Identifying Gaze Behavior Evolution via Temporal Fully-Weighted Scanpath Graphs

EDUARDO DAVALOS, Vanderbilt University, USA
CALEB VATRAL, Vanderbilt University, USA
CLAYTON COHN, Vanderbilt University, USA
JOYCE HORN FONTELES, Vanderbilt University, USA
GAUTAM BISWAS, Vanderbilt University, USA
NAVEEDUDDIN MOHAMMED, Vanderbilt University, USA
MADISON J. LEE, Vanderbilt University, USA
DANIEL T. LEVIN, Vanderbilt University, USA

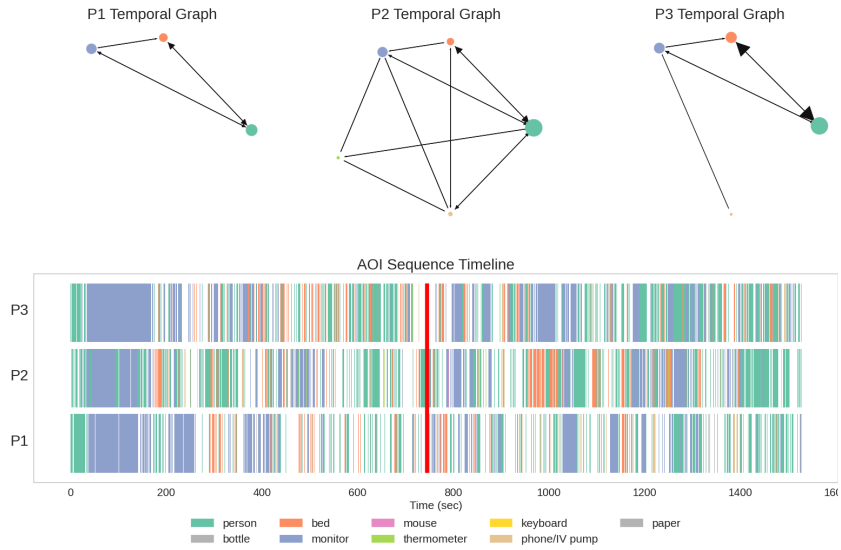

Fig. 1. A snapshot of the temporal fully-weighted graphs of a simulation nurse triad taken at $t \simeq 760$ seconds, indicated by the red vertical line. At the top of the figure, the participating nurses (P1 to P3) have their own graph built based on their gaze data that has been mapped to the mixed-reality nursing scenario by matching gaze to relevant objects. These graphs model gaze evolution and decay by altering the nodes' and edges' weight over time. Node weight represents accumulated fixation duration and edge weight indicates the accumulated number of transitions between two objects. The nurses' gaze over the chosen time interval, presented at the bottom of the figure, illustrates the temporal change in gaze fixation for the three nurses across the objects.

Manuscript submitted to ACM

Eye-tracking technology has expanded our ability to quantitatively measure human perception. This rich data source has been widely used to characterize human behavior and cognition. However, eye-tracking analysis has been limited in its applicability, as contextualizing gaze to environmental artifacts is non-trivial. Moreover, the temporal evolution of gaze behavior through open-ended environments where learners are alternating between tasks often remains unclear. In this paper, we propose temporal fully-weighted scanpath graphs as a novel representation of gaze behavior and combine it with a clustering scheme to obtain high-level gaze summaries that can be mapped to cognitive tasks via network metrics and cluster mean graphs. In a case study with nurse simulation-based team training, our approach was able to explain changes in gaze behavior with respect to key events during the simulation. By identifying cognitive tasks via gaze behavior, learners' strategies can be evaluated to create online performance metrics and personalized feedback.

CCS Concepts: • **Applied computing** → **Collaborative learning**; • **Computing methodologies** → **Cluster analysis**.

Additional Key Words and Phrases: eye-tracking, network analysis, temporal, learning analytics, simulation-based training

ACM Reference Format:
Eduardo Davalos, Caleb Vatral, Clayton Cohn, Joyce Horn Fonteles, Gautam Biswas, Naveeduddin Mohammed, Madison J. Lee, and Daniel T. Levin. 2023. Identifying Gaze Behavior Evolution via Temporal Fully-Weighted Scanpath Graphs. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3576050.3576117

1 INTRODUCTION

With the adoption of innovative technologies and new care practices in the healthcare sector, healthcare educators and teaching institutions are now turning their attention to crafting efficient and effective curricula that improve the training experiences of learners. An established yet evolving instructional approach that bridges the gap between classroom knowledge and real-world readiness is simulation-based training that can replicate real-world conditions with high fidelity [30]. In particular, nursing instruction has adopted simulation-based training as a pedagogical approach that combines testing of nurses' clinical judgment with procedures they will execute in the real world [16, 18].

Simulation-based training is popular and has been widely adopted by nursing educational institutions, but robust analytics-based performance feedback that helps students achieve proficiency is still not well-developed [18]. This is because most analyses of nurse progress are qualitative and somewhat ad hoc in nature, and systematic documentation of students' progress as they go through various scenarios are rare [15]. These issues are further exacerbated by the open-ended nature of simulation-based training. This creates a situation where there is more than one way to approach a patient care problem, making it difficult to generate analytics that is valid across the range of possible approaches that nurses may take. Further, multiple nurses may collaborate in a simulation. Although this is closer to reality, this collaboration makes it harder for nursing instructors to assess individual and team performance, especially because nurses can switch between a variety of roles and tasks as a scenario unfolds. These complexities make it difficult for instructors to develop a nuanced analysis of the activities that nurses perform during the simulation, and this motivates the need to develop automated artificial intelligence (AI)-based algorithms and analytics that can support nursing instructors and students.

In the past, individual and team performance have been measured using techniques such as interviews, surveys, and pre-post tests. However, these methods fail to capture in-the-moment decisions that nurses make and the details of the procedures they perform. Moreover, these approaches are restricted to after-the-fact post-simulation evaluation only. Measures of real-time team metrics that rely on more quantitative data sources, such as nurses' activities, movements, emotions, behaviors, and visual attention, are currently active areas of research. A common barrier to collecting activity and behavior data is the lack of domain contextualization whereby the observed behavior is situated to domain-specific

information [38]. This problem is often caused by a focus on a single data source, e.g., tracking the nurses' activity data in the simulation log files. Such data are often insufficient to capture the full extent of the nurses' behavior and decision-making, because the environmental context in which actions are being performed may not be available [31]. Current advances in multimodal data collection and analysis [7, 8, 37] provide approaches to address these limitations. Multimodal Learning Analytics (MMLA) has gathered significant research interest in recent years for its ability to explain learners' performance and behaviors in a more holistic manner. In particular, eye-tracking has been at the forefront of cognition, neuroscience, and MMLA [19] in providing mechanisms to better understand students' performance [14], communication and coordination, and overall team collaboration [1, 34].

