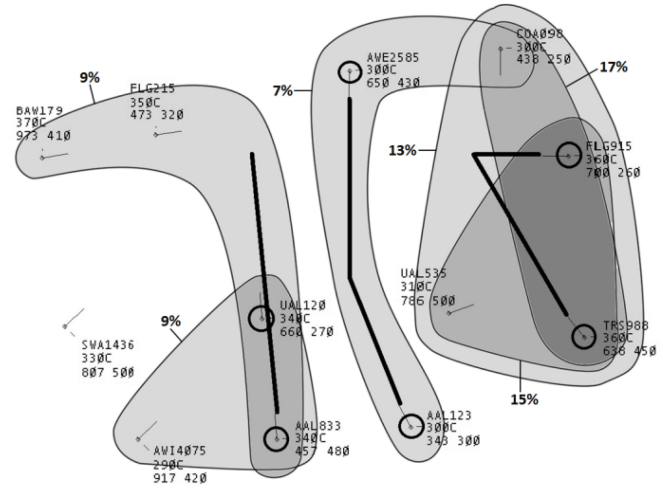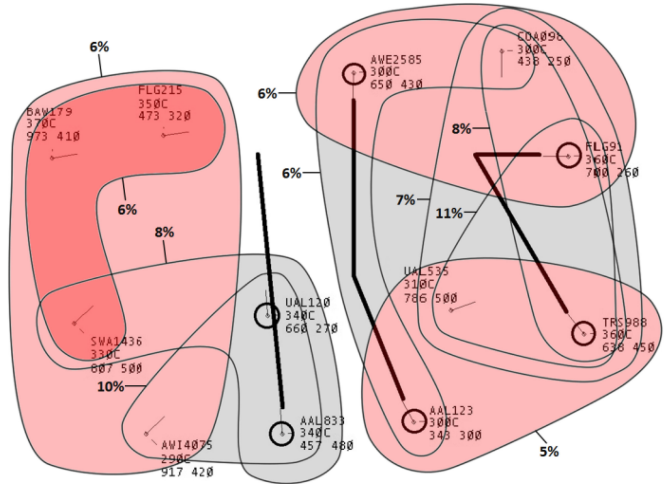
additional VG sets around multiple aircraft to the left side of the display (highlighted in red) and fewer VG sets around the aircraft on the right side of the display. In addition, the experts created VG sets around the same altitudes either at 34 000 ft or 37 000 ft, whereas novices did not.

The correct detection rates for the moderate scenario for novices were 62.5%, 45%, and 30% for the conflict pairs $\{B, L\}$, $\{C, D\}$, and $\{E, I\}$, respectively. The correct detection rate was the lowest for the pair $\{E, I\}$. This result largely accorded with the VG result, instead of forming VG sets that contained $\{E, I\}$, novices created a high percentage of the VG set $\{E, F, G\}$ (21%) on the upper right part of the screen, which did not contain conflict pairs.

