Toward Multimodal Situated Analysis

Chreston Miller
Center for HCl, Virginia Tech
2202 Kraft Drive, Blacksburg, Va. 24060, USA
chmille3@vt.edu

Francis Quek
Center for HCI, Virginia Tech
2202 Kraft Drive, Blacksburg, Va. 24060, USA
quek@vt.edu

ABSTRACT

Multimodal analysis of human behavior is ultimately situated. The situated context of an instance of a behavior phenomenon informs its analysis. Starting with some initial (user-supplied) descriptive model of a phenomenon, accessing and studying instances in the data that are matches or near matches to the model is essential to refine the model to account for variations in the phenomenon. This inquiry requires viewing the instances within-context to judge their relevance. In this paper, we propose an automatic processing approach that supports this need for situated analysis in multimodal data. We process events on a semi-interval level to provide detailed temporal ordering of events with respect to instances of a phenomenon. We demonstrate the results of our approach and how it facilitates and allows for situated multimodal analysis.

Categories and Subject Descriptors
I.2.4 [Artificial Intelligence]: Knowledge Representation
Formalisms and Methods—relation systems, temporal logic

General Terms
Algorithms, Design

Keywords
Situated Analysis, Multimodal Analysis, Temporal Relations

1. INTRODUCTION

Multimodal analysis of human behavior is ultimately situated. Consider an analysis of several students solving a math problem together with a focus on how the students ‘co-construct’ space together in the course of their discussion. This co-construction is done through pointing (where spatial reference may be initiated or referenced), gaze (as students co-attend to the shared space that now has meaning), and other gestural forms (that shape the space). The students’ gaze, speech, and gesture channels are recorded into a database. During their session, a door slams outside their workspace, at which point all the students turn around. The question is: “Is this very strong, temporally coherent alignment of gaze important?” Looking at the gaze data alone, the strong temporal and spatial alignment may suggest an important event has occurred, but for co-construction of space, it is probably unimportant. The only way to judge relevance is to look at the specific instance.

However, what if the analysis is focused on how an external attention-synchronizing event may alter shared decision making, e.g., clearing out irrelevant distractor concepts when the group returns and picks up where they left off. In this case, the door-slam incident and gaze alignment phenomenon is very relevant. One analyst’s data is another analyst’s noise. The tension is that we want to have machine help to filter events and focus analysis.

The question is how to solve the problem of filtering and focusing analysis. Such detailed analysis is characterized by focused attention on particular instances of behavior within-context where inter-relations (temporal order) among events in the channels for a particular instance are analyzed. This process is time consuming and very tedious as some form of manual inspection referencing the original media context (e.g., video) is necessary. The situated contexts of individual instances are very important - an instance being a particular occurrence of a phenomenon within the data. We propose a situated analysis approach that supports viewing instances of a phenomenon within-context and automates processing of temporal ordered relationships across channels. Our approach is to allow the analyst to define a model of a phenomenon, guided by what is in the data, then, to identify and locate instances of the model (and sub-structures) within the data and show the context(s) of the occurrence(s). Typically, the data consists of time intervals describing observed events, e.g. “Student 1 begins a gaze event at Student 2 at time $t_i$ and ends at time $t_j$”, for $i ≥ 0$ and $j > i$. Flexible representation of events (and the model) are provided through semi-interval processing [9] where a time interval of an event is viewed as comprising of beginning and end atomic units (semi-intervals). The semi-intervals are viewed as base representative units as opposed to a complete interval. We process these time intervals to extract the necessary information that supports our approach.

Multimodal analysis integrates multiple data channels (e.g., gaze, gesture, and speech) where incorporation of such information provides more accurate analysis results [35]. These data channels are conduits of data streams that contain...
events describing actions in the data-set. We extract “within-context temporal relations” with respect to the model from these data streams. We abstract away from the data streams and focus on the situated events that describe the observed actions. Then for each instance found, the situated events can be viewed, providing the analyst the opportunity to judge the relevance of the instance with respect to their analysis goal and to further investigate other identified instances.