Recently, one particularly popular approach to assessing eye tracking data generated by complex real-world tasks has been Network analysis [1, 25, 34, 43], using graph representations, attributes, and metrics to link eye-tracking data with learning theories [10]. However, in previous research eye-tracking graphs are typically generated at the end of the experiment as an accumulated representation of the gaze behaviors, thereby losing important information in the temporal dimension of gaze patterns and gaze shifts that occur in the training scenario. This makes it difficult to interpret and evaluate learners' activities and behaviors in the training scenario.

To address this gap, we focus on the temporal evolution of nurses' eye-tracking behaviors using a network analysis approach and then use this information to analyze nursing students' activities and their coordination behaviors in mixed-reality simulation-based training scenarios. In this paper, we address the following research question:

RQ1: **How can eye-tracking data be used to provide contextualized online information about nurses' activities and behaviors in a complex mixed-reality simulation-based training environment?**

The rest of this paper is organized as follows: Section 2 provides a brief literature review that combines learning theory, data-driven methods, and network analysis to create meaningful depictions and analyses of eye-tracking data. Section 3 presents the approaches we develop to address these goals. Section 4 starts with a case study that evaluates how our method allows gaze behavior to support interpretations of key events in a nursing simulation using a task model we have generated in conjunction with our instructors by applying cognitive task analysis. Section 5 provides the conclusions of this paper along with limitations and future work. Section 6 provides the paper's meta information.

2 BACKGROUND

In this section, we begin with a brief literature review of our cognitive task analysis framework appended with a distributed cognition approach for analyzing teamwork behaviors. In parallel, we review current MMLA applied to learning and training environments and then focus on eye-tracking and network analytics algorithms to derive scan paths that form our core representation for analyzing nurses' teamwork behaviors.

2.1 Cognitive Learning Theory

Classical cognitive theories model human learning by linking external factors with an individual's internal mental processes [41]. Many of these approaches explain how new information is obtained, processed, and stored in a symbolic database. In the lens of traditional cognitive theory, the unit of analysis is an individual's cognitive and metacognitive processes [2]. Clark [12] concisely described how classical cognitive theory models the mind as a central logic engine that fetches information from memory through a symbolic database. The logical inference is applied to problem-solving, where the environment is the problem domain and the body is the sensor for collecting information from the environment [21].

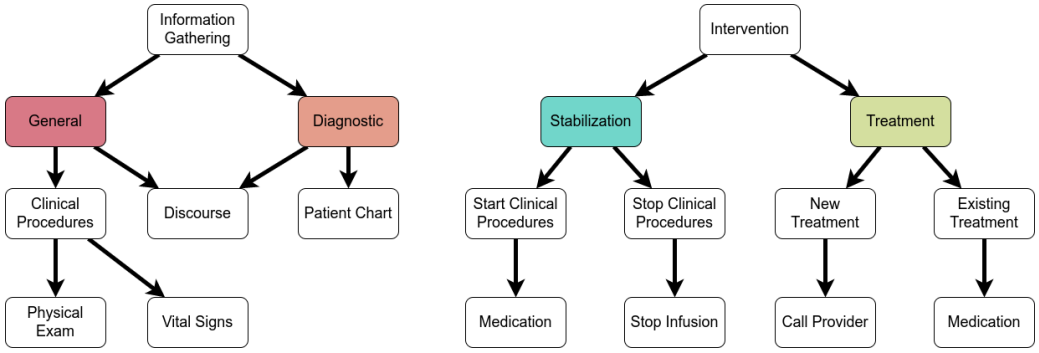

Fig. 2. Cognitive Task Model for simulation-based training scenarios in nursing. The top level of the task model includes domain-general tasks: information gathering and intervention. The further we descend in the model, the tasks become increasingly domain-specific (e.g. checking vital signs).

2.1.1 Cognitive Task Modeling. oo gain a better understanding of the learners' activities and behaviors, we adopt a cognitive task analysis approach [26] to create a hierarchical task structure that captures the set of primary tasks that define our training or learning domain. A task model for the nursing domain is shown in Fig. 2. These concepts map the task domain onto domain-relevant cognitive processes, psychomotor skills, and affective states. For example, learning and training domains typically include high-level cognitive processes, such as information acquisition, solution construction, and evaluation. Although these processes may reflect some invariant properties across domains, their execution likely differs depending on the training scenario and the domain under consideration.

Tasks within a domain are modeled by a top-down hierarchical structure, where the highest levels include domain-general tasks of information gathering, solution construction, and solution assessment [4]. The subsequent subtasks are more domain-specific (e.g., in nursing: checking vitals; providing an intervention). The lowest levels of the hierarchy represent observable actions (e.g., administering oxygen). Betty's Brain [5], a computer-based learning environment that teaches students to model scientific processes (e.g., effects of global warming) as causal relationships between entities (e.g. increased temperature decreases the amount of sea ice), uses a cognitive task model to interpret learner actions in the environment in terms of cognitive processes represented as tasks and subtasks (information seeking, model building, and model checking). It leverages event logs to directly map actions to subtasks in the task model and track the learners' progress and strategies.

2.1.2 Distributed Cognition. Classical cognitive theory often deemphasizes the social component of learning even though real-world learning environments can include other humans who may join an individual in collective interactions with instruments, tools, and other objects. Therefore, distributed cognition theories account for peer-to-peer interactions, and the use of artifacts, such as whiteboards, notebooks, and computers, that are present in the environment. The propagation of information and knowledge within a network largely contributes to learning [40]. Various cognitive scientists, including Hutchins, Cole, and Clark [11, 13, 20], recognized this gap in cognitive theory. To address this gap, Hutchins [21] developed distributed cognition with the view that learning is a joint activity by expanding the unit of analysis to encompass more than a single mind, including networks of individuals, environmental artifacts, and technologies.

With the success of distributed cognition in analyzing behavior within social networks, its adoption has been accompanied by new methodologies that mold the framework to better fit new domains and applications. For our study, we adopt the qualitative *Distributed Cognition for Teamwork (DiCoT)* model [6]. DiCoT has been commonly used within the nursing simulation-based training environment literature [29, 32, 36] to derive qualitative measures and models for training behaviors. The DiCoT model is broken down into 5 themes: (1) physical layout, (2) information flow, (3) design and use of artifacts, (4) social interactions, and (5) temporal evolution. Using these models, we can ground our methods and expand our interpretation and understanding of our data and results.

2.2 Multimodal Learning Analytics

MMLA is a growing field that includes contributions from many researchers, such as Worsley [39], Blikstein [7, 8], Schneider [34], and Ochoa [14], to name a few. The field of MMLA is situated at the intersection of learning analytics, multimodal data, and computer-supported analysis [37]. MMLA has been applied to analyze learning behaviors in complex environments across a variety of applications that have used different sensory data arrangements depending on the needs and the intent of the analyses. The argument for multimodal data analysis, in spite of the expenses and complexity of collecting and analyzing the data, is its ability to provide rich and in-depth measures to characterize complex human behavior.