D. Multilevel Visual Grouping Sets for the Hard Difficulty Scenario

Chi-square tests for the hard difficulty scenario showed no significant differences for the level 1 VG sets, but a significant
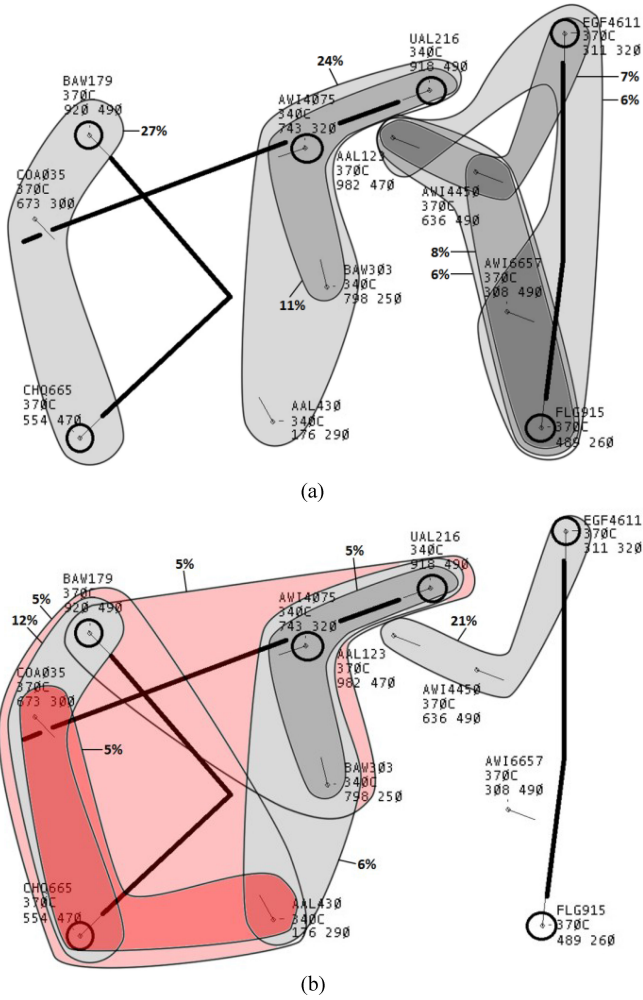
(a)

(b)

Fig. 16.   Level 2 VGs from moderate difficulty scenario. (a) Experts. (b) Novices.

(a)

(b)

Fig. 17.   Level 2 VGs from hard difficulty scenario. (a) Experts. (b) Novices.

difference for the level 2 VG sets, between the experts and the novices ($\chi^2\,1,191 = 24.874$, $p < 0.001$).

For level 1 VG sets, again the novices created VG sets around the nonconflicting pairs that were converging and in close proximity, but were safely separated by altitude; however, the experts also formed many VG sets around nonconflicting converging pairs that were to be at similar altitude in the near future.

For level 2 VG sets, Fig. 17 shows how experts formed correct VG sets, whereas novices did not. Experts correctly created VGs around the conflict pairs $\{A, K\}$, $\{F, G\}$, and $\{I, K\}$ [see Fig. 17(a)], whereas novices created VG sets around the nonconflicting aircraft highlighted in red [see Fig. 17(b)].

The correct detection rates for the hard level of difficulty scenario for novices were 30%, 17.5%, and 47.5% for the conflict pairs $\{A, K\}$, $\{F, G\}$, and $\{I, K\}$, respectively. The correct detection rate was the lowest for the pair $\{F, G\}$. This result largely accorded with the VG result, instead of forming VG sets that contained $\{F, G\}$, novices created high proportions of the VG sets for $\{C, D, G\}$ (14%) and $\{C, D, E, G\}$ (18%) in the upper right part of the screen, which did not contain conflict pairs.
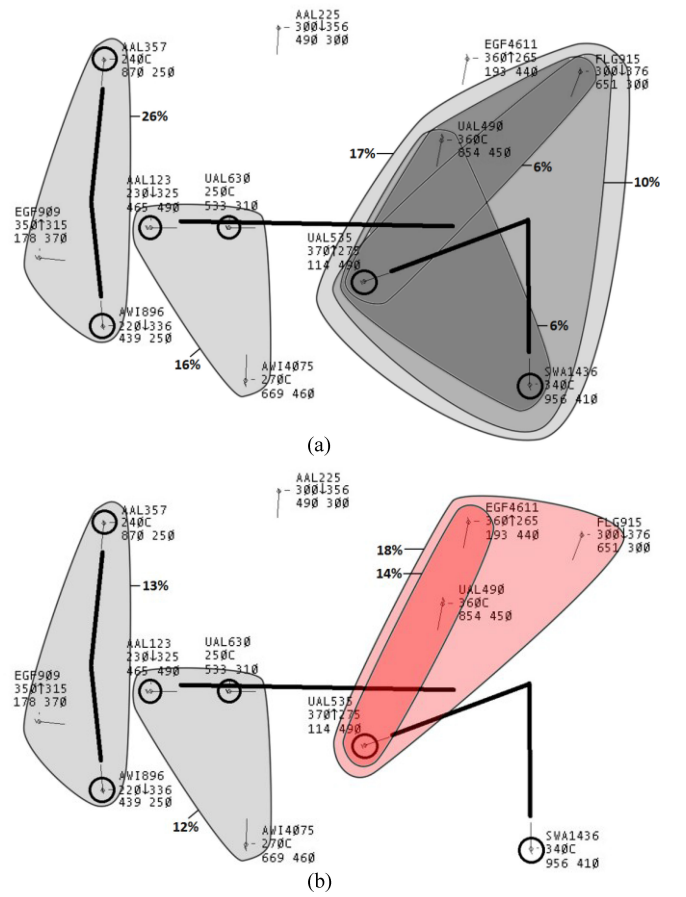
VII. DISCUSSION