In §2 we provide an overview of our approach. We then review related work in §3. §4 describes multimodal data streams in terms of the events they contain and how we abstract away from streams and focus on viewing the events they contain. Afterwards, §5 describes how our approach extracts temporally ordered relations between events in all data streams and the process of identifying specific instances of a phenomenon. §6 details how we support situated analysis through automated means by combining the components described. §7 follows with a discussion of example analyses using our approach on multimodal human behavioral corpus. Lastly, §8 provides conclusions and discussion of future work.

2. APPROACH OVERVIEW

Here we describe an overview to our approach that will serve as the foundation to the rest of the paper.

Our approach specifically seeks to incorporate the knowledge and guidance of the domain expert (such as a psycholinguist, or other behavioral expert) in human behavior analysis tasks. Conversely, we want automatic processing to support domain experts as they apply their knowledge in exploring multimodal data. Other attempts at providing automation support for analysis include [40] and [5]. As models of phenomena are created and discovered (through exploration and refinement), the identification of other instances of the models is needed. Our approach is to provide automatic processing of temporal relations between data stream events and view such relations situated (within-context) within instances of a phenomenon.

We provide support in three ways. First, we view the analysis process as beginning with the domain expert advancing a hypothesis (model) of a behavior phenomenon and proceeding to explore instance(s) of the phenomenon (point of focused attention). How the phenomenon is present in the data may be unknown at the start of analysis, hence, the expert starts to explore the data with an initial idea. This initial idea is likely to be an incomplete model but a general structure of the phenomenon. Second, our approach can identify instances of this model in the data and present related events to each specific instance. Related events refer to events found in context to a specific instance and their respective temporal relations to the instance. The related events presented are from all data streams enabling situated analysis as we extract within-context information of a domain expert’s model. Third, all instances of the model are identified, allowing a comparison of relations across contexts which may help in formulating new insights. With these three components, our approach allows a situated view of cross-stream relations with respect to a phenomenon.

3. BACKGROUND AND RELATED WORK

Temporal Event Data: The ordered relations that are inherent to temporal event data is an active area of research. Allen in [1] formulated thirteen relationship principles that express all the possible orderings between two event intervals. These orderings describe how events between data streams relate. Research, such as [11, 13, 26, 27, 38], has focused on processing numerical or descriptive univariate, multivariate, or symbolic temporal data with the goal of discovering events and temporal relationships from temporal data. Others have explored the discovery of temporal knowledge and associated patterns, such as [12, 19, 28].

The stream data of multimodal data-sets is a mixture of ordinal and nominal data. As per the temporal data models reported in [28], gesture traces are a collection of univariate numerical time series (ordinal) while speech and gaze target fixations are multivariate symbolic interval series in which their meaning is categorically descriptive (nominal). Explorations into the relational structures of nominal data can be seen in [6, 20, 43, 44]. More specifically, our prior research [25] explores how to structure nominal event data based on semi-interval processing using n-gram techniques from speech processing. Other research [9, 29] has explored the concept of breaking temporal events into core units of beginnings and endings through semi-interval temporal ordered processing. Previously mentioned research ([1, 13, 27, 38]) has investigated patterns and relations among interval data irrespective of the data being ordinal or nominal.

Multimodal Analysis: Multimodal analysis of human behavior typically involves multiple time-synchronized data streams which consist of different observable human behavior signals (e.g. different gesture types, gaze fixations, and speech). The streams are then annotated either by hand, automatically, or a mixture of both. Then manual analysis, normally with some machine automated help, of the streams is performed. Example application and results of this process can be seen in [5, 4, 16, 18, 23, 31, 34, 42].