Using MMLA within a DiCoT framework to analyze mixed-reality simulation-based team training allows for a rigorous mixed-methods approach, as demonstrated in [36]. MMLA has also been successfully applied in collaborative settings. The observable and measurable interactions between students or trainees and their environment are a key focus in MMLA, as a means to relate data to theory constructs. As a preliminary step to understanding complex and noisy environments, dashboards and other types of visualizations have been used to better understand the relationship between the multimodal data and activities in the learning domain [27].

2.2.1 Eye-tracking. As a robust measure of perceptual and cognitive processes, eye-tracking has been an influential tool in the field of learning analytics. It is based on the eye-mind hypothesis [33], in which visual attention is related to the mind's new information processing. To start making the connections between gaze and visual attention, eye-tracking analytics have to conform to and be embedded into the learning context. In our study, the raw gaze information comes in the form of (x, y) coordinates with respect to the individual's egocentric view of the training space, i.e., the simulated hospital room. These data can be analyzed to compute generic metrics, such as average fixation time and saccade frequency. However, more informative analyses must be based on a direct mapping of gaze coordinates to the objects in the person's view, and these objects can be further contextualized in terms of their roles and functionality in the domain and setting. Therefore, it is essential to create a mapping between the raw eye-tracking data and the objects observed in the domain to draw contextualized conclusions of the subject's gaze and link it to their information acquisition, diagnostic inferences, and intervention tasks as implied by the cognitive task model. Area-of-Interest (AOI) encoding is commonly used to facilitate the identification of objects or areas in the environment that provide rich context and help interpret domain concepts, as well as the nurse's interactions with the patient and with each other [17].

There are two distinct types of AOIs: static and dynamic. The availability of AOI information and type is largely dependent on the environment. Environments can be found on a spectrum ranging from physical to virtual, with mixed reality that includes both physical and virtual components in the training scenario, located in between. Static AOIs are commonly used in laboratory settings where the AOI placement, usually in still images, can be fixed, and where the participant's viewpoint does not change relative to a stimulus-display screen. By using static AOIs, contextualizing

eye-tracking can be easily achieved. On the other hand, real-world scenarios, in which gaze is often tracked using head-mounted eye trackers, require dynamic AOI encoding which is a non-trivial task. Using eye-tracking in 3-dimensional environments that the participant moves through poses a large set of computational challenges, especially if researchers wish to avoid time-intensive hand-coding of AOIs.

2.2.2 Network Analytics. Once AOI encoding is achieved, trainees' gaze sequences on AOIs can be articulated and further processed to extract rich information. Various analysis methods can be applied to these sequences, such as sequence mining [23], machine learning (ML) regression or classification [14], time series analysis, and network analysis [25, 43]. In collaborative settings where multiple eye-tracking devices are used, network analysis has been heavily used to represent joint attention [1, 34]. The node and edges of a graph can be used as powerful methods for representing complex relationships between AOIs [25].

3 METHODOLOGY

The approach employed in this paper analyzes nursing training exercises in a mixed-reality environment that represents a simulated hospital room equipped with standard medical devices and monitors for information display and communication of the providers' orders. The patient is represented by a high-fidelity manikin that exhibits distress symptoms and a deteriorating health state.

The instructors monitor the simulation from behind a one-way glass partition, allowing them to observe the student nurses' activities, conversations, and interventions. Then, based on the nurses' specific actions (or lack of actions), the instructor may make real-time modifications to the simulation on the LLEAP software. The instructor can also play the role of the patient by speaking through a microphone in the control room, which can be heard through speakers in the manikin. Therefore, the patient's state, utterances, and health displays are controlled by the patient simulator, whereas the nurses' conversations and activities take place in physical space in and around the simulated hospital room. Thus, the analysis of the nurses' activities and behaviors takes place in the context of the evolving scenario. This constraint requires us to develop mixed-reality methods that combine video, speech, gaze, and log file analysis.

3.0.1 Participants. For our study, we collected data from five separate simulation training sessions run at Vanderbilt's Nursing Simulation lab in Spring 2022. Each simulation involved training of a nurse triad. One out of five of the recorded simulations had eye-tracking data for all three nurses. In the case study discussed in this paper, we analyze the training scenario where all three nurses wore eye-tracking glasses. With the nurses' informed consent, we recorded and analyzed the simulation scenarios using data from the eye-tracker glasses and the overhead camera recordings. The study was approved by the Vanderbilt University Institutional Review Board.

3.0.2 Data Sources. Our eye-tracking data was collected using Tobii Glasses 3. These units record four data streams: egocentric video, audio, eye-tracking, and inertial measurement units. Our second data source, two overhead cameras provided different top-down views of the simulation environment and included both video and audio data.

Participants donned eye-tracking glasses and then completed Tobii's one-point calibration routine. Then, the student nurses entered the simulated hospital room to start their training. As students performed their training exercise, we collected (1) video data from two overhead cameras to analyze the nurses' physical movement and activities of the nurses in the room; (2) audio data from both the tracking glasses and overhead camera videos to capture the nurses' dialogue with each other, the patient, and the provider; and (3) nurses' gaze coordinates using the eye tracking glasses

Fig. 3. Example of aligned and synchronized video feeds from our data collection and preprocessing steps. To collectively analyze the team's interactions & collaboration across a simulation, we aligned the 3 nurses' head-mounted eye-tracking glasses recordings.

as they worked through the simulated scenario. All of the raw data streams were stored during recording and later analyzed offline as we describe below.

3.1 Data Analysis

For our data analysis, a preprocessing step involving temporally aligning the multiple data sources was required. Once the data sources were aligned and synchronized, the egocentric videos were processed to track the dynamic gaze patterns in terms of the AOIs. The fixations were matched with the detected and tracked AOIs in the video generating a sequence of AOIs with timestamps over the period of the training simulation. The AOI sequence of each nurse was then used to construct a temporal fully-weighted graph, acting as a model of attention and memory [9]. Each of the steps in this process is discussed in greater detail below.

3.1.1 Preprocessing. For our multimodal data alignment and synchronization, the videos' start times were used to initiate the timeline. For the alignment of the different data streams, we manually adjusted the individual time stamps to account for time differences between the different computers that collected each data stream. The data streams stemming from the eye tracking glasses were automatically aligned with the egocentric video by Tobii's recording software.

