Multimodal Analysis of Video Collections: Visual Exploration of Presentation Techniques in TED Talks

Aoyu Wu and Huamin Qu, *Member, IEEE*

Abstract—While much research in the educational field has revealed many presentation techniques, they often overlap and are even occasionally contradictory. Exploring presentation techniques used in TED Talks could provide evidence for a practical guideline. This study aims to explore the verbal and non-verbal presentation techniques from a collection of TED Talks. However, such analysis is challenging due to the difficulties of analyzing multimodal video collections consisted of frame images, text, and metadata. This paper proposes a visual analytic system to analyze multimodal content in video collections. The system features three views at different levels: the Projection view with novel glyphs to facilitate cluster analysis regarding presentation styles; the Comparison View to present temporal distribution and concurrences of presentation techniques and support intra-cluster analysis; and the Video View to enable contextualized exploration of a video. We conduct a case study with language education experts and university students to provide anecdotal evidence about the effectiveness of our approach, and report new findings about presentation techniques in TED Talks. Quantitative feedback from a user study confirms the usefulness of our visual system for multimodal analysis of video collections.

Index Terms—Visual Analytics, Multimedia Visualization, Multimodal Analysis

1 Introduction

While it is usually easy to judge a presentation, it is far more difficult to explain what makes a presentation excellent. Regarding this, much research in the educational field has suggested many presentation techniques. However, they often overlap and are even occasionally contradictory, thus puzzling learners. For example, presentation expert Currie [1] criticizes incessant hand movements, while Khoury [2] asserts that "*Thou Shalt Not Leave Hands Idle*". There has been little quantitative analysis reported on the actual usage of presentation techniques in a collection of good presentations, which could help gain empirical insights into effective presentation delivery.

Computer-based solutions to analyze presentation techniques have being receiving considerable attention since the early 2010s when presentations were unprecedentedly finding outlets online. In this vein, one of the most impacting disseminators is the TED (Technology, Entertainment, Design) conference and its associated website (TED Talks), where scholars disseminate information to the lay public through a condensed presentation often within 18 minutes [**?**]. TED Talks are released under a Creative Commons BY-NC-ND license for unrestricted use. To provide insights into the functional aspects of presentation techniques, recent research has sought to explore characteristics of TED Talks by computer-based solutions on a large scale. Prior research has analyzed speaker demographics [3], [4], metadata [5], [6], user comments [7], prosody [8], and simple linguistic characteristics [9].

Nevertheless, information about presentation styles and non-verbal communication, while advocated as cornerstones of successful presentations by most domain experts, remains largely absent from such large-scale automatic analysis. There is also a paucity of research on the interplay between verbal and non-verbal presentation techniques. In addition, existing research is constricted by limited data mining techniques such as statistical hypothesis testing and regression, therefore failing to reveal hidden patterns. For instance, clustering analysis is desirable to alleviate the effects of individual differences, helping discern techniques among different presentation styles. To this end, we make a multimodal and integrated investigation into presentation techniques from many TED Talks. However, the term presentation techniques is used in a rather loose sense, making a comprehensive probe infeasible. We hence collaborate with three university language education professionals and identify three paramount aspects of presentation techniques: rhetorical modes, body postures, and gestures. These three aspects and their interplay form the focus of our analysis.

Visualization techniques can be of significant assistance in understanding those aspects by exploring video collections. The interest in video visualization has grown rapidly over a wide spectrum including sports analysis [10], [11], entertainment [12], [13], traffic and area surveillance [14], medical endoscopy [15], and to a lesser extent, educational videos [16]. In those systems, visual analytic tasks are placed on a single modality such as visual content, text annotation or metadata. For instance, visualization systems for traffic surveillance are typically only concerned with the visual mode (the videos). Nevertheless, our study considers multimodal content in videos including frame images, texts and metadata simultaneously. Moreover, analyzing video collections popes a challenge on visualization approaches [17]. To conclude, our work is challenging due to a variety of factors, including the integration of relevant techniques to extract and process multimedia data, and the complexity of visual analytics approaches for multimodal content in video collections.

In this study, we present a novel visualization system to support interactive and insightful analysis into presentation techniques in many TED Talks (Fig.1). Through an iterative design

- *A. Wu and H. Qu are with the Hong Kong University of Science and Technology, Hong Kong, China. E-mail: awuac, huamin@ust.hk*

Manuscript received XX XX, 2018; revised XX XX, 2018.