We were able to characterize, compare, and analyze complex eye movements for a multielement target tracking task. Specifically, VGs sets were created using the MTAHC algorithm that showed interrogated pairs or groups of targets. The MTAHC algorithm (1) transformed the scanpaths into an unordered transition matrix, and (2) used an adapted hierarchical clustering algorithm to produce multilevel VGs. VG sets enabled 1) effective quantification and comparison of eye movements, 2) performance evaluation, and 3) analysis of the possible different underlying cognitive processes for different VG sets. These contributions should enable researchers and practitioners to effectively compare scanpaths across different individuals, applications, or treatments.

A. Experiment on Conflict Detection Task

Different multilevel VGs formed between the experts and the novices. First, multilevel VG sets indicated different underlying cognitive processes between the experts and the novices. The experts created multilevel VGs around aircraft that were at the same altitude or would have similar altitude in the near future, whereas the novices tended to create multilevel VGs around aircraft that are converging and in close proximity without considering altitudes similarities.

This result extends the finding that participants observed altitudes first when evaluating conflicts for one pair of converging aircraft [31], [38]. The experiment in this paper further shows how multiple moving aircraft are interrogated differently based on expertise.

One reason the novices did not effectively identify conflicts could be that they simply did not know that grouping aircraft by altitude instead of convergence could be more effective when many aircraft are presented on screen. A radar screen is a 2-D representation of a 3-D airspace in which the vertical location (altitude) of an aircraft is presented as data values, and the vertical separations among aircraft cannot be visually observed, unlike the horizontal separations. Therefore, it seems that the experts were well trained to immediately consider altitudes first whereas novices did not.

Second, multilevel VG sets provided a way to utilize eye movements for performance analysis by finding out whether VG sets were created around the conflict pairs. A high proportion of VG sets around conflict pairs can be a good indicator for identifying whether the participant was correctly paying attention to the conflict pairs. For example, as the difficulty of the scenario increased, the obtained multilevel VG sets showed that the novices tended not to interrogate aircraft on the left lower side of the screen.

Level 1 VGs showed that novices created additional VGs around unnecessary aircraft pairs that were safely separated by altitudes; however, the comparison tests for the moderate and hard difficulty scenarios were not significant. The reason is that many converging nonconflict pairs existed, which were or would have similar altitudes in the near future. Therefore, both groups had to evaluate those pairs to determine whether conflict would occur or not, which was not the case for the easy difficulty scenario.

However, significant differences from level 2 VG sets, as well as the accompanied illustrations show how VG sets are explicitly formed around the conflict pairs for the experts and not for the novices. Specifically, in all three scenarios, the illustrations of level 2 VG sets show that experts largely formed three clusters around three conflicts, indicating their awareness of the given task, whereas the novices did not show this tendency.

B. Maximum Transition-Based Agglomerative Hierarchical Clustering Algorithm

The MTAHC algorithm was able to characterize long scanpaths that have large deviations. The algorithm created multilevel VG sets based on transitions among the targets, and the VG sets were mapped to the cognitive processes. We have shown that hierarchically and iteratively grouping multiple targets that have higher transitions compared with other targets can provide insights to better evaluate performance and understand the underlying cognitive processes.