3.1.2 Dynamic AOI Tracking and Matching. To link gaze with AOI's defined by relevant objects in the nursing domain, we used YOLOv5 [22] to perform object detection. We used the deep learning algorithm (YOLOv5) to find the dynamic AOIs through video sequences in spite of the complex layout of the physical environment. We used the pretrained weights and biases provided by the YOLOv5 GitHub repository for the "yolov5s" model. This YOLOv5 model was trained

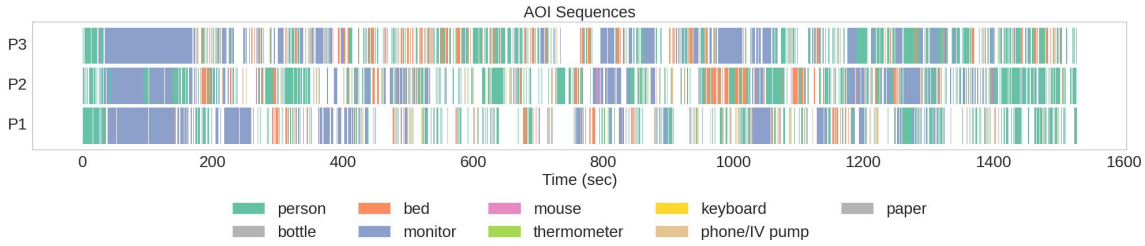

Fig. 4. Timeline of AOI Sequences of Nurse Triad in Simulation A. The fixation-AOI matching is performed on all gaze data over the period of a simulation to obtain the figure above. With the development of this timeline, we were able to detect gaze patterns over the period of the simulation. This provides evidence that the nurses' eye-gaze behavior is correlated with their actions, as it changes as the actions performed change over time.

on the widely-used COCO dataset [24] that contains 91 different annotated categories of everyday objects and items in images. Not all categories detected by YOLOv5 are useful for the nursing domain, but a subset of these categories mapped well with the objects of interest in the simulation environment: person, bottle, bed, tv, laptop, mouse, remote, keyboard, cell phone, and book. These labels were then renamed to better match the objects located in the room. The renaming included the following mappings: "tv" and "laptop" to "monitor", "remote" to "thermometer", "cell phone" to "phone/IV pump", and "book" to "paper". In total, we have 9 AOIs tracked via the YOLOv5 model.

The video was then processed with the YOLOv5 model to detect, track, and match dynamic AOIs through the period of the training simulation for each nurse's egocentric view. An example of a sequence of processed frames with the YOLOv5 detector is shown in Fig. 3. The object detection results of each frame were used to match the current frame's fixation. For every detected AOI in a video frame, we determined if the fixation was found within the AOI's bounding box. In the simplest case, only a single AOI bounding box captured the fixation, making the AOI-to-fixation matching trivial. More challenging situations occurred when the fixation was located across two or more AOI bounding boxes. In such cases, we adopted a naive strategy of selecting the AOI with the minimum euclidean distance to the fixation point. This worked in our examples because the detected AOIs had, for the most part, small intersections.

The matching process between per-frame fixations and detected AOIs was then applied to the entirety of the training video. Using this approach, a timeline of AOI sequences was articulated to further observe gaze patterns during the simulation. In the past, learning analytics methods, such as sequence mining and other sequence-based analyses would have been used. However, these methods typically apply to event and action logs and may not be very suitable for eye-tracking data, which requires analyses at much lower levels of granularity. Also, gaze sequences can be noisy (see Fig. 4). Given these limitations in the traditional sequence analysis methods, we decided to use *network analysis* as an alternative technique.

3.1.3 Network Analysis. Previous work by Zhu & Feng, Schneider, Clariana, Ma, & Andrist [1, 10, 25, 34, 43] have applied network analysis to analyze eye-tracking data after AOI encoding. A graph is used to capture the transitions between AOIs and the fixation time associated with an AOI. In previous eye-tracking studies, the AOI transition graph is an accumulated representation of the experiment, where the total sums of fixation time and transition counts are mapped to node and edge weights. This approach is limited by the compression of the temporal dimension as it is unable to characterize the evolution of gaze behavior over the period of an experiment. This is especially important

in collaborative and complex environments, where the gaze behavior is expected to change to match new needs (e.g., information gathering versus interventions).

In this paper, we propose a novel representation for temporal analysis of eye-tracking data inspired by network analysis using the notion of *temporal fully-weighted graphs*. To take advantage of the expressiveness of these graphs, we applied temporal clustering using the standard K-Means algorithm to graph snapshots [42]. Time segments based on cluster ids were then characterized by the cluster centroids. The analytics were then mapped back to activities and tasks in the cognitive task model using network metrics applied to the centroids of the AOI-transition graphs.

Temporal Dimension: Evolution and Decay. Our method for accounting for the evolution and decay of the graph is through mutation of nodes and edges' weights. Before discussing the graph, it is important to consider that the construction of the temporal graphs includes hyperparameters for decay rate and sliding window size. For our study, we used a sliding window size of 15 seconds and a decay rate of 4%. These hyperparameters were qualitatively determined through a manual hyperparameter search, based on the frequency of gaze transitions and the stability of the graphs. Variations in the sliding window parameter had a minimal impact on the graphs' structures, as long as the size was less than 30 seconds. For the decay rate, too rapid a rate resulted in a diminishing graph, and too slow a rate caused the graph to continue accumulating edges and stagnated the network's typology. The selected hyperparameters were chosen based on maximizing the distance from the aforementioned extremes.

The directed graph is first initialized with all 9 AOI nodes with node weights and edges assigned 0 values. This corresponds to an empty adjacency matrix. Working with all nodes, even when they have zero weights, makes graph embedding and clustering easier by keeping the dimensions consistent. Through a non-overlapping sliding window approach, the temporal AOI sequences were added to the dynamic graph by summing the fixation time for each AOI to the node weights. The counts of transitions between AOIs were summed to the edge weights between the two corresponding nodes. If an AOI was not found within the sliding window, then its node and edge weights decayed at a constant rate. The decay, after initial focus, models how the nurses' attention to an AOI decreases over time as their gaze directions change.

Graph Embedding and Clustering. After developing the temporal graph, we used the K-Means algorithm from the scikit-learn library [28] to cluster 2250 snapshots of temporal graphs generated from all five training simulations so that we could generalize across simulation scenarios and participants. The goal was to group time segments by distinct gaze behaviors, and then find interpretations that may map the aggregated gaze behaviors to task model activities that nurses were performing in the simulated scenario. Each snapshot represented a static graph sample generated from the non-overlapping sliding window algorithm.

For clustering, given that our goal was to group gaze segments and represent each group by aggregated gaze behaviors, we chose not solely rely on overall network metrics, such as centrality and transitivity, which might make it harder to derive an interpretation for each group that can be linked to task model activities. Instead, we decided to use a lossless graph embedding as input to the clustering algorithm to later use the clusters' mean to construct a graph. The edges' weights from the adjacency matrix were concatenated with the nodes' weight vector into a single vector of length 91. Therefore, each of the 2250 data points was represented by 91 feature values. For interpretation purposes, we computed a mean 91-valued vector for each group so we could construct a graph with the corresponding node and edge weights representing each of the derived groups or clusters. These constructed graphs were then mapped to specific activities in our task model.