There has been much work in automatically detecting and recognizing certain behavior, a few examples being [10, 30, 41]. However, it is the more detailed analysis prominently exemplified by the math students example that is much more difficult to automate and is the kind of analysis where human guidance is needed. The goal of this detailed process is to explore the data and to identify and understand instances of phenomena. This is done through inspection of the temporal relations of the events within the streams. However, several challenges surface in this process. The identification of instances of a phenomenon that are relevant to the analysis being conducted can be difficult. Plus, coordination between the multiple streams requires careful attention as seen in [5, 35]. Both of these challenges were exemplified in the previously described analysis scenario of math students.

In terms of the analysis focus, we have observed two viewpoints employed for multimodal analysis of human behavior. The first focuses on identifying behavior instances within sections of recorded data (e.g. recorded video/audio sessions). This is a useful and powerful means as it shows the context in which the behavior occurs. However, what if an expert is interested in identifying instances of a particular behavior (e.g., co-construction of space) and viewing all the contexts in which they are situated. The second viewpoint is the reverse of the first by showing the behavior within-context. This is a plausible approach since the focus of some analyses is instances of a specific behavior. Examples of this can be seen in [5] where cues were identified within a group meeting setting that could signal floor
control events. Other research [34] has found cues in gaze and gestures that aid in discourse analysis. These cues can be flags in the identification of specific behavior of interest for which the expert would want to view the different instances within their respected situated contexts. This prior research exemplifies that the difficulty is in identifying how phenomena are manifested within the corpus. By taking the viewpoint of behavior within-context, we can begin with a general structure (model), or sub-structure, that describes the phenomenon and see how events within the corpus are related to the structure.

4. MULTIMODAL DATA TO EVENTS

The first component of our approach has the expert start with a hypothesis with which to identify instances. Hence, we need a way to describe the multimodal data to allow for defining a model of the hypothesis. We want to abstract away from data streams and view the descriptive events. In this section we define data streams and the events they produce and how this desired abstraction can be obtained.

Multimodal data typically exist in the form of data streams (time series of data points and signals) and music score data. Music score data are multivariate time series data that describe event intervals in time that possibly overlap. A number of multimodal analysis tools (e.g., [17, 36]) employ this form of multimodal data representation. Our first step is to convert these data streams into a homogeneous stream of events. An event is an interval in time representing when an action begins and ends. It comprises of a beginning, end, and description. Figure 1A shows three examples of events, A, B, and C, with their respective beginning (start) and end times. Viewing events with respect to their beginning and end is known as semi-interval processing [9] where a beginning or end is a semi-interval. Sometimes it is difficult to determine what is an event within a data stream (e.g., deciding exactly when an event begins and ends) however, we assume such detection occurs prior to our processing. Our focus in this paper is not on how to extract events from data, but on how to interpret these events after they are extracted.

Multimodal analysis comprises multiple data streams by definition. Hence, there will be overlaps between the multiple streams and cross-modal relations supportive of multimodal analysis. In Figure 1B, we see three data streams, Student 1’s speech, Student 2’s gaze, and Student 3’s gestures, with respective events. It can easily be observed how these represented events occur in overlap.

We abstract away from various specific representations in the separate data streams and focus on the events they contain while keeping the context (data stream) from which they originate. The strict notion of data streams fades into the background as the events become a prevalent part of the analysis. It is the events and what they represent that aid in the analysis. This approach allows the flexibility of processing that is mode independent where the kind and number of modes used does not matter.

Figure 1C shows an example model of a hypothesis an expert might use to find instances of attention-synchronizing events during the math student session described earlier. Assuming three students are involved, the expert specifies a model where the student’s gaze change at the same time. This involves specifying the temporal order of start gaze change semi-intervals for each student. In this case, the order represents concurrent action (equality). This will identify instances where the students have a synchronized gaze change. This example shows how we abstract away from the idea of three gaze data streams and focus on the gaze events of the students.