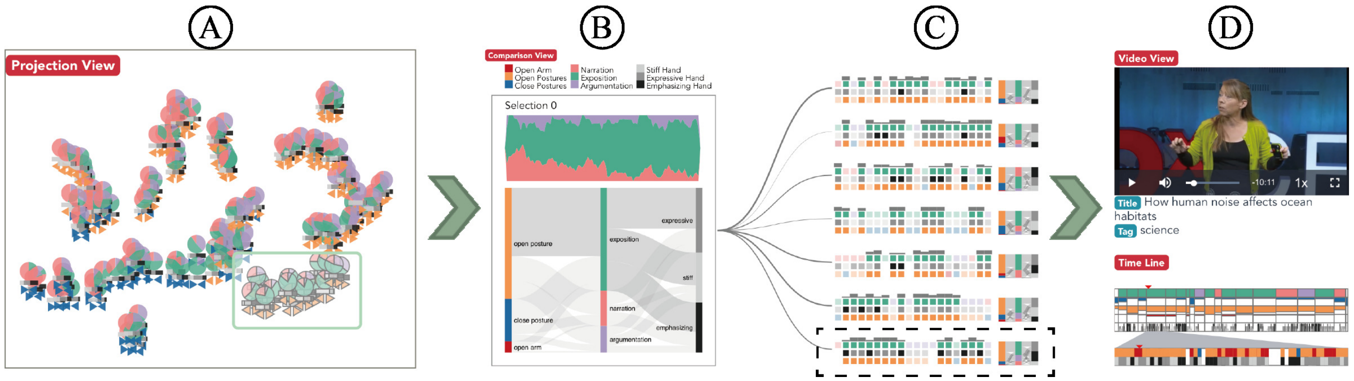


Fig. 1. Our approach features an interactive visualization for multimodal analysis of presentation techniques in TED Talks. This figure illustrates the analysis pipeline at different levels of detail: (A) The Projection View helps to inspect clusters of presentation styles and select those of interest; (B) The Comparison View responsively summarizes the temporal usage of verbal presentation techniques and their interplay with non-verbal techniques; (C) Meanwhile, the Presentation Fingerprinting provides an at-a-glance overview and highlights major patterns to facilitate intra-cluster comparison. It also serves as a navigation tool for case investigation; (D) Navigated by an elastic timeline, the Video View enables effective, contextualized exploration of a TED Talk and fosters user trust regarding data.

study with language education experts, we first elicit domain-specific analytical goals and apply appropriate computer vision and natural language processing methods to capture presentation techniques. We further formulate the visualization tasks for analyzing multimodal content in video collections. We then perform a case study with domain experts and general users on 146 TED Talks. The findings conform to and further supplement the existing theories in the education domain, providing anecdotal evidence about the effectiveness of our system. Finally, we conduct a user study to validate the usefulness of our visualization system for analyzing multimodal content in video collections. In summary, the contributions of our work are:

- A visualization system which integrates well-established techniques and several novel designs to analyze multimodal content in video collections.
- A study of the temporal distribution of presentation techniques as well as their interplay used in TED Talks.
- A novel glyph design to show multiple attributes of presentation techniques for easy identification of speakers' features.
- A case study which reports the gained insights, and a user study which demonstrates the effectiveness of the visualization design.

2 Literature Review

Our work is closely related to analysis and browsing interfaces of presentation videos, as well as video visualization.

2.1 Analysis of Presentation Techniques

In recent years, some research has analyzed presentation techniques by computer-based approaches. For this discussion, we classify related work into verbal and non-verbal aspects.

For verbal techniques, Tsai [8] compared the prosody characteristics of TED speakers and university professors, revealing discriminative features such as deeper voices. Kravvaris and Kermanidis [9] studied simple language characteristics in the more popular TED videos and those less popular ones. Their analysis shows that popular TED talks feature faster pace and higher sentence complexity. Recently, Tanveer et al. [18] analyzed the narrative trajectories in over 2000 TED Talks. They identified the impact of narrative trajectories on the subjective ratings of the audience. Nevertheless, little work has delved into high-level semantic information, for example, whether the speaker is expressing ideas, giving examples or telling stories, though they are advocated as cornerstones of successful presentations in much presentation literature [19]. Inspired by this, we seek to explore the rhetorical modes in TED talks. Specifically, our work follows the state-of-the-art method [20] to characterize snippets of presentation scripts into three primary modes suggested by Rozakis [21] (*i.e.*, narration, exposition and argumentation).

Closely related to our work, several systems have analyzed non-verbal techniques such as body gestures and examined their relationship with semantic significance. A prior system [22] automatically identifies instructor's gestures during lectures whereby user studies are conducted to verify the hypothesis that gestures yield significant pedagogy. Okada and Otsuka [23] proposed a framework to associate spoken words with hand motion data observed from optical devices. Our work is different from those on three aspects. We adopt the state-of-the-art system OpenPose [24] to detect human body key-points, which significantly improves the accuracy. We involve domain knowledge to characterize postures and gestures into the known taxonomy [25], while existing approaches only calculate numerical hand motion. More importantly, we use visual analysis to bring in human expertise in analyzing multimedia data, rather than pure statistical analysis.

2.2 Interfaces for Browsing Single Presentation Video

Several visual interfaces have been proposed to facilitate searching and browsing within a single presentation video. TalkMiner [16] is an early system for retrieving keywords from slides and metadata within lecture video archives. To infer higher-level semantics, MMToc [26] automatically performs topic segmentation and creates a table of content based on word salience from multimodal cues. Pavel et al. [27] presented tools to create video digests to afford browsing by segmenting videos and summarizing those segments. These systems focus on video segmentation and supporting browsing at an aggregate level. In our study, we adopt the predefined video segments labeled by TED to better represent the video semantics, and extend browsing support to a finer level.

Fewer interfaces have been proposed to navigate presentation videos at a second-by-second level. Haubold and Kender [28] proposed a visual interface for segmented multimodal content with

TABLE 1
Research Scope of Presentation Techniques.