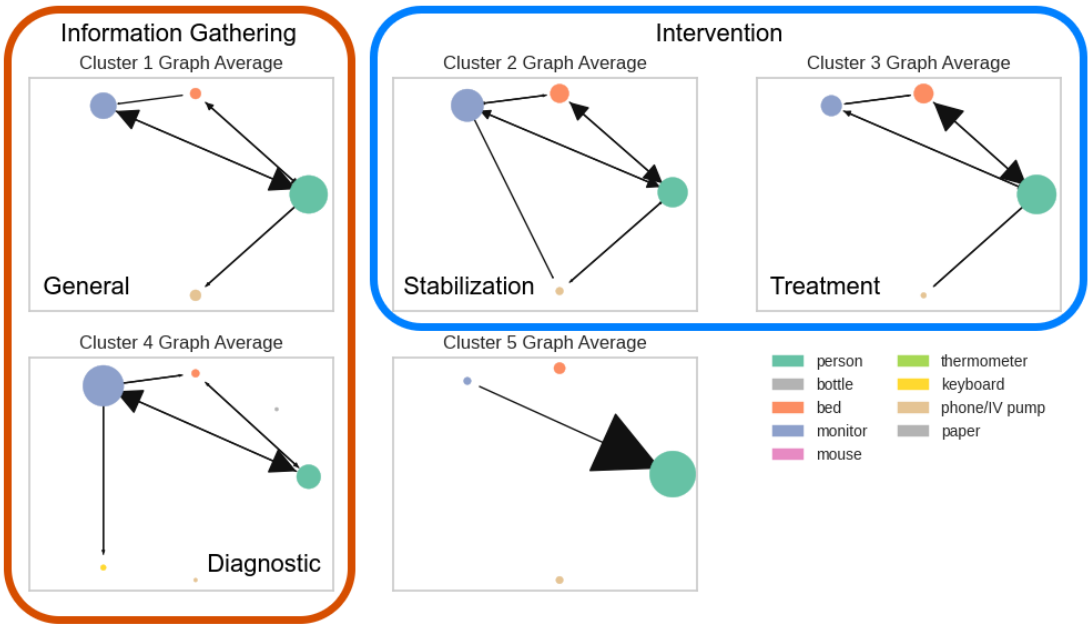

Fig. 5. Clustering 2250 snapshots of the temporal graphs with the K-Means algorithm produced five clusters. The clusters' mean vectors are then used to reconstruct a graph that represents the clusters' mean graphs. Through the clusters' mean graphs, domain-specific relationships can be more easily observed and identified, such as in Cluster 3's graph average, where there is a large edge between *person* and *bed*.

Cluster Number	Density	Centrality	Local Efficiency	Transitivity	Global Efficiency
1	0.097	0.089	0.259	0.375	0.139
2	0.125	0.071	0.407	0.643	0.153
3	0.111	0.071	0.259	0.600	0.139
4	0.111	0.071	0.259	0.600	0.139
5	0.014	0.000	0.000	0.000	0.028

Table 1. Network metrics for each of the clusters' means. Network analysis based on network metrics assists in graph interpretation through an aggregated domain-general lens; the general structure of the graph has been correlated to general cognitive strategies and behaviors [10]. For example, the centrality metric that ranges from $[0, 1]$ and its implications in the network typology has been correlated with a spectrum of strategies starting from naive (low centrality) to goal-focused (high centrality).

In any unsupervised clustering algorithm, the number of clusters can be considered a hyperparameter with an associated criterion function that is used to discover the optimal number of clusters for the data. We used the K-Elbow approach [35] that was readily available via the YellowBricks package [3] to find the optimal number of clusters. This approach produced five as the optimal number of clusters. Each of the cluster means was then computed and the resulting network graphs are displayed in Fig. 5. The aggregated network metrics are shown in Table 1.

Mapping Clusters to Cognitive Tasks. At first glance, it seems that the aggregated metrics for some of the clusters are identical (e.g., see feature values for clusters 2 and 3 in Table 1). But Fig. 5 shows clear differences based on edges and size of the nodes. This further emphasizes that network metrics may be insufficient to disambiguate clusters. Overall,

combining the graph structures in Fig. 5 with the aggregated statistical values from Table 1, we can derive general interpretations of the type of gaze behavior for each cluster and relate them to the nurses' activities and performance.

Cluster 1 and **Cluster 4** fall into information gathering since the cluster mean graphs have large node and edge weights between the *monitor* and *person* nodes, as these objects are the primary data sources available to the nurses. The monitors display the patient's vitals and medical history as well as the physician's orders that govern the intervention. The *person* category can imply both nurse-to-nurse and nurse-to-patient interactions. These gaze exchanges reflect visual inspection, attention, and discourse to obtain relevant information. The size of the *person* and *monitor* nodes is what differentiates **Cluster 1** and **Cluster 4** into general or diagnostic information gathering. **Cluster 1**'s larger *person* node shows a focus on generic procedures, such as conversation with the patient about his or her medical history, symptoms, and other generic information. In contrast, **Cluster 4**'s larger *monitor* node reflects a diagnostic approach to information gathering, as nurses read patients' charts and consider real-time vitals to hypothesize a diagnosis.

On the other branch of the cognitive task model, **Cluster 2** and **Cluster 3** represent intervention tasks, as their mean vectors have large node weights for *bed*, and the edge weight between *bed* and *person* is large. Note that attention to the *bed* node is analogous to gazing at the patient. The *bed* and *person* in this context can be associated with direct patient interactions, either by conversations or by physical interaction (e.g., administering medicine). In addition, **Cluster 2** and **Cluster 3** have a more even distribution of gaze between the *monitor*, *person*, and *bed* nodes, as the nurses track the intervention's impact on these three objects. **Cluster 2** and **Cluster 3** differ from one another in their distribution (connected to network density and centrality) of gaze and transitions along their graphs. **Cluster 2** has a more even weight distribution among its nodes and edges, implying an even distribution of attention and focus by a nurse. This aligns well with stabilization, as the nurses are performing intervention tasks on the patient and monitoring their effects to get the patient to a more comfortable state. **Cluster 3** differs from **Cluster 2** because of its greater weight on the *bed* to *person* edge, which is representative of actions that involve initiating new treatments, and at the same time, discussing these treatments with the patient, fellow nurses, and providers.

4 CASE STUDY

To evaluate our approach, we applied our proposed temporal graph representation, clustering, and mapping to the cognitive task model to the simulation scenario where all three nurses wore eye-tracking glasses as they trained in the scenario. In this analysis, we summarize our interpretation of the simulation from the eye-tracking data and then discuss a list of important training events that we derived by matching against our interpreted cluster segments that were then mapped back to understand the temporal sequence of the nurses' activities.

4.1 Simulation Summary

The simulation scenario began with the instructor assigning roles (P1 and P2 were assigned to be the primary nurses, and P3 was assigned the role of a care partner), giving a brief overview of the patient case, and recommending that the nurses start by reading the chart. The manikin patient was experiencing painful coughing episodes and had difficulty breathing. The nurses were tasked with providing general medical care, stabilizing the patient by easing his pain, and starting a treatment plan. In Table 2, key events in the simulation are presented and briefly described. Overall, the nurses were able to navigate the problem presented in the simulation and successfully resolve the medical issues that the patient was experiencing.