5. MODEL ASPECTS

Following from our first component, temporal relations between the model and related events is needed. The coordination between events within multiple data streams is necessary to understand a specific model instance. The temporal order in which events occur is key in describing a specific phenomenon, e.g., the coordinated gaze in response to a door-slam describes an attention-synchronizing phenomenon. This leads to the need of our second component which is identifying an instance of a model and extracting within-context temporal information between the model and related events. First we will address extracting the temporal information, after which, we will describe the process for identifying a model instance.

5.1 Event Temporal Relations

The challenge is temporal relations describing phenomena are structural in nature and not a set of parameters. Such a structure must be learned. Consider the scenario where a group of four History students use a horizontal, high-resolution, interactive tabletop display [24] to perform analysis of information snippets on virtual notecards (displayed on the table) [2]. The students were not given any instructions for making sense of the data, however, over time their actions around the shared space evolved into a ratification process in which changes to the notecard placement and utilization of the display space could be made. This process begins with the wielder of the interaction device announcing a statement or question around a piece of data (announcement) which lead to the coordinated gaze of the other three students (co-gaze). The wielder and other student(s) then proceeded to discuss a piece of data and decide (ratify) what to do with it.

The students’ interaction resulted in a sequence of proposal, echo, and act with enunciation that produced a ratification process which advanced the joint project and common ground. This temporal sequence exemplifies the structural nature of temporal relations in multimodal data. Other approaches, such as HMMs [37] and n-grams [3, 39], can be viewed as piece-wise parametric models that get some sense of overall structural relations. The problem is that approaches such as these ‘linearize’ behavior structure of these piece-wise models and do not address how to combine these...
linearized pieces structurally. Research in the numerical domain has also investigated structural means to numerical model modification, for example [8, 14, 21].

We approach representing temporal relations structurally by viewing events from the semi-interval level and how such atomic units of events relate temporally. Other research has investigated the use of semi-intervals for its representational flexibility with respect to incomplete or partial knowledge in identifying event interval sequences of interest [29]. Our prior work [25] investigated the use of n-gram processing of semi-intervals for providing comparison means for events in event sequences in which the processing takes advantage of the structure of the events.

With this structural approach, we provide automatically extracted temporally ordered relations within and between streams. This extracted information includes how semi-intervals relate with respect to order (e.g., “<”, and “=”), and their likelihood with respect to a model. Given a model consisting of semi-intervals, related events to each semi-interval are classified occurring either before, current (concurrent), or after. This is the same classification as used in [25] with the addition of current.

To obtain this information, we first segment all interval events into semi-intervals. Next, all the events from all data streams are serialized into one linear encoding with temporal ordering preserved through the use of ordered relations. An example of this can be seen in Figure 2 where the events from 1A and B are linearized. Due to this serialization process, results will contain temporal ordered relations within and between data streams as seen in Figure 2B. Here, the semi-interval events from multiple data streams are integrated into one sequence allowing comparison across streams and handling of overlap. Hence, given a focal point (model) in the data, one is able to view events across streams that relate to the focal point within-context. Semi-intervals that co-occur are set to be equal (‘=’) preserving their co-occurrence (e.g., B₁ and Cₛ in Figure 2).

![Figure 2: A) Example of linearizing events and preserving their temporally ordered relations. B) Example of how serialization of multiple streams allows for comparison across them.](image)

### 5.2 Model Instance Identification

The process of how we identify an instance of a model and subsequently provide related event semi-intervals to the model can be seen in Figure 3. To identify an instance given a model (in terms of semi-intervals), we search the data-set for matches to the semi-intervals (1). In this step, multiple matches may be found as can be seen from the two potential models. This occurs because events may repeat. A first approximation may be to assume that the closer of two events of the same type is the better candidate. Hence, we search for the match where all the semi-intervals are temporally closest. This mainly applies to semi-intervals of a different event type, e.g., Aₛ and Bₛ or a gaze event versus a speech event. Hence, Potential Model 2 is chosen. We also process sub-structures (sections) of the model to detect overlap, e.g., Dₛ overlaps with A. Note that identified instances may have semi-intervals that “interrupt” the instance, e.g., event H occurs between A and B. After identification is complete, the related semi-intervals to the model are extracted using n-gram processing at the semi-interval level (2). The set of related semi-intervals are then presented as relating to the model (3). This set also includes events that may only relate to a substructure of the phenomenon but are part of the same instance. This allows for identification of overlap and semi-intervals not seen by viewing the whole model solely, as seen in Figure 3 step (2).