Presentation Technique	Subcategory	Description
Rhetorical Mode	Narration	To tell a story or narrate an event or series of events
	Exposition	To explain, inform, or describe.
	Argumentation	To prove an idea, or point of view, by presenting reasoning.
Body Posture	Close Posture	Includes crossed arms and wringing hands, which are considered to communicate defensiveness.
	Open Palm	Convey openness at the risk of being overenthusiastic and offensive.
	Open Posture	Refers to keeping hands within the "strike zone" between being closed and having open palms.
Body Gesture	Stiff	Refers to no or slight hand movement.
	Expressive	Refers to hand movement no more than the length of the forearm.
	Jazz	Refers to exaggerated hand movements.

four parallel timeline graphs. Their interface does not contain a video player or a text viewer due to the limited screen space, which could degrade users' experiences. Kim et al. [29] presented novel interaction techniques such as dynamic timelines and interactive transcripts to optimize navigation. While they improve user performance and experience in browsing presentations, they could not be extended to multivariate finer-grained data. Our work addresses these issues with an elastic timeline which reduces the screen space. Meanwhile, our interface includes and vertically aligns the video browser, the elastic timeline, and the text viewer to encourage contextualized exploration.

2.3 Visualization Support in Multimedia Analysis

Information visualization has been a valuable tool for multimedia analysis, easing the investigation of multimodal content for various tasks. We focus on the most relevant work that addresses multimodality and scalability.

Multimodality. Much work has looked at ways to visually encode multimodal content, and we classify related work into computer vision based and abstract approaches. Computer vision based systems [10], [15] typically reconstruct scenes or extract trajectories from videos, to which data from other modes are visually mapped. While they enable a tightly integrated visual exploration within the video context, they fail to provide a temporal overview. Regarding this, abstract techniques [30], [31] usually aggregate multimodal content into a single-variate feature and visualize its temporal dynamics by line-based charts. For instance, Story Explorer [31] uses a story curve to visualize the temporal evolution of movie narrative order derived from scripts and metadata. To visualize multivariate features, Kurzhals et al. [13] used matrix-based visualization on multiple hierarchical levels for investigating various descriptive features in movie scenes and dialogues. We adopt a similar design and integrate such abstract visualization into a browsing environment in a more screen-space-effective and responsive manner.

Scalability. In addition to visualizing the multimodal content of a single video, visualizing video collections poses a challenge [17]. Recent research in the multimedia field has proposed a number of interfaces for analyzing video collections. These systems, as proposed in the Video Browser Showdown competition [32], are mainly designed for retrieval and navigation tasks in large databases and typically consist of: (1) a video player; (2) a control panel; and (3) a result view. However, they do not include sufficient visualization for exploratory or analytical tasks. Regarding this, some systems [12], [17], [33] augment the result view with visualization such as timelines, grids, or plan texts for each video. Such visual summary of a single video cannot direct analytics to identify patterns at an aggregated level. Our work embeds those components into an analytic framework to provide visual access to aggregate results at different levels of details. Specifically, we introduce the Presentation Fingerprinting as a hierarchical visual abstraction and add a novel glyph to support cluster analysis, inter- and intra-cluster comparison.

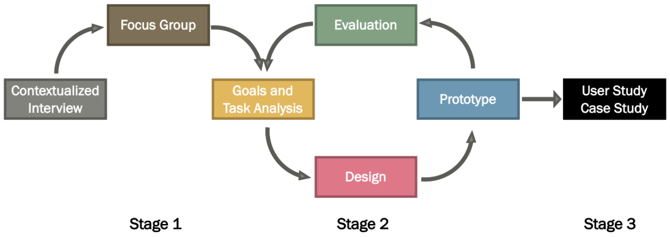


Fig. 2. The design process consists of three stages: a preliminary stage with contextual interviews and focus group study to contextualize design, an iteration stage with three rounds of design-prototype-evaluation refinement, and a final evaluation stage including two studies.

Perhaps most related to our work is the Video Lens [17] system, which is designed to find relevant sections (*event*) by metadata and play them back in rapid succession. However, our system is different as we treat videos as a fundamental element instead of an *event*, which offers opportunities for comparing videos and performing clustering analysis. In addition, we extend the modalities from metadata to text and visual mode, and support video browsing at both the section and second-by-second levels.

3 User-Centered Design

Our primary goal was to investigate the usage of presentation techniques from excellent presentation videos. We adopted a user-centered design methodology which was divided into three stages, as depicted in Fig. 2. In this section, we first describe how experts identified and updated research questions through the preliminary stage and the iteration stage, and summarize the derived analytical goals, as well as the visualization tasks.

3.1 Design Process

We closely worked with three domain experts (E0-2) in university language education. They have been engaged in English teaching for an average of 13.7 years and accumulated rich experience in teaching presentation techniques, which account for roughly half of their teaching content. Currently, E0 is in charge of compiling

TABLE 2
Visualization Tasks with Category and Corresponding Analytical Goals.