Event ID #	Timestamp (sec)	Event Description
I	0	Instructor discusses scenario information
II	40	Simulation start
III	168	Patient requests help
IV	212	Nurses implicitly split tasks
V	319	A nurse initiates nebulizer treatment but Patient contests
VI	371	Nurses consult then call respiratory therapist
VII	426	Nurses start administering IV fluids
VIII	771	Nurses try to provide medications but Patient contests
IX	866	IV pump administration complete
X	999	Nurses convince Patient and start administering antibiotic
XI	1404	Simulation end

Table 2. The key events that occurred in the simulation case study. Events I-IV were routine initial events in the simulation; event V marked the simulation timeline where the nurses faced new unexpected challenges. From event V to event X, large time intervals were composed of the nurses gathering equipment, requesting assistance from other medical personnel, and waiting for the completion of medication administration. Event XI marked the end of the simulation, after which the instructors discussed simulation outcomes and performance with the nurses.

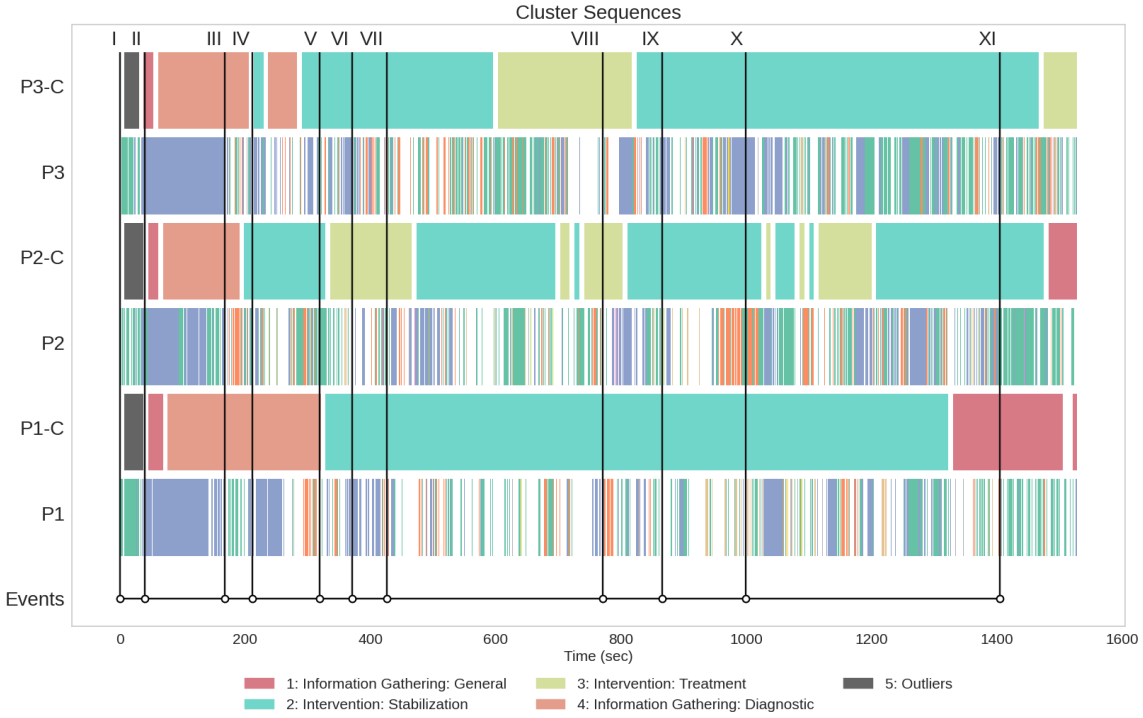

Fig. 6. Graph clustering applied to the case study simulation scenario along with significant events that occurred during the training exercise. The P1, P2, and P3 timelines are raw gaze, with PX-C being the gaze counterpart's time segments based on the clusters. These time segments are obtained by identifying the cluster that the gaze graph belongs to using our sliding window approach (size = 15 seconds). Through the simulation, there was a gradual transition from information gathering to intervention, and the nurses switched from stabilization to treatment and attempt to resolve the patient's medical problems.

4.2 Applying and Interpreting Graph Clusters

As a first step, we analyzed the simulation data to obtain cluster segments, as shown in Fig. 6. The events' ID and timestamp are included in the cluster segments figure to observe the transitions between cluster-defined activities. Also, key events that made the nurses rethink their approach were also observed. In the interval between Events I and II, we can see how Cluster 5 (outliers) represents the instructor's introduction to the scenario and the nurses' attention focused on the instructor. In the interval from Event II to III, all nurse activities fall into the realm of Clusters 1 (generic information gathering) and 4 (diagnostic information gathering). When the transition to Event III happened, and the patient asked for help, the nurses finished their information gathering and switched their attention to the patient, and administered interventions. During Event IV and V, all nurses independently transitioned to Cluster 2 (stabilization intervention). After Event V occurred, the nurses' tasks diverged with Nurse 2 switching to more stabilizing or treatment-focused interventions. The multiple occurrences of Cluster 3 (treatment intervention) matched events VI, VII, and VIII, which were all goal-oriented tasks that did not require monitoring. The simulation officially concluded with Event XI.

5 CONCLUSION

In this paper, we presented a novel approach for scanpath network analysis by temporally parsing simulation gaze events using a combined evolution and decay modeling approach. Our mixed-reality simulation environment includes the physical layout of a hospital room, where the nurses move about to monitor and administer interventions to stabilize the patient. All of this results in dynamic AOIs that we capture as egocentric views for each nurse using the eye-tracking glasses they wear. The YOLOv5 model was used to detect and track AOIs that were matched with fixations to construct AOI sequences that evolved in time as the simulation progressed. The temporal weighted graphs (with a node weight corresponding to gaze durations and edge weight corresponding to the number of gaze transitions between nodes) were generated by segmenting the duration of the entire simulation into a sequence of non-overlapping segments using a sliding window approach.

Snapshots of these temporal graphs from all five simulation scenarios with different participants produced a total of 2250 data points (each one derived from a weighted graph) with 91 feature values to preserve edge and node information. We clustered these data points using the K-Means algorithm. The K-elbow approach provided the optimal number of clusters, which was five. Each cluster was then represented by its mean vector as its constructed into a fully-weighted graph. For the constructed mean graphs, we computed individual network metrics, contextualized them using their node and edge weights, and mapped the clusters to task and activity components in the cognitive task model.

To demonstrate our approach, we conducted a case study of the simulation scenario where all three nurses wore eyeglasses during their training. We used the methods described above to segment the timeline into cluster segments. Key events were then used to interpret and evaluate how the gaze behaviors represented by each cluster explain the activities that the nurses conducted in that time interval.