### 6. Assisted Situated Analysis

The identification of an instance of the hypothesis (model) allows a within-context view of related events. However, the expert is interested in all instances of the hypothesis, leading to a need to view the multiple instances of the hypothesis within context and, potentially, to compare them. In this section, we address our third component through initial discussion of identifying all instances of a phenomenon, how an expert can toggle between them, and the resulting aggregate view of all the results. After which, we discuss how our approach facilitates comparison across contexts.

To identify all instances of a phenomenon, the process described in Figure 3 is repeated until all instances have been identified. The expert can then step through all unique instances, observing the related events and the differences and similarities. Through this process we can support analysis of the interrelationships among events with respect to the model wherever it occurs in situ, hence providing relations within-context and across data streams. We call this assisted situated analysis as we use automation to identify instances and the associated temporal relations for each instance and provide support to the expert in viewing and exploring this information.

We also allow the expert to view the related events from all contexts at once. An example of this can be seen in Figure 4 where a model and sub-structures are identified in three contexts and the related events within each context are aggregated to allow a view of all related events. With the aggregation of all results, the likelihood that the events occurred given all the contexts is provided when comparing the related events. The likelihood calculation is with respect to the whole model and its sub-structures and follows the same procedure of our prior work [25]. This aids in comparison among the instances extracted and also provides information on the event frequency with respect to the model. The number of semi-intervals used in calculating the likelihood is given as a confidence level as it is a measure of the amount of information used in the calculation. We are more confident in a result that incorporated more information.

An interesting artifact of this process can be seen in Figure 4 where the frequency of G may suggest an event of interest. However, the frequency is only seen through comparing multiple contexts. The aggregation of the results from all instances of the model and sub-structures allows for the opportunity to discover new temporal relations between events and the model. Alternatively, H occurs only once but may hold equally or more important value than G’s frequency (frequent vs. infrequent). Presenting the results from all contexts provides the opportunities for the researcher to investigate variations of the model (through inspection of their
original context) and decide their relevance to the current analysis. Hence, given the above procedure, we are able to provide assisted support to multimodal situated analysis.

<table>
<thead>
<tr>
<th>Model of Phenomenon</th>
<th>Aggregated Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: Identification of a model and sub-structures in several contexts and the aggregation of related events across contexts. Overlap of G after model event B represents its multiple occurrence.

7. Implementation and Use

Here we present example analyses that illustrate how our approach is able to provide support for multimodal situated analysis. A real data-set was chosen to exemplify our approach. The data-set consists of meetings of U.S. Air Force officers from the Air Force Institute of Technology (AFIT) engaging in military wargaming scenarios [4]. In this section we discuss the implementation of our approach, the data-set used, and example analyses.

7.1 Implementation

The core algorithm of our approach is based on the system described in [25] with a number of updates and enhancements. An interactive layer was added to allow user interaction and viewing of the context of any specific model instance and its respective related events. Implementation is in Python using QT 4.5 [33] for the user interface, and the Pymunk physics engine [32] (a Python wrapper of the C library Chipmunk [7]) is used as a means of collision detection for creating the spatial layout of the graphical semi-interval instances. Although it is necessary to show our user interface to explain our results, the interface is not the focus of the current paper. Our focus is on the event processing.