Category	Visualization Task		Analytical Goals
Visual Mapping	T1	To present temporal proportion and distribution of data.	G1, G4
	T2	To find temporal concurrences among multimodal data.	G2, G4
Relation & Comparison	T3	To support cluster analysis and inter-cluster comparson.	G3, G5
	T4	To compare videos at intra-cluster level.	G3
Navigation	T5	To enable rapid video browsing guided by multiple cues.	G4, G5
	T6	To allow faceted search to identify examples and similar videos in video collections.	G5, G6
Overview + Detail	T7	To display data at different levels of detail and support user interaction.	G3, G5
Feature Specification	T8	To support selecting interesting data or feature space.	G3, G6
Automated Aggregation	T9	To algorithmically extract meaningful patterns and suppress irrelevant details.	G2, G3

textbooks and E1-2 are responsible for teaching language courses for undergraduate students.

Preliminary stage. Our user-centered design process started with individual *contextual interviews* with experts to understand current practices and domain problems. E0 reported that their research was mainly driven by case-based evidence rather than large-scale automatic analysis, which has often led to ongoing controversy. For example, while many experts suggest minimal body gestures to avoid distracting audiences, a recent study [34] suggests that popular presentations feature speakers with the most vigorous hand gestures. Regarding this, she highly valued a quantitative analysis approach for the actual usage of presentation techniques from TED Talks, which could offer empirical insights (G1). E1-2 appreciated the “example and Non-example Learning” strategy, and mentioned that they usually need to manually search domain literature or browse lengthy TED talks to find examples and non-examples. They therefore expected an interactive interface for searching and browsing presentation videos (G4, G5).

During the individual interviews, we noticed slight differences among the scope of presentation techniques suggested by each expert. We therefore conducted a *focus group* to try to reach consensus about the research scope. Before the focus group, we surveyed presentation techniques in the related literature, as well as corresponding state-of-the-art methods for extracting those techniques, which resulted into 14 candidates which were both mentioned in the domain literature and quantifiable by computer algorithms. The focus group was subsequently conducted to inquire experts’ opinions based on their professional views on the significance of each candidate, which was coded as *very*, *moderately*, or *slightly significant*. Meanwhile, we introduced the feasibility and accuracy of extraction algorithms, which were coded similarly. At last, they came to an agreement on important presentation techniques which were both *very* significant and feasible: rhetorical modes, body postures, and gestures. For each technique, together we enumerated subcategories according to the domain literature [21], [25], as described in Table 1. We then conducted *task analyses* to determine domain goals and visualization tasks.

Iteration stage. We then iteratively performed paper-based *designing* and code-based *prototyping* to display the summary of usage on presentation techniques at both the individual and collective level. During three rounds of *evaluation*, our experts showed excitement about the power of visualization, which in turn triggered new questions and design requirements. They posed problems on the interplay among verbal and non-verbal techniques, for the purpose of a systematic and integrated understanding (G2). Particularly, they exhibited curiosity about the semantic context for various body postures and gestures. Moreover, they would like to explore individual and group variations, so as to identify different presentation styles based on the usage of techniques (G3). They also proposed an extension of study factors such as presentation topics, allowing users to freely explore the data from a wider perspective (G6). As we discuss later in Section 5, we modified and improved our design in line with their feedback.

3.2 Analytical Goals

Experts’ analytical goals went broader along with a deepening understanding gained from prototyping. We hereby formulate the analytical goals according to the complexity level.

G1: To reveal the temporal distribution of each presentation technique. The temporal distribution captures the fundamental characteristics of a presentation, since presentations typically follow design structures. Understanding the temporal distribution and dynamics is therefore a vital prerequisite for assessing and classifying presentation techniques. For instance, how is narration temporally distributed among the presentations?

G2: To inspect the concurrences of verbal and non-verbal presentation techniques. While presentation techniques can be separately utilized in verbal or non-verbal forms, their simultaneous interplay is proven to offer a combined power. Our experts are particularly interested in hand gestures and rhetorical semantics, because hand gestures can communicate different emotional messages suitable for certain semantics. For example, the degree of hand movement might vary from narration to argumentation. Depicting their concurrences could provide intriguing insights into presentation techniques from a more comprehensive perspective.

G3: To identify presentation styles reflected by technique usage and compare the patterns. Despite much well-acknowledged guidance for presentation techniques, their actual usage might vary on a case-by-case or collective basis. A single aggregate summary would therefore be flimsy to draw conclusions, demonstrating the necessity of cluster and outlier analysis. Examining the variations among different individuals or groups helps comprehend presentation techniques in a systematic sense.

G4: To support guided navigation and rapid playback of video content. Although exploring the summarized features of presentation techniques enables effective visual analysis, it could be abstract and difficult to verify. Therefore, the experts suggest that it is essential to browse the original video and scripts in response within up to one second. Moreover, they expected additional information on top of textual content such as signposting

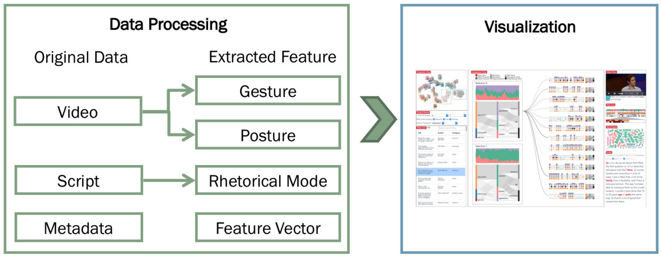


Fig. 3. Overview of the system architecture.

functionality, in order to obtain a deeper understanding of the verbal and non-verbal interplay.