5.1 Addressing Research Question

To address our research question, **RQ1**, we decompose the question into two elements: (1) contextualization and (2) in-the-moment interpretation. In terms of contextualization, we showed how we can elevate low-level eye-tracking data systematically to higher levels of abstraction to interpret nurse interactions and activities. This involved AOI encoding that mapped gaze (x, y) coordinates to objects of interest in the environment. For our physical environment, we used

the YOLOv5 object detection algorithm to track dynamic AOIs. The second step used a graph representation to model the gaze information in terms of AOI sequences that captured the change in attention between objects.

For in-the-moment information representation, we differentiated the temporal representations by regenerating the graph representation using evolution and decay parameters. Through our observations, the temporal fully-weighted graph yielded a direct mapping from the nurses' gaze behaviors to tasks and activities that we defined by cognitive task analysis of the simulation scenarios.

5.2 Limitations & Future Work

We have noted several limitations in our first approach to using multimodal data, primarily the egocentric eye-tracking data collected to interpret, analyze, and evaluate nurse training behaviors in our mixed-reality training environments. The initial step, dynamic AOI tracking using YOLOv5, produced semantic category representations that did not disambiguate between instances of the same object category. For example, aggregating fixations to the general category of *person* missed relevant social interaction information, as it did not distinguish between the different nurses. In future work, we hope to match *person* AOIs to specific individuals (e.g., patients, nurses P1, P2, and P3), to better capture the information flow and social interactions. Second, YOLOv5 frequently misclassified various essential yet challenging objects such as IV lines, syringes, and other medical equipment. A fine-tuned YOLOv5 for the nursing domain would have produced considerably better results than the domain-general pre-trained model that we used. Last, some of the network metrics used in our paper (e.g., centrality, transitivity, and density) only used edge weights and the network typology but did not account for the node weights. Fully-weighted network metrics are not as widely supported in graph software packages and are still an active field of research in network analysis. The use of fully-weighted network metrics would help further characterize this type of graph and allow us to derive clearer interpretations.

6 AUTHOR CONTRIBUTIONS

ED, the primary author of the paper, developed, programmed, and applied the proposed approach and wrote the manuscript's initial draft. He also led the effort on revising the paper to address the reviewers' comments. ML, CV, CC, and ED were responsible for data collection, annotation, and curation. GB and DL are the principal investigators, responsible for leading the study's vision and scope. DL contributed to the overall framework for interacting with the nursing school, funding, and IRB approval. GB provided guidance in the development of the overall technical framework and the machine learning-based analysis methods for the paper. GB also put in a considerable amount of effort in revising the paper and addressing all of the primary comments made by the reviewers. All authors contributed to brainstorming, approach development, and manuscript revision. All authors approved the submitted version.

ACKNOWLEDGMENTS

The authors wish to show their appreciation to the Nursing Simulation Lab instructors, including Eric Hall, Jo Ellen Holt, and Mary Ann Jessee, for inviting us to collect data and providing their nursing-domain expertise. Additionally, we thank the paper's anonymous reviewers for their insightful comments.

This study is partially funded by Army Research Laboratory Award 1150 W912CG2220001, NSF Cyberlearning Award 2017000, and Vanderbilt Chancellor's funding of the LIVE (Learning Incubator: A Vanderbilt Enterprise) Initiative, a joint program between the Peabody College of Education and the Department of Computer Science.

REFERENCES

[1] Sean Andrist, Wesley Collier, Michael Gleicher, Bilge Mutlu, and David Shaffer. 2015. Look together: analyzing gaze coordination with epistemic network analysis. *Frontiers in Psychology* 6 (July 2015). https://doi.org/10.3389/fpsyg.2015.01016

[2] Roger Azevedo, Daniel C Moos, Amy M Johnson, and Amber D Chauncey. 2010. Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. *Educational psychologist* 45, 4 (2010), 210–223.

[3] Benjamin Bengfort, Rebecca Bilbro, Nathan Danielsen, Larry Gray, Kristen McIntyre, Prema Roman, Zijie Poh, et al. 2018. *Yellowbrick*. https://doi.org/10.5281/zenodo.1206264

[4] Gautam Biswas, Ramkumar Rajendran, Naveeduddin Mohammed, Benjamin S. Goldberg, Robert A. Sottilare, Keith Brawner, and Michael Hoffman. 2020. Multilevel Learner Modeling in Training Environments for Complex Decision Making. *IEEE Transactions on Learning Technologies* 13, 1 (Jan. 2020), 172–185. https://doi.org/10.1109/TLT.2019.2923352

[5] Gautam Biswas, James R. Segedy, and Kritya Bunchongchit. 2016. From Design to Implementation to Practice a Learning by Teaching System: Betty's Brain. *International Journal of Artificial Intelligence in Education* 26, 1 (March 2016), 350–364. https://doi.org/10.1007/s40593-015-0057-9

[6] Ann Blandford and Dominic Furniss. 2006. DiCoT: A Methodology for Applying Distributed Cognition to the Design of Teamworking Systems. In *Interactive Systems. Design, Specification, and Verification*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Stephen W. Gilroy, and Michael D. Harrison (Eds.). Vol. 3941. Springer Berlin Heidelberg, Berlin, Heidelberg, 26–38. https://doi.org/10.1007/11752707_3 Series Title: Lecture Notes in Computer Science.

[7] Paulo Blikstein. 2013. Multimodal learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13*. ACM Press, Leuven, Belgium, 102. https://doi.org/10.1145/2460296.2460316

[8] Paulo Blikstein and Marcelo Worsley. 2016. Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics* 3, 2 (Sept. 2016), 220–238. https://doi.org/10.18608/jla.2016.32.11

[9] Nichol Castro and Cynthia S. Q. Siew. 2020. Contributions of modern network science to the cognitive sciences: revisiting research spirals of representation and process. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 476, 2238 (June 2020), 20190825. https://doi.org/10.1098/rspa.2019.0825

[10] Roy B. Clariana, Tanja Engelmann, and Wu Yu. 2013. Using centrality of concept maps as a measure of problem space states in computer-supported collaborative problem solving. *Educational Technology Research and Development* 61, 3 (June 2013), 423–442. https://doi.org/10.1007/s11423-013-9293-6

[11] Andy Clark. 1997. *Being there*. MIT Press Cambridge, MA.

[12] Richard E. Clark and Fred Estes. 1996. Cognitive task analysis for training. *International Journal of Educational Research* 25, 5 (Jan. 1996), 403–417. https://doi.org/10.1016/S0883-0355(97)81235-9

[13] Michael Cole. 1998. *Cultural psychology: A once and future discipline*. Harvard university press.

[14] Cristina Conati, Vincent Aleven, and Antonija Mitrovic. [n. d.]. CHAPTER 21 –Eye-Tracking for Student Modelling in Intelligent Tutoring Systems. 1 ([n. d.]), 10.