7.2 Demonstration Domain

We demonstrate key aspects of our approach with a comparison of an original manual analysis of the AFIT data. There are multiple days of data recordings that have been carefully hand annotated by behavior domain experts. We discuss a study focused on exploring one session in which the officers (labeled C, D, E, F, and G) are discussing potential candidates for a scholarship. The scenario is that C, D, F, and G are department heads meeting with the institute commandant E to select three scholarship award recipients. It was discovered that in such meetings, the social interaction among the participants have as much to do with the outcome of the meeting as the specific merits of the scholarship candidates being discussed. The participants dynamically formed coalitions to support each-other’s candidates through a process of mutual gaze fixations and back-channel expressions of support during discussions [29].

This session is approximately 45 minutes long. A coalition to support a proposed idea is initialized when the proponent of the idea seeks to make eye-contact with other participants while he is presenting the idea. Participants who supported the idea would return the eye-contact, while those who disagreed with the idea would avert gaze. When a return gaze is attained, the presenter’s gaze moves to another member. This phenomena was recorded within this scholarship session and we want to compare the results of our extraction methods against this real world recorded scenario. This scenario describes C’s sequence of gaze events starting with one to G, then D, then E. During this sequence, D and E’s gaze is fixated on C.

There are three key areas we want to demonstrate. The first is our ability to automate the processing of cross-modal relations. Given a snapshot of a real analysis scenario, we want to compare our automatically extracted relation information with the relation information employed in the original analysis. Second, closely tied to the first, we would like to show how our approach is able to report related events that are within-context to a respective model. Then, switch between the different identified contexts to view different related events. Lastly, we will show how we can view the related events of all contexts at the same time allowing an aggregate view. Through this we will show how we can discover variations not seen in the original analysis.

7.3 Analyses Examples

Single Context Cross-modal Relations: First, we demonstrate the ability to automate processing of cross-modal relations. The emphasis in the original analysis was observing the gaze behavior. However, there are many other data channels available of which our approach also handles. One of the difficulties for this detailed data-set is there are many channels of information (approximately 37). Viewing the relevant information across all the streams within these channels is a challenge of wading through visual clutter. For this example (and subsequent ones) we use 22 channels consisting of F-formations, gaze fixations, nodding gestures, and gesture phrases. F-Formations, or focus formations, were first identified by Adam Kendon to be units of topical cohesion marked by gaze cohesion of participants to common
objects or spaces [15]. Gesture phrases are atomic gestural motions marking a single motion trajectory that typically coincide with atomic speech phrases [22]. The data streams used numbered 6,500+ semi-intervals.

Table 1: Results for start of AFIT gaze model

<table>
<thead>
<tr>
<th>Context</th>
<th>Semi-intervals</th>
<th>Actor/Source</th>
<th>Probability</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>After</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>C gaze to papers</td>
<td>C gaze to E</td>
<td>0.05</td>
<td>0.75</td>
</tr>
<tr>
<td>2</td>
<td>C gaze to E</td>
<td>C gaze to F</td>
<td>0.06</td>
<td>0.85</td>
</tr>
<tr>
<td>3</td>
<td>C gaze to F</td>
<td>E gaze to E</td>
<td>0.07</td>
<td>0.95</td>
</tr>
<tr>
<td>4</td>
<td>C gaze to D</td>
<td>E gaze to papers</td>
<td>0.08</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>F gaze to C</td>
<td>E gaze to papers</td>
<td>0.09</td>
<td>1.00</td>
</tr>
</tbody>
</table>

The model for this example details the order of C’s gaze events to G, then D, and then E where each event has a start and end semi-interval. For our demonstration, we use the same situated instance from the original analysis. Figure 5 illustrates the connection between the annotation data and the results of our approach. In (A) is a visualization of the event data in the multimodal analysis tool MacVisSTA [36] (used for the original analysis). (B) graphically shows the model constructed using semi-intervals: [C gaze to G:s < C gaze to G:e = C gaze to D:s < C gaze to D:e = C gaze to E:s < C gaze to E:e], where ‘<’ or ‘=’ represents the start or end semi-interval, respectively. The first event, C’s gaze to G, is highlighted showing three sets of semi-intervals seen amongst the 22 data streams, allowing a quick view of these events with respect to the instance. These reported events are indeed events seen holding relation to the instance used (context 1) for the original analysis. With this information, an analyst can judge if the instance is deemed interesting enough to jump to the video and view the instance further. Also, identifying semi-intervals that adhere to equality (‘=’) seen in the model is not strictly upheld to allow for more flexibility in instance identification.