G5: To facilitate searching in video collections. The experts wish to rapidly identify examples and non-examples of each presentation technique from a video collection. In addition, they want to search similar presentations based on the techniques used.

G6: To examine presentation techniques from different perspectives and provide faceted search. While the system will be capable of the aforementioned presentation techniques, our experts advocate that it should encourage data exploration from wider perspectives such as presentation tags. In addition, they suggest faceted search to allow free exploration.

3.3 Visualization Tasks

We have derived 9 visualization tasks from the analytical goals, as shown in Table 2. They are organized according to the categorization of visual analysis techniques for multimodal data suggested by Kehrer and Hauser [35]. T1-2 pertain to visual mapping, *i.e.*, how to represent the data, which is fundamental for probing into presentation techniques and support guided video browsing. Tasks T3 through T8 mainly focus on interaction, *i.e.*, how to link views for systematic analysis, leading to insights into presentation techniques at different scales. T9 relates to computational analysis, *i.e.*, what the main characteristics of data are, which could facilitate visual analysis by highlighting meaningful patterns.

4 System and Data

Fig. 3 illustrates the architecture of our system. It consists of two major phases: data processing and visualization. As a vital prerequisite, the data processing phase collects TED Talks and extracts presentation techniques. The latter phase presents extracted information in an interactive visual analytic environment for deriving insights.

4.1 Data Processing

The data processing module comprises data collection and feature extraction. Initially our system automatically collects pertinent information of TED Talks from the official website in the chronological order, including videos, segmented transcripts and metadata. Feature extraction is then performed separately on the verbal and non-verbal mode.

Verbal techniques. Based on the domain scope discussed in Section 3.1, we develop an automatic framework to extract verbal presentation techniques. The data input is the original transcripts, which have been segmented into snippets roughly within one minute according to the discourse structure by the TED website. We automatically label each snippet with three rhetorical modes (narration, exposition and argumentation) using the state-of-the-art method [20], which is a neural sequence labeling model with an average F1-score of 0.7.

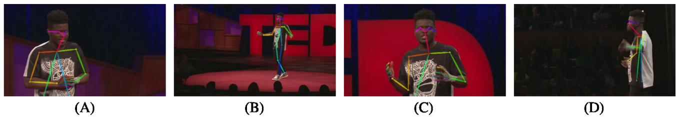


Fig. 4. Illustration of extracting body gestures based on body key points: (A) Close postures owing to overlapping hand key points. (B) Open arms since both the elbow points cross the torso region and wrist points go outermost. (C) Open postures as the hands fall in the torso region. (D) Rejected because the Face++ API does not return a frontal face.

Non-verbal techniques. We employ OpenPose [24], which is placed first in the inaugural COCO 2016 key points challenge, to detect the body keypoints, whereby gestures and postures are classified into three categories, respectively (Table 1). To accelerate the detection of body keypoints, we enable GPU acceleration with a GTX 1080 graphics card with an Intel Core i7-6600U 2.81GHz processor, achieving a processing rate of 4.4 frames-per-second on videos with a resolution of 640×480. The whole process takes more than 80 hours. Furthermore, we utilize the Face++ Landmarks API [36] to filter out non-speakers and frames where the yaw angle of speaker's head is over 30 degrees to ensure data quality. Accordingly, we apply rule-based methods to classify the postures into *open arm*, *close posture* and *open posture*, as illustrated in Fig. 4. Moreover, we experimentally adopt the thresholds for classifying hand movements as discussed in Section 3.1 into *stiff*, *expressive* and *jazz*. We allow users to specify the thresholds later in the interactive visualization module.

