MIVA: Multimodal Interactions for Facilitating Visual Analysis with Multiple Coordinated Views

Imran Chowdhury, Abdul Moeid
Chittagong University of Engineering & Technology
Chittagong, Bangladesh
{u1704107, u1704101}@student.cuet.ac.bd

Enamul Hoque
York University
Toronto, Canada
enamulh@yorku.ca

Muhammad Ashad Kabir
Charles Sturt University
NSW, Australia
akabir@csu.edu.au

Md. Sabir Hossain
Chittagong University of Engineering & Technology
Chittagong, Bangladesh
sabir.cse@cuet.ac.bd

Mohammad Mainul Islam
Verizon Media
Sunnyvale, CA, USA
mohammad.islam@verizonmedia.com

2020 24th International Conference Information Visualisation (IV) | 978-1-7281-9134-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/IV51561.2020.00124

***Abstract*—Typically, people perform visual data analysis using mouse and touch interactions. While such interactions are often easy to use, they can be inadequate for users to express complex information and may require many steps to complete a task. Recently natural language interaction has emerged as a promising technique for supporting exploration with visualization, as the user can express a complex analytical question more easily. In this paper, we investigate how to synergistically combine language and mouse-based direct manipulations so that weakness of one modality can be complemented by the other. To this end, we have developed a novel system, named Multimodal Interactions System for Visual Analysis (MIVA), that allows user to provide input using both natural language (e.g., through speech) and direct manipulation (e.g., through mouse or touch) and presents the answer accordingly. To answer the current question in the context of past interactions, the system incorporates previous utterances and direct manipulations made by the user within a finite-state model. We tested the applicability of *MIVA* on several dashboards including a COVID-19 dashboard that visualizes coronavirus cases around the globe. Our demonstration provides initial indication that the *MIVA* system enhances the flow of visual analysis by enabling fluid, iterative exploration and refinement of data in a dashboard with multiple-coordinated views.**

***Index Terms*—Multimodal interaction, multiple-coordinated views, natural language interface, direct manipulation**

I. Introduction

Traditionally people interact with data visualization using mouse and/or touch-based interaction techniques. Often the user points to different objects of interests by direct manipulations (e.g., making lasso selection in a map chart using a mouse). While such interaction techniques are effective, they can be inefficient when the information need is ambiguous, complex or require many steps to complete a task [15].

Natural language (NL) interaction has emerged as a promising technique for performing analytical tasks with visualizations [7]. Using natural language, a user can express a complex question more easily through text or speech. Accordingly, the system can respond to the question by showing the answer within existing visualizations or when necessary by creating a new visualization. Natural language interaction can offer numerous advantages such as ease of use, convenience, and accessibility to novice users, facilitating the flow of analysis for novices and experts alike [15].

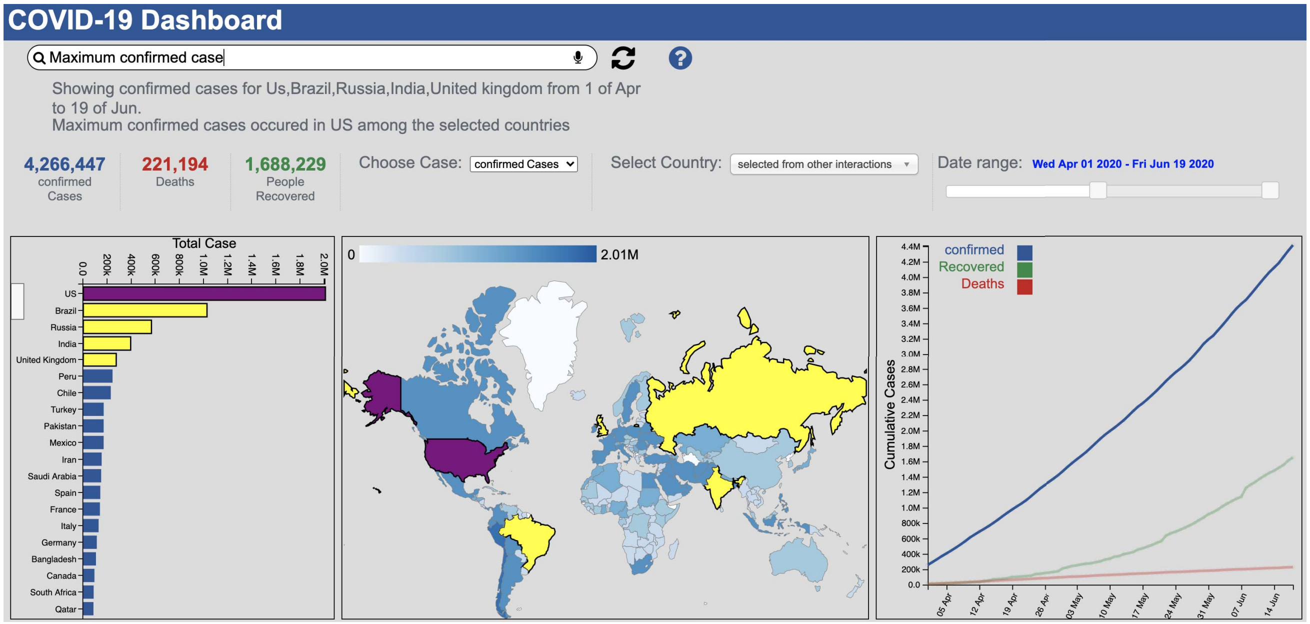
While direct manipulation (using mouse and/or touch-based interactions) and natural language interaction techniques have their own advantages, we argue that we can amplify their utilities if we synergistically combine the two techniques. In this way, both interaction modalities can complement each other. In fact, a previous study found that when speech and pen input are combined people make significantly fewer mistakes compared to using either speech or pen individually [13]. More recent system Orko also confirms the utility of multimodal interactions for visual data analysis [17]. However, multimodal interaction techniques for data visualizations are still in their nascency. Existing works have demonstrated the capabilities of multimodal interactions in a limited context (e.g., focusing on a single visualization only).