Multiple Context Cross-modal Relations: Our second example is motivated by judging the relevance of an instance. Here we show the viewing of and switching between results for multiple contexts. There are a total of 21 instances of C’s gaze sequence model. To exemplify the varying contexts, we list results from five of these in Table 1. By switching between contexts, it can be seen that the related events vary providing interesting variations of context. Each context can be used to query the video content and allow a more detailed look. Switching between the contexts can show the differing events that occur and may aid the analyst in judging the relevance between multiple instances. Some results contain more than one related semi-interval, e.g. context 2, Before. This shows when either multiple events started and/or ended at the same time marking a transition between events. For the case of context 2, Before, C’s gaze to the papers ends then C’s gaze changes to E.

Aggregated Results: Our third and last example takes the results from all contexts and aggregates them into one view to allow statistical comparison among the contexts and identification of variations. Here, instead of viewing one set of results at a time (e.g. one row of Table 1), we can view all sets (e.g. all rows) together. This combines the results into one view where the likelihood of each related event is presented as a comparison metric allowing a statistical comparison view of all related events across all contexts. One powerful aspect of this approach is that results from both complete and partial matches to the model are presented together. This allows the analyst to view instances that have some commonality to that of the model of interest and may prove informative in the analysis as illustrated in Figure 4 and related discussion.

An example of this aggregation can be seen in Figure 6 where the presented results correlate to a sub-set of the before results for C gaze to G:s. The first column (left) represents the actor/source of an event with the percentage fill of the representative graphical rectangle being the to-
tal number of semi-intervals associated with the respective actor/source. When one actor/source is selected, a second column is displayed which illustrates all the semi-intervals related to the actor/source that occur before C’s gaze to Gs. This second column displays each semi-interval’s description and associated conditional probability and confidence. The probability is mapped to the percentage fill of the graphical rectangle where more fill equals a greater probability. A bold bar is displayed on either end of the graphical rectangle depicting whether the semi-interval is a start (left side) or end (right side).

Given this aggregated view of related events, we can identify variations of related events to the model. A variation is described as a sequence of semi-intervals not recorded in the original analysis or that differs from the original model of interest. The reasoning for these variations was exemplified in Figure 4 and related discussion. In this example, three variations were identified that were of interest. The first stemmed from the question of if G returns C’s gaze at some point. This is a valid point of interest as such gaze exchange is an important indicator for coalition building, as mentioned earlier. Upon inspection of related events in the aggregate view of C’s gaze sequence detailed earlier, we see a gaze return from G during C’s gaze to G and ending while C’s gaze was on D. The resulting model of this variation (variation in bold) is: [C gaze to G:s < G gaze to C:s < C gaze to G:e = C gaze to D:s < G gaze to C:e < C gaze to D:e = C gaze to Es < D gaze to E:e]. Using this modified model, we performed another query of the data-set and found two instances of this model. At this point the analyst can view the related events of those instances and jump to their situated context in the video data.