4.2 Data

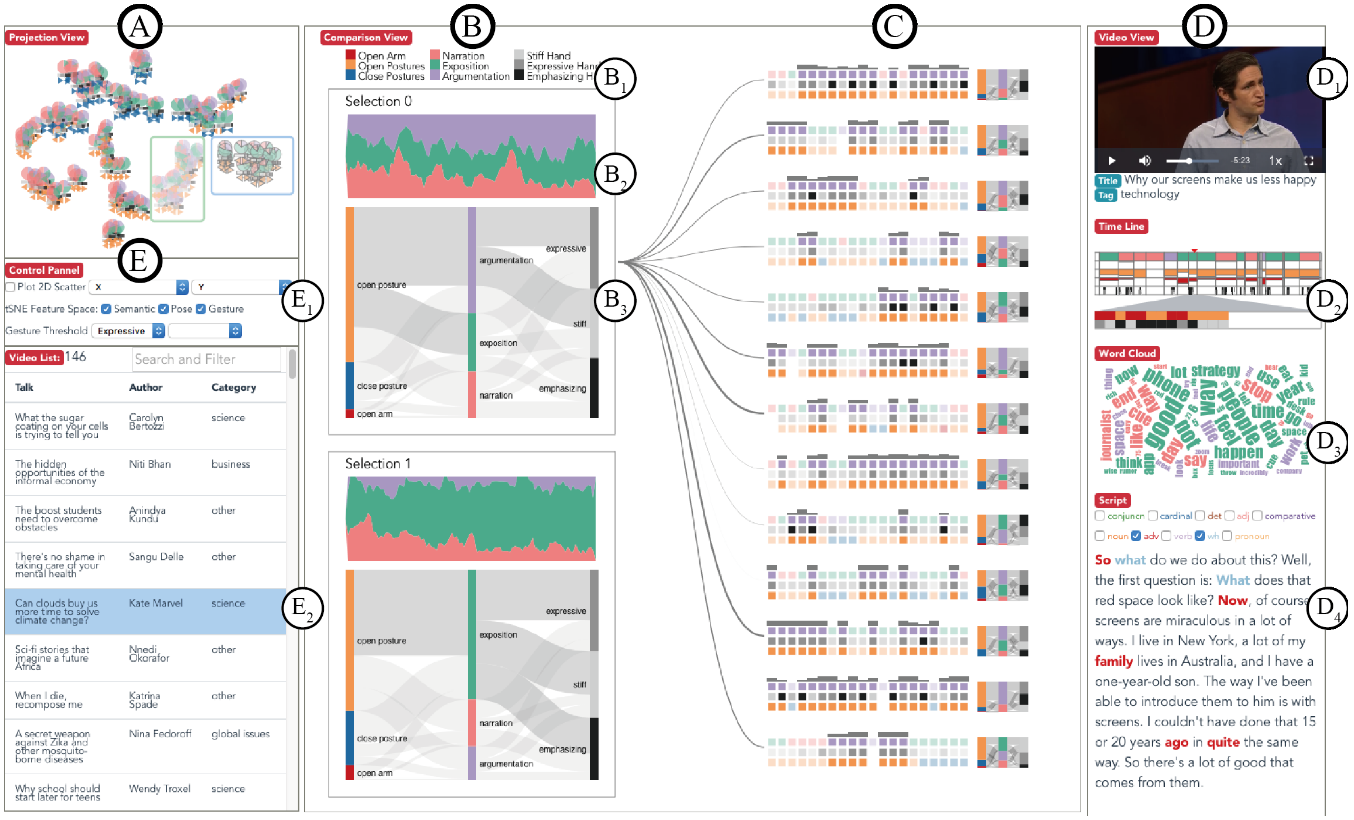
Extracted information from the processing step is a collection of data, which is included in the visualization system for analysis. The entire collection includes 146 TED Talks. As illustrated in Fig. 3, each TED Talk consists of: 1) the original video; 2) the scripts; 3) metadata such as video tags; 4) tags for body gestures per half second; 5) tags for body postures per half second; 6) verbal presentation techniques including rhetorical modes for each snippet of transcripts; and 7) a feature vector of size 9×1 describing the temporal proportion of each of the nine techniques.

5 Visual Design

The final visual design has four components (Fig. 5): the Projection View, the Comparison View, the Video View and the Control Panel. In this section, we describe the visual representations and interactions in detail and discuss our considerations in making these design decisions.

5.1 Projection View

The Projection View is motivated by the experts' suggestion for cluster analysis (T3) after the first prototype demonstration. We apply a dimension reduction method to map presentation videos, described by the feature vector, into a 2D space. Particularly, we adopt t-distributed stochastic neighbor embedding (t-SNE) [37], because it is suitable for embedding high-dimensional data into two-dimensional space and places points by similarity. As a result, the TED Talks are mapped to two-dimensional points in a way



□ Plot 2D Scatter X Y

tSNE Feature Space: ☑ Semantic ☑ Pose ☑ Gesture

Gesture Threshold Expressive

Video List: 146 Search and Filter

Talk	Author	Category
What the sugar coating on your cells is trying to tell you	Carolyn Bertozzi	science
The hidden opportunities of the informal economy	Niti Bhan	business
The boost students need to overcome obstacles	Anindya Kundu	other
There's no shame in taking care of your mental health	Sangu Delle	other
Can clouds buy us more time to solve climate change?	Kate Marvel	science
Sci-fi stories that imagine a future Africa	Nnedi Okorafor	other
When I die, recompose me	Katrina Spade	other
A secret weapon against Zika and other mosquito-borne diseases	Nina Fedoroff	global issues
Why school should start later for teens	Wendy Troxel	science

Title Why our screens make us less happy
Tag technology

Script
□ conjuncn □ cardinal □ det □ adj □ comparative
□ noun ☑ adv □ verb ☑ wh □ pronoun

So what do we do about this? Well, the first question is: What does that red space look like? Now, of course screens are miraculous in a lot of ways. I live in New York, a lot of my family lives in Australia, and I have a one-year-old son. The way I've been able to introduce them to him is with screens. I couldn't have done that 15 or 20 years ago in quite the same way. So there's a lot of good that comes from them.

Fig. 5. The interface comprises five linked components with a unified color theme (B_1): (A) the Projection View provides a holistic view of clusters; (B) the Comparison View juxtaposes two selected clusters of interest for inter-cluster comparison. It summarizes the temporal distribution of presentation techniques (B_2) and their interplay (B_3); (C) when users navigate to finer-level representations, the Presentation Fingerprinting offers a quick overview of each TED Talk and user interaction for intra-cluster comparison; (D) linked with Presentation Fingerprinting, the Video View enables detailed and contextualized case investigation of a TED Talk (D_1), supplemented by an elastic timeline (D_2), a word cloud (D_3) and an annotated transcript viewer (D_4); and (E) the Control Panel supports feature filtering (E_1) and faceted search (E_2).

that similar videos are placed nearby. We link it interactively to the Comparison View to guide overview+detail exploration (T7). When users drag or click in the Projection View, the Comparison View will automatically display corresponding features of selected TED Talks and their aggregate results. Meanwhile, we allow users to specify the feature space for dimension reduction to explore the interplay among designated techniques (T8).