To address the above mentioned limitations, in this paper, we explore how to combine speech and direct manipulation as complementary modalities for visually analyzing data presented in multiple coordinated views. Our propose system, named Multimodal Interactions System for Visual Analysis (MIVA), allows the users to type queries as well as to make mouse-based selections in various views within a dashboard. The underlying mechanism of our system uses a frame-based dialog management tool that detects the user's intent (e.g., 'find maximum value') and slots (e.g., 'among European countries') and manages contexts across multiple input modalities.

Figure 1 illustrates how our system helps the user to perform multimodal interactions within a dashboard. Here, the dashboard is showing the coronavirus cases around the world based on the Johns Hopkins dataset[1]. The user decided to choose some countries and a time-range through lasso selection followed by typing the query "Find maximum cases". In response, the system analyses all the selections and the query to generate the answer. The primary contributions of our work are two-fold:

[1] https://github.com/CSSEGISandData/COVID-19

DOI 10.1109/IV51561.2020.00124



COVID-19 Dashboard

Maximum confirmed case

Showing confirmed cases for Us,Brazil,Russia,India,United kingdom from 1 of Apr to 19 of Jun.
Maximum confirmed cases occured in US among the selected countries

4,266,447 confirmed Cases | 221,194 Deaths | 1,688,229 People Recovered

Choose Case: confirmed Cases

Select Country: selected from other interactions

Date range: Wed Apr 01 2020 - Fri Jun 19 2020

Fig. 1. A screenshot of the *MIVA* with a dataset of coronavirus cases around the world. The dashboard consists of three coordinated views: A bar chart showing countries with top cases, a map chart showing the frequency of cases using color saturation and a line chart showing how cumulative cases evolved over time. Here, the user types the query "find maximum confirmed cases" and then selects a date range from the slider widget and some countries from the bar chart. In response, the system displays the results accordingly.

1) A frame-based dialog approach that detects the user's intent and slots from multimodal inputs (speech and direct manipulations).
2) We have developed *MIVA*, that supports the multimodal interactions with dashboards having several views. Our system allows selection within multiple views of the dashboard and then uses the frame-based dialog approach and data analysis techniques to generate the answer. It also employs a finite-state model to capture the context – both the past utterances and the direct manipulations in multiple views in a dashboard – to fulfill the current information needs of the user.

II. Related Work

A. Natural Language Interactions with Visualizations

Natural language interfaces for data visualization have received considerable attention recently. Typically, these interfaces respond to user queries by creating a new visualization (e.g., DataTone [5]) and/or by highlighting answers within an existing visualization (e.g., Eviza [16]). Some systems enable follow-up data queries from users with limited support for pragmatics [2], [3], [19]. Commercial systems such as AskData [1], [7] have also been developed for similar purpose. Generally, Such systems understood the importance of providing feedback on how the system interprets queries and allowing users to rectify misunderstandings. However, all these works largely depend on heuristics or grammar-based parsing techniques which are incapable of handling questions that are compositional or otherwise incomplete and ambiguous.

There has been a recent surge in research on conversational AI [4], which focuses on three broad areas namely automatic question answering [4], [8], [10], task-oriented dialog systems (e.g., Siri) and social bots (e.g., Xiaoice [20]). While still in early stage, this emerging research shows promising results thanks to a large amount of training data and the breakthrough in deep learning and reinforcement learning. Methods for question answering with semi-structured tables [14] demonstrate how highly compositional queries can be translated into logical forms to generate answers using neural network models [6], [10]. Despite their promise, the current level of accuracy of these techniques is inadequate. More importantly, they are not designed to deal with questions that are unique to visualizations, for example, questions that refer to visualization properties (e.g., colors, marks, axes).