Next, we looked to see if D returned C’s gaze. Similar to the prior case, D’s return gaze begins during C’s gaze to D and ends afterwards. This is in contrast to the original analysis where D’s gaze was already fixated on C prior to the start of C’s gaze sequence. Using this new model, we performed a query and found partial matches to the variation. No instances of this model variation were seen in the data-set. The only “instance” seen was matches found through identifying sub-structures of the model within the data (as seen in Figure 4). The resulting matches were very close to that of the desired variation leading to the identification of a variation not originally conceived in the data. This variation represents a new structure of how a coalition can be formed a query and found partial matches to the variation. The reasoning for these variations was exemplified in Figure 4 and related discussion. In this example, three variations were identified that were of interest. The first stemmed from the question of if G returns C’s gaze at some point. This is a valid point of interest as such gaze exchange is an important indicator for coalition building, as mentioned earlier. Upon inspection of related events in the aggregate view of C’s gaze sequence detailed earlier, we see a gaze return from G during C’s gaze to G and ending while C’s gaze was on D. The resulting model of this variation (variation in bold) is: [C gaze to G:s < G gaze to C:s < C gaze to G:e = C gaze to D:s < G gaze to C:e < C gaze to D:e = C gaze to Es < D gaze to E:e]. Using this modified model, we performed another query of the data-set and found two instances of this model. At this point the analyst can view the related events of those instances and jump to their situated context in the video data.

Next, we looked to see if D returned C’s gaze. Similar to the prior case, D’s return gaze begins during C’s gaze to D and ends afterwards. This is in contrast to the original analysis where D’s gaze was already fixated on C prior to the start of C’s gaze sequence. Using this new model, we performed a query and found partial matches to the variation. No instances of this model variation were seen in the data-set. The only “instance” seen was matches found through identifying sub-structures of the model within the data (as seen in Figure 4). The resulting matches were very close to that of the desired variation leading to the identification of a variation not originally conceived in the data. This variation represents a new structure of how a coalition can be constructed with which such knowledge can be applied to other data-sets (e.g. the other AFIT sessions).

The detection of a coalition can be strengthened through using other data streams of interaction such as acknowledge-ment gestures in response to gaze. Hence, we identified another variation where G responds to C’s gaze with a gesture nod (absent of G’s gaze fixation). However, no instances of this model variation were seen in the data-set. The only “instance” seen was found through identifying sub-structures (same as the previous variation). Hence, this variation represents another model structure of a potential coalition that can be used in searching other data-sets.

7.4 Discussion

Given a set of events describing multimodal behavior of multiple individuals, we were able to identify events that were potentially relevant to particular instances of a phenomenon model. In doing so, we support expert exploration of large corpora of multimodal behavior to understand and extend behavioral models. Our approach is situated in that the expert works with particular instances of the behavior in the data and is able to inspect the original video/audio record of the behavior. However, our approach transcends the confines of the current instance in that it reveals other occurrences of the behavior being studied in the data. It allows the user to explore these occurrences within their contexts to determine if recurring event patterns are indeed relevant to the behavior being studied. Hence, this situated analysis also allows the comparison of phenomenon instances across different contexts.

The aggregated view of results adds an extra level to situated analysis as it provides a means to view all the results to all instances at once. This allowed identification of variations not originally conceived, providing opportunities for new thought constructions of models to be used in future analyses.

8. CONCLUSION AND FUTURE WORK

We were able to facilitate the situated analysis of multimodal corpus of human behavior. The expert is able to view related events of different instances of a phenomenon, view the cross-stream relations, compare across instances, and be assisted in the discovery of variations of phenomenon. We were able to illustrate how our approach is a beneficial aid in analysis through application to a data-set with known ground truth.

Through this investigation of situated analysis, we see areas where our processing can be improved to provide advantageous benefits to the analyst. The first is expanding the processing of our results to include more than boundary relations to events. Currently, our processing reports related semi-intervals with respect to the boundary (start/end) of events. Other equally important information is events occurring during a model but their start or end semi-interval is not temporally close. These events are on-going during instance(s) of the model. We are also interested in allowing more flexibility in identifying instances of a model with respect to how the order of model semi-intervals are realized in the data-set. For example, the hypothesis representation in Figure 1B may not surface in the data as all the gaze semi-intervals occurring at the exact same time, but very close to each other. This exemplifies the idea of temporal constraints, an area of research for this kind of processing and analysis discussed in [25]. Also, semi-intervals occurring in a slightly different order could pose a beneficial instance match, hence flexibility for this would be beneficial.

9. REFERENCES