Initially, each presentation was denoted by a point. However, this was not sufficient in providing visual guidance to identify patterns, since experts needed to manually select points before they could see their characteristics. This made their analytical goals cumbersome to achieve, i.e., to find examples of presentation techniques. Thus, we added a novel glyph design to offer an overview of each cluster's characteristic to ease inter-cluster comparison. The glyph encodes the same feature vector as utilized for dimension reduction, so as to completely reflect the similarity.

During the design process, we considered several alternatives as shown in Fig. 6. Although a treemap-based design (Fig. 6 (A)) fully represents the feature vector, it needs more space to achieve good legibility compared with other types. As for the pie chart design (Fig. 6 (B)), it has an inside-out visual hierarchy which could cause perception differences of visual importance. While the radar chart design (Fig. 6 (C)) avoids visual hierarchy and encodes additional information about the dominant technique by the link color, it is hard to see glyphs clearly in the Projection View. In addition, those designs suffer from high cognitive burden due to the complicated feature categories.

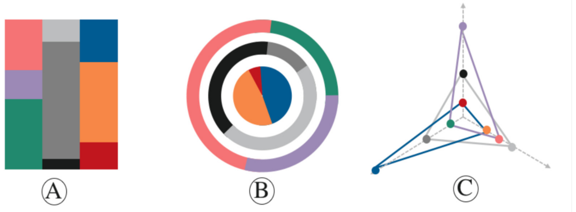


Fig. 6. Three design alternatives for the glyph where the proportions of presentation techniques are encoded by: (A) a treemap-based design; (B) a nested pie chart design; (C) a radar chart design.

Our final design (Fig. 7) is built on a metaphor of the human upper-body and consists of three independent parts. The head area, as the human organism for analyzing semantics, is represented by a pie chart encoding the proportion of rhetorical modes. A bar chart, which demonstrates the percentage of gestures, is arranged around the shoulders. Finally we use two triangles to indicate the most frequent hand posture. In some sense, this design could alleviate the cognitive pressure for digesting the encoding scheme, because each part seamlessly conforms to the widely known, entrenched functionality of corresponding human organisms. Meanwhile, each part is independently encoded with different methods to avoid confusion. Furthermore, this glyph does not have certain visual hierarchy, diminishing potential perception differences of visual importance.

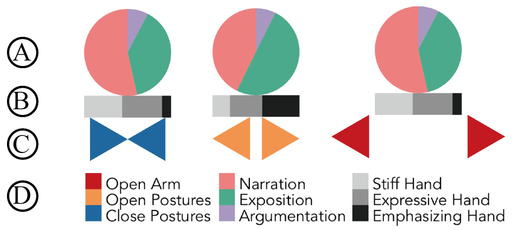


Fig. 7. TED Talk glyphs using human body metaphors. (A) The pie chart in the head area encodes rhetorical modes. (B) Hand movement is represented by the bar chart in the shoulder area. (C) The triangles encode hand postures. (D) The color-map for each category is designed in consideration of color psychology.

In addition, we carefully design the colorway for easy comprehension and memorization. Specifically, we adopt colors for gestures in line with the conveyed emotional message. For instance, the close posture is filled with cool colors since it is believed to communicate defensiveness. The three levels of hand movements are matched with suitable saturation, where increasing saturation corresponds to a larger movement. As for rhetorical modes, no common colorway exists. We therefore pick up the color scheme for qualitative data from the well acknowledged tool ColorBrewer [38], while avoiding hues that are already adopted for gestures and implying the color psychology. Specifically, we adopt pink (symbolizing life) for narration, green (reliability) for exposition, and purple (wisdom) for argumentation, according to the culture-related color psychology [39]. We integrate this unified colorway into the whole system (Fig. 7 (D)).

During the first round of design process, we observed visual clutter because t-SNE projection places similar items nearby. Although it did not impede conducting a comparison at the inter-cluster level, we found obstacles remained in two visualization tasks during the third round of iterative development. First, users had difficulty clicking onto the intended glyph due to the overlap. Regarding this, we added user interaction to attenuate the glyph capacity by default and recover it on hovering, which made users aware of items needing to be clicked. Second, it was difficult to compare nearby glyphs. Thus, we added panning and zooming user interactions into the prototype.

5.2 Comparison View

The Comparison View is the primary component since it embodies the implication of major analytic tasks (T1-4,6-7,9). During the design process, we found it challenging to display many details without aggravating the cognitive burden. Thus, we followed paper- and code-based prototyping approaches to refine this system according to the experts' feedback. In this section, we summarize the design considerations, and describe the details of each component.

5.2.1 Design Consideration

The core design consideration is grounded in the complexity of visual elements. The system should reveal both temporal aspects of various presentation techniques and their interplay. It should also present the information above for single TED Talks and their aggregation. Regarding this, we summarize design considerations based on our experts' feedback:

DC1: Prioritize aggregate results. Experts report that aggregated information carries statistical implications of technique usage, helping identify cluster-wise patterns. Aggregate results should be prioritized with large visual significance.