B. Multimodal Interactions with Visualizations

Prior work in the HCI community suggests that multimodal interfaces can significantly enhance user experience and usability of a system [13]. However, research on interaction techniques with visualizations has mostly explored a single input modality such as mouse, touch, or more recently, natural language. Evizeon [7] investigates the potential benefits of multimodal interactions for visualizations by combining language and mouse-based direct manipulations. However, the direct-manipulation based interaction of Evizeon was limited to a map chart only. Orko [18] demonstrates the value of combining mouse and touch with natural language interactions for network data visualization. Again the multimodal interaction was limited to a network diagram only.

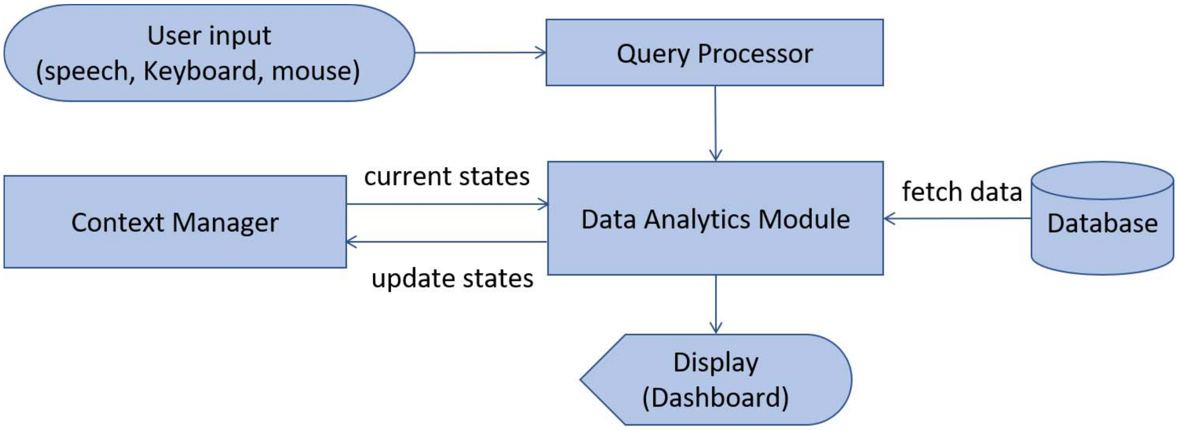


Fig. 2. The *MIVA* system architecture

While these works suggest that multimodal interaction offers strong promise, the potential challenges and opportunities in integrating different modalities have not been deeply investigated yet [11]. In this paper, we particularly focus on promoting and enabling multimodal interactions for facilitating visual analysis with multiple-coordinated views.

III. MIVA System

The *MIVA*[2] system is designed as a web-based tool that runs on a browser at the client side. Figure 2 illustrates the architecture of the system. Below we describe how the major components of our system work and how they communicate with each other:

A. *Query Processor*

Given a textual query, the system first tokenizes the text to find the list of words in the query [12] followed by performing stemming and stop word removal in that list. Note that if the user provides the input using speech, the system uses the Webkit[3] speech recognition library to convert the input into text.

In a goal oriented dialog system, a set of actions that the system supports is defined as the intent set. In the context of the COVID-19 dashboard shown in Figure 1 the example set of intents could be $I = \{$‘Find maximum’, ‘Show’, ‘Sort’, ...$\}$. We build this list of intents based on the low-level visualization tasks identified by Amar et al. [9] as well as by analyzing common queries asked by people while exploring visualizations [15].

Usually there are one or more data variables that are involved with each intent, which are called slots. For example, given the query ‘Find maximum confirmed cases’, the intent is ‘find maximum’ while the slot is ‘confirmed cases’. Similarly, given the query ‘Show cases in US and China between March and April’, the intent is ‘Show’, while the slots are ‘cases in US and China’ and ‘between March and April’.

In order to detect slots, the system extracts the list of column names and the corresponding cell values from the data tables in the database to build a dynamic lexicon. For example, for the dataset visualized in Figure 1, the column names include ‘date’, ‘country’, ‘case type’ and so on. The cell values like ‘China’ and ‘US’ belongs to the ‘country’ column. Given the query ‘Show cases in US and China’, the system searches through the dynamic lexicon to find that the slots involved here are ‘US’ and ‘China’ which are two cell values in the country column. So the system detects the slots as $S = \{$‘Country == US’, ‘Country == China’$\}$.