The multilevel VGs obtained from the MTAHC algorithm can be useful, where transitions among many targets require investigation. For example, if a small number of targets are investigated, then either applying the binomial test from the transition

matrix [4] or comparing whole matrices [29] may be sufficient. However, if the number of targets increases, the size of the matrix grows accordingly, making it difficult to apply the methods due to either too many tests to perform or to a decrease of sensitivity. The MTAHC algorithm is not affected by these issues.

Next, multilevel VG sets from eye movements can serve as a complement to analyzing verbal protocols for a conflict detection task. For example, if an individual was interrogating nonconflicting targets, multilevel VG sets can identify those targets. In detail, an individual can interrogate nonconflicting targets for a long duration only to find out that the targets would not conflict and choose to not answer. It is possible to obtain this information through a critical thinking process using a talk aloud protocol [39]. However, the talk aloud protocol could affect the task performance since an individual must articulate her/his actions during the task [40], whereas multilevel VGs are obtained with no interference.

Finally, the MTAHC algorithm can be used for spatially fixed targets, moving targets, or a combination of these targets. The algorithm uses sequences of enumerated types from defined targets, and the sequences are used to create unordered transition matrices to compute multilevel VGs. Therefore, the algorithm is not affected by dynamic changes to the display or scenery.

C. Limitations and Future Research

Although we showed that MTAHC can be used to evaluate performance, we have not investigated whether the algorithm can be used to quantify workload. Previous research works have investigated eye movements to quantify workload through analyzing sensitivity of spatial dispersion indices obtained from visual search patterns [41], [42]. It may be possible that the hierarchical structure of the MTAHC algorithm may contribute to the above research. If hierarchical structure does not exist in the matrix, then it could be an indicator of low workload meaning that interrogation of multiple targets are being performed in a random fashion since there are no conflicts.

The analysis is based on goal-oriented or top-down perspective. Although a goal-based task revealed that multilevel VGs could be used to understand different cognitive processes, it is still unclear whether and how much Gestalt effects [2], [43] may have taken effect between the groups of different expertise.

Finally, MTAHC may be used in conjunction with other approaches, such as the multimatch method [23], [24] that combines several dimensions of eye movement characteristics. For example, FDs (which is also the basis for the heatmaps), as well as vector information, may be accompanied by multilevel VGs to better understand and analyze the overall eye movements.

REFERENCES

[1] R. Pieters, E. Rosbergen, and M. Wedel, "Visual attention to repeated print advertising: A test of scanpath theory," *J. Marketing Res.*, vol. 36, no. 4, pp. 424–438, 1999.

[2] S. J. Landry, T. B. Sheridan, and Y. M. Yufik, "A methodology for studying cognitive groupings in a target tracking task," *IEEE Trans. Intell. Transp. Syst.*, vol. 2, no. 2, pp. 92–100, Jun. 2001.

[3] P. Chapman, G. Underwood, and K. Roberts, "Visual search patterns in trained and untrained novice drivers," *Transp. Res. Part F, Psychol. Behav.*, vol. 5, no. 2, pp. 157–167, 2002.

[4] G. Underwood, P. Chapman, N. Brocklehurst, J. Underwood, and D. Crundall, "Visual attention while driving: Sequences of eye fixations made by experienced and novice drivers," *Ergonomics*, vol. 46, no. 6, pp. 629–646, 2003.

[5] A. Çöltekin, S. I., Fabrikant, and M. Lacayo, "Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings," *Int. J. Geog. Inf. Sci.*, vol. 24, no. 10, pp. 1559–1575, 2010.

[6] J. Tchalenko, "Segmentation and accuracy in copying and drawing: Experts and beginners," *Vis. Res.*, vol. 49, no. 8, pp. 791–800, 2009.

[7] W. J. Gibbs and R. S. Bernas, "Visual attention in newspaper versus TV-oriented news websites," *J. Usability Stud.*, vol. 4, no. 4, pp. 147–165, 2009.

[8] C. Holland and O. Komogortsev, "Biometric identification via eye movement scanpaths in reading," in *Proc. Int. Joint Conf. Biometrics*, 2011, pp. 1–8.

[9] C. Krischer and W. Zangemeister, "Scanpaths in reading and picture viewing: Computer-assisted optimization of display conditions," *Comput. Biol. Med.*, vol. 37, no. 7, pp. 947–956, 2007.