[15] J B Cooper and V R Taqueti. 2008. A brief history of the development of mannequin simulators for clinical education and training. *Postgraduate Medical Journal* 84, 997 (Nov. 2008), 563–570. https://doi.org/10.1136/qshc.2004.009886

[16] Philip Dickison, Katie A. Haerling, and Kathie Lasater. 2019. Integrating the National Council of State Boards of Nursing Clinical Judgment Model Into Nursing Educational Frameworks. *Journal of Nursing Education* 58, 2 (Feb. 2019), 72–78. https://doi.org/10.3928/01484834-20190122-03

[17] Upamanyu Ghose, Arvind A. Srinivasan, W. Paul Boyce, Hong Xu, and Eng Siong Chng. 2020. PyTrack: An end-to-end analysis toolkit for eye tracking. *Behavior Research Methods* 52, 6 (Dec. 2020), 2588–2603. https://doi.org/10.3758/s13428-020-01392-6

[18] Pål A. Hegland, Hege Aarlie, Hilde Strømme, and Gro Jamtvedt. 2017. Simulation-based training for nurses: Systematic review and meta-analysis. *Nurse Education Today* 54 (July 2017), 6–20. https://doi.org/10.1016/j.nedt.2017.04.004

[19] Roy S. Hessels and Ignace T.C. Hooge. 2019. Eye tracking in developmental cognitive neuroscience – The good, the bad and the ugly. *Developmental Cognitive Neuroscience* 40 (Dec. 2019), 100710. https://doi.org/10.1016/j.dcn.2019.100710

[20] Edwin Hutchins. 1995. *Cognition in the Wild*. MIT press.

[21] Edwin Hutchins. 2000. Distributed Cognition.

[22] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Kalen Michael, Jiacong Fang, imyhxy, Lorna, Colin Wong, Zeng Yifu, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Max Strobel, Mrinal Jain, Lorenzo Mammana, and xylieong. 2022. *ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations*. https://doi.org/10.5281/zenodo.7002879

[23] John S Kinnebrew, Kirk M Loretz, and Gautam Biswas. 2013. A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining* 5, 1 (2013), 190–219.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. http://arxiv.org/abs/1405.0312 arXiv:1405.0312 [cs].

[25] Xiaochuan Ma, Yikang Liu, Roy Clariana, Chanyuan Gu, and Ping Li. 2022. From eye movements to scanpath networks: A method for studying individual differences in expository text reading. *Behavior Research Methods* (April 2022). https://doi.org/10.3758/s13428-022-01842-3

[26] Laura G. Militello and Robert J. B. Hutton. 1998. Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands. *Ergonomics* 41, 11 (Nov. 1998), 1618–1641. https://doi.org/10.1080/001401398186108

[27] René Noël, Diego Miranda, Cristian Cechinel, Fabián Riquelme, Tiago Thompsen Primo, and Roberto Munoz. 2022. Visualizing Collaboration in Teamwork: A Multimodal Learning Analytics Platform for Non-Verbal Communication. *Applied Sciences* 12, 15 (July 2022), 7499. https://doi.org/10.3390/app12157499

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[29] Christoph Pimmer, Norbert Pachler, and Urs Genewein. 2013. Reframing Clinical Workplace Learning Using the Theory of Distributed Cognition:. *Academic Medicine* 88, 9 (Sept. 2013), 1239–1245. https://doi.org/10.1097/ACM.0b013e31829eec0a

[30] Patricia Ravert. 2002. An Integrative Review of Computer-based Simulation in the Education Process:. *CIN: Computers, Informatics, Nursing* 20, 5 (Sept. 2002), 203–208. https://doi.org/10.1097/00024665-200209000-00013

[31] Ognjen Rudovic, Meiru Zhang, Bjorn Schuller, and Rosalind W. Picard. 2019. Multi-modal Active Learning From Human Data: A Deep Reinforcement Learning Approach. http://arxiv.org/abs/1906.03098 arXiv:1906.03098 [cs, stat].

[32] Jonas Rybing, Erik Prytz, Johan Hornwall, Heléne Nilsson, Carl-Oscar Jonson, and Magnus Bang. 2017. Designing a Digital Medical Management Training Simulator Using Distributed Cognition Theory. *Simulation & Gaming* 48, 1 (Feb. 2017), 131–152. https://doi.org/10.1177/1046878116676511

[33] Maike Schindler and Achim J. Lilienthal. 2019. Domain-specific interpretation of eye tracking data: towards a refined use of the eye-mind hypothesis for the field of geometry. *Educational Studies in Mathematics* 101, 1 (May 2019), 123–139. https://doi.org/10.1007/s10649-019-9878-z

[34] Bertrand Schneider, Sami Abu-El-Haija, Jim Reesman, and Roy Pea. 2013. Toward collaboration sensing: applying network analysis techniques to collaborative eye-tracking data. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13*. ACM Press, Leuven, Belgium, 107. https://doi.org/10.1145/2460296.2460317

[35] Edy Umargono, Jatmiko Endro Suseno, and S.K Vincensius Gunawan. 2020. K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula:. In *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)*. Atlantis Press, Yogyakarta, Indonesia. https://doi.org/10.2991/assehr.k.201010.019

[36] Caleb Vatral, Gautam Biswas, Clayton Cohn, Eduardo Davalos, and Naveeduddin Mohammed. 2022. Using the DiCoT framework for integrated multimodal analysis in mixed-reality training environments. *Frontiers in Artificial Intelligence* 5 (July 2022), 941825. https://doi.org/10.3389/frai.2022.941825

[37] Marcelo Worsley. [n. d.]. Multimodal Learning Analytics' Past, Present, and, Potential Futures. ([n. d.]), 16.

[38] Marcelo Worsley, Dor Abrahamson, Paulo Blikstein, Shuchi Grover, Bertrand Schneider, and Mike Tissenbaum. 2016. Situating Multimodal Learning Analytics. (2016), 5.

[39] Marcelo Bonilla Worsley. 2012. Multimodal learning analytics: enabling the future of learning through multimodal data analysis and interfaces. In *ICMI '12*.

[40] Peter C Wright. [n. d.]. ANALYSING HUMAN-COMPUTER INTERACTION AS DISTRIBUTED COGNITION: THE RESOURCES MODEL. ([n. d.]), 58.

[41] Kaya Yilmaz. 2011. The cognitive perspective on learning: Its theoretical underpinnings and implications for classroom practices. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas* 84, 5 (2011), 204–212.

[42] Chunhui Yuan and Haitao Yang. 2019. Research on K-Value Selection Method of K-Means Clustering Algorithm. *J* 2, 2 (June 2019), 226–235. https://doi.org/10.3390/j2020016

[43] Mengxiao Zhu and Gary Feng. 2015. An exploratory study using social network analysis to model eye movements in mathematics problem solving. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. ACM, Poughkeepsie New York, 383–387. https://doi.org/10.1145/2723576.2723591