Detecting slots from direct manipulation based interactions is rather straightforward. For each chart, the system maintains the visual encoding information describing how the data attribute is mapped to visual attributes (e.g., area, height, fill-color) of marks (e.g., bar, circle). As such, whenever the user makes a selection by clicking, dragging, or lasso selection on graphical marks (e.g., selecting some bars or a line segment), the system automatically identifies the corresponding data attribute and the values. For example, if the user selects a bar from the bar chart in Figure 1 that represents ‘US’, the system identifies that ‘US’ is cell value for the data attribute ‘Country’ therefore, $S = \{$‘Country == US’$\}$.

B. *Data analytics module*

The data analytics module takes the slots and the intentions as an input and executes corresponding analytical functions on the rows in the data table to generate the answer. For example, for the query in Figure 1, the system applies filtering functions based on the selected countries and date range and then sort the countries by cases to find the country with the maximum cases. The result is then presented in the visual interface as a response to the query.

C. *Context manager*

The context manager communicates with the data analytics module to ensure that the system executes and generates the

[2] A video demo of the *MIVA* system is available at: https://www.yorku.ca/enamulh/video/MIVA.mp4

[3] https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API

result based on the past interactions made by the user. One of the design challenges is that the system needs to consider all past interactions generated from multiple input modalities. For the convenience of integrating all modalities, we represent all slots detected in textual format. After every interaction made by the user, the system updates the intentions and slots which are represented in the context manager as finite state models. The intentions and slots are represented as a finite state machine with three transition functions: addition, deletion and change operations.

IV. Conclusion and Future Work

We present *MIVA*, a system that synergistically integrates multimodal interactions for facilitating visual data analysis of multiple coordinated views in a dashboard. For this purpose, the system follows an architecture that takes various forms of input (speech, mouse, keyboard, etc.) from users and then answers the query in the context of all past interactions. Based on the initial evidence, we argue that combining different interaction modalities is a very promising research direction in achieving natural and fluid exploration, and refinement of data in complex dashboards.

In the future, we would like to incorporate more input modalities like hand gestures and display factors (e.g., large display and virtual reality) to see how such new modalities and form factors can emerge to a new kind of human-interaction paradigm for exploring information visualization. We will also enhance our natural language processing methods for detecting intentions and slots to capture the complexity, ambiguity and nuances of human language using recent deep neural network based sequence-to-sequence model. Finally, we will further evaluate our system using a lab-based summative evaluation as well as more long-term case studies to understand the potential utilities and trade-offs of introducing multimodal interactions for dashboards with multiple coordinated views.

References

[1] Tableau's Ask Data. https://www.tableau.com/products/new-features/ask-data, 2020.

[2] K. Cox, R. E. Grinter, S. L. Hibino, L. J. Jagadeesan, and D. Mantilla. A multi-modal natural language interface to an information visualization environment. *International Journal of Speech Technology*, 4(3-4):297–314, 2001.

[3] K. Dhamdhere, K. S. McCurley, R. Nahmias, M. Sundararajan, and Q. Yan. Analyza: Exploring data with conversation. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 493–504. ACM, 2017.

[4] J. Gao, M. Galley, and L. Li. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374. ACM, 2018.

[5] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 489–500. ACM, 2015.

[6] T. Haug, O. Ganea, and P. Grnarova. Neural multi-step reasoning for question answering on semi-structured tables. *CoRR*, abs/1702.06589, 2017.

[7] E. Hoque, V. Setlur, M. Tory, and I. Dykeman. Applying pragmatics principles for interaction with visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):309–318, 2018.

[8] K. Kafle, C. Scott, P. Brian, and C. Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5648–5656, 2018.

[9] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Visual analytics: Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer, 2008.

[10] J. Krishnamurthy, P. Dasigi, and M. Gardner. Neural semantic parsing with type constraints for semi-structured tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1526, 2017.

[11] B. Lee, A. Srinivasan, J. Stasko, M. Tory, and V. Setlur. Multimodal interaction for data visualization. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, AVI '18, pages 11:1–11:3, 2018.

[12] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[13] S. Oviatt. Multimodal interactive maps: Designing for human performance. *Human-computer interaction*, 12(1):93–129, 1997.

[14] P. Pasupat and P. Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1470–1480, 2015.

[15] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST 2016, pages 365–377, New York, NY, USA, 2016. ACM.

[16] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 365–377. ACM, 2016.

[17] A. Srinivasan and J. Stasko. Natural language interfaces for data analysis with visualization: Considering what has and could be asked. In *Proceedings of EuroVis*, volume 17, pages 55–59, 2017.

[18] A. Srinivasan and J. Stasko. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE transactions on visualization and computer graphics*, 24(1):511–521, 2018.

[19] Y. Sun, J. Leigh, A. Johnson, and S. Lee. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In *International Symposium on Smart Graphics*, pages 184–195. Springer, 2010.

[20] Xiaoice. Microsoft xiaoice bot, 2018.