[10] A. Poole, L. J. Ball, and P. Phillips, "In search of salience: A response-time and eye-movement analysis of bookmark recognition," in *People and Computers XVIII-Design for Life*, S. Fincher, P. Markopoulos, D. Moore, and R. Ruddle, Eds. London, U.K.: Springer, 2005, pp. 363–378.

[11] R. Caldara and S. Miellet, "iMap: A novel method for statistical fixation mapping of eye movement data," *Behav. Res. Methods*, vol. 43, no. 3, pp. 864–878, 2011.

[12] R. W. Reeder, P. L. Pirolli, and S. K. Card, "WebEyeMapper and WebLogger: Tools for analyzing eye tracking data collected in web-use studies," in *Proc. ACM Conf. Human Factors Comput. Syst.*, New York, NY, USA, 2001, pp. 19–20.

[13] F. Cristino, S. Mathôt, J. Theeuwes, and I. D. Gilchrist, "ScanMatch: A novel method for comparing fixation sequences," *Behav. Res. Methods*, vol. 42, no. 3, pp. 692–700, 2010.

[14] J. Goldberg and X. Kotval, "Computer interface evaluation using eye movements: Methods and constructs," *Int. J. Ind. Ergonom.*, vol. 24, no. 6, pp. 631–645, 1999.

[15] D. Noton and L. Stark, "Eye movements and visual perception," *Sci. Amer.*, vol. 224, no. 6, pp. 34–43, 1971.

[16] S. Brandt and L. Stark, "Spontaneous eye movements during visual imagery reflect the content of the visual scene," *J. Cognit. Neurosci.*, vol. 9, no. 1, pp. 27–38, 1997.

[17] A. T. Duchowski, J. Driver, S. Jolaoso, W. Tan, B. N. Ramey, and A. Robbins, "Scanpath comparison revisited," in *Proc. Symp. Eye-Tracking Res. Appl.*, New York, NY, USA, 2010, pp. 219–226.

[18] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, Sep. 2000.

[19] P. Hejmady and N. H. Narayanan, "Visual attention patterns during program debugging with an IDE," in *Proc. Symp. Eye Tracking Res. Appl.*, 2012, pp. 197–200.

[20] J. H. Goldberg and J. I. Helfman, "Scanpath clustering and aggregation," in *Proc. Symp. Eye-Tracking Res Appl.*, 2010, pp. 227–234.

[21] S. Mannan, K. Ruddock, and D. Woodling, "The relationship between the locations of spatial features and those of fixations made," *Spatial Vis.*, vol. 10, pp. 65–188, 1996.

[22] S. Mathôt, F. Cristino, I. D. Gilchrist, and J. Theeuwes, "A simple way to estimate similarity between pairs of eye movement sequences," *J. Eye Mov. Res.*, vol. 5, no. 1, pp. 1–15, 2012.

[23] R. Dewhurst, J. Jarodzka, K. Holmqvist, T. Foulsham, and M. Nyström, "A new method for comparing scanpaths based on vectors and dimensions," *J. Vision*, vol. 11, no. 11, pp. 502–502, 2011.

[24] R. Dewhurst, M. Nyström, H. Jarodzka, T. Foulsham, R. Johansson, and K. Holmqvist, "It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach," *Behav. Res. Methods*, vol. 44, no. 4, pp. 1079–1100, 2012.

[25] J. West, A. Haake, E. Rozanski, and K. Karn, "eyePatterns: Software for identifying patterns and similarities across fixation sequences," in *Proc. Symp. Eye Tracking Res. Appl.*, San Diego, CA, USA, 2006, pp. 149–154.

[26] M. Just and P. Carpenter, "Eye fixations and cognitive processes," *Cognit. Psychol.*, vol. 8, no. 4, pp. 441–480, 1976.

[27] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation," in *Proc. 8h ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.*, 2002, pp. 429–435.

[28] G. Underwood, T. Foulsham, and K. Humphrey, "Saliency and scan patterns in the inspection of real-world scenes: Eye movements during encoding and recognition," *Vis. Cognit.*, vol. 17, nos. 6/7, pp. 812–834, 2009.

[29] Z. Kang and S. J. Landry, "Capturing and analyzing visual groupings of multiple moving targets in an aircraft conflict detection task using eye movements," in *Proc. 54th Annu. Meet. Human Factors Ergonom. Soc.*, San Francisco, CA, USA, 2010, pp. 1906–1910.

[30] S. Loft, P. Sanderson, A. Neal, and M. Mooij, "Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 49, no. 3, pp. 376–399. 2007.

[31] Z. Kang, E. J. Bass, and D. W. Lee, "Air traffic controllers' visual scanning, aircraft selection, and comparison strategies in support of conflict detection," presented at the 58th Annu. Meet. Human Factors Ergonom. Soc., Chicago, IL, USA, 2014.

[32] S. Kotsiantis and P. Pintelas, "Recent advances in clustering: A brief survey," *WSEAS Trans. Inf. Sci. Appl.*, vol. 1, no. 1, pp. 73–81, 2004.

[33] R. Carlson. (2014, Sep. 15). SimTarget and SimScope. Ver. 1.0.117, [Computer software]. [Online]. Available: http://www. metacraft.com/, 2008.

[34] P. Olsson, "Real-time and offline filters for eye tracking," M.S. thesis, Dept. Elect. Eng., KTH Royal Inst. Technol., Stockholm, Sweden, Apr. 2007.

[35] O. V. Komogortsev, D. V. Gobert, S. Jayarathna, D. H. Koh, and S. M. Gowda, "Standardization of automated analyses of oculomotor fixation and saccadic behaviors," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 11, pp. 2635–2645, Nov. 2010.

[36] F. Papenmeier and M. Huff, "DynAOI: A tool for matching eye-movement data with dynamic areas of interest in animations and movies," *Behav. Res. Methods*, vol. 42, no. 1, pp. 179–187, 2010.

[37] Z. Kang and E. J. Bass, "Supporting the eye tracking analysis of multiple moving targets: Design concept and algorithm," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, San Diego, CA, USA, 2014, pp. 3191–3196.

[38] E. Rantanen and A. Nunes, "Hierarchical conflict detection in air traffic control," *Int. J. Aviat. Psycol.*, vol. 15, no. 4, pp. 339–362, 2005.

[39] M. W. Van Someren, Y. F. Barnard, and J. A. Sandberg, *The Think Aloud Method: A Practical Guide to Modeling Cognitive Processes*. London, U.K.: Academic Press, 1994.

[40] M. Van Den Haak, M. De Jong, and P. Jan Schellens, "Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue," *Behav. Inf. Technol.*, vol. 22, no. 5, pp. 339–351, 2003.

[41] F. Di Nocera, M. Camilli, and M. Terenzi, "A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload," *J. Cognit. Eng. Decision Making*, vol. 1, no. 3, pp. 271–285, 2007.

[42] B. Hilburn, P. G. Jorna, E. A. Byrne, and R. Parasuraman, "The effect of adaptive air traffic control (ATC) decision aiding on controller mental workload," in *Human-Automation Interaction: Research and Practice*, M. Mouloua and J. Koonce, Eds. Mahwah, NJ, USA: Erlbaum, 1997, pp. 84–91.

[43] S. Yantis, "Multielement visual tracking: attention and perceptual organization," *Cognit. Psychol.*, vol. 24, pp. 295–340, 1992.

Ziho Kang received the Ph.D. degree from the School of Industrial Engineering, Purdue University, West Lafayette, IN, USA, in 2010.

He is currently an Assistant Professor with the School of Industrial and Systems Engineering, University of Oklahoma, Norman, OK, USA. His research interests include human-integrated systems engineering, and his specialty is in eye tracking research.

Steven J. Landry received the Ph.D. degree from the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2004.

He is currently an Associate Professor and Associate Head with the School of Industrial Engineering, Purdue University, West Lafayette, IN, USA. He is specialized in the area of human factors and air transportation systems engineering..