DC2: Enhance comparative visualization. Techniques such as juxtaposition and computed relationship should be employed to ease the inter- and intra-cluster comparison.

DC3: Summarize single TED Talk. It is necessary to present information of a single TED Talk in a concise and screen-space-efficient manner. The visual component should convey additional information besides that were already encoded in the glyph.

DC4: Adopt consistent visual encoding methods. The visual islands for single talk and aggregate results should embody consistent or at least similar encoding methods to avoid confusion.

5.2.2 Aggregate View

We juxtapose the aggregate characteristics of two user-selected clusters, empowering users to compare inter-cluster patterns (T3). Specifically, we adopt a streamgraph to encode the temporal dynamics of presentation techniques, and their concurrences are presented by a Sankey diagram to give an overview of their interplay. In the first prototype, this view only displayed the information of one selected cluster, which made inter-cluster comparison tedious. We therefore adopted the juxtaposition approach to allow comparing clusters side-by-side (DC2). Meanwhile, we increased the visual space for the Aggregate View to prioritize aggregate results (DC1).

Streamgraph. We use the streamgraph chart, a widely implemented continuous data visualization technique, to show the temporal distribution of presentation techniques (T1). The visual encoding is designed around the rhetorical modes, because it is temporally continuous, while other techniques could be non-retrievable from certain video clips.

Sankey diagram. We adopt the Sankey diagram to display the concurrences among presentation techniques (T2). Initially we investigate the design space for encoding categorical concurrences. The sunburst diagram and concurrence matrix chart are excluded because they are not screen-space-efficient. We further consider the parallel set chart and the Sankey diagram, among which our experts prefer the latter because it does not assume a hierarchical order. In the Sankey diagram, we put rhetorical modes in the middle bar-set to represent their interplay with non-verbal techniques. We automatically calculate the top three concurrence tuples and render corresponding Sankey links with higher saturation (T9). This computed relationship approach helps suppress irrelevant details and facilitates easy comparison (DC2).

5.2.3 Presentation Fingerprinting

We propose a hierarchical visual component called Presentation Fingerprinting to create a holistic view of presentation techniques used in TED Talks and facilitate the intra-cluster comparison (T4). The visual encoding is extended from that of the aggregate view to avoid an overwhelming visualization that could confuse users (DC4). Akin to the Aggregate View, we separately summarize and encode the temporal distribution of presentation techniques and their interplay (DC3). However, we substitute a tabular design for the streamgraph to display the temporal distribution.

The tabular design is inspired by Kurzhals et al's approach for visual movie analysis [13], which depicts heterogeneous features of a video as rows composed of tabular cells. While there are several alternatives such as the story curve, streamgraph and time series, we adopt this design due to two considerations. First, it does not assume a scale along the vertical axis, which is usually

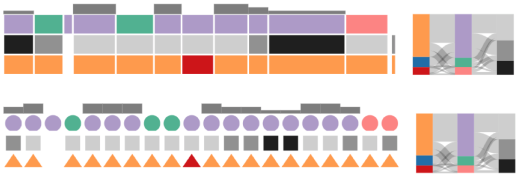


Fig. 8. Design alternatives for Presentation Fingerprinting. Top: Time grid is based on predefined script segments; Bottom: The shape channel encodes the technique categories in the same way as the aforementioned glyph.

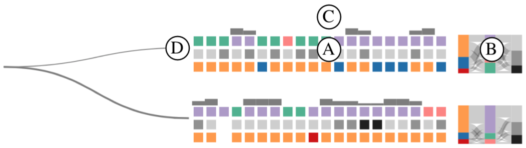


Fig. 9. Presentation Fingerprinting representing each TED Talk: (A) the tabular diagram encodes the temporal distribution of techniques during a presentation; (B) the Sankey diagram encodes their concurrences; (C) a bar chart is embedded onto the tabular design to depict the top concurrences of presentation techniques; (D) the edge visually links individual talks with aggregate results.

required in those line-based designs; Second, it is particularly well-suited for displaying multivariate heterogeneous data, since each variable is independently encoded by a row without clutter. In our design, each row from top to bottom encodes the predominately used rhetorical modes, gestures and postures within a time interval, respectively.

For the first prototype, we set the time grid by the predefined script segment (Fig. 8 top). These lead to a list of tabular grids with a varying grid, which our experts commented was messy and thus not easy to conduct intra-cluster comparison. Therefore, we adopt a uniform time interval of five percent of the TED Talk duration. Besides, we also consider encoding the technique categories by the shape channel in the same way as the aforementioned glyph to reduce the memory burden (Fig. 8 bottom). We allow users to specify encoding methods.

To ease intra-cluster comparison, we borrow the idea of Heterogeneous Embedded Data Attribute (HEDA) [40], which is a tabular visualization building block which embeds and extends common, familiar visualization techniques. We consider our tabular design as a HEDA component and integrate it with bar charts and links. The bar chart is embedded onto the tabular design to illustrate the top concurrence tuples (DC2). By default, the top three are encoded where the bar height encodes the frequency. For example, in Fig. 9(C), we can observe that the concurrence tuple (argumentation, expressive hand, and open posture) appears the most in this cluster. Moreover, users can highlight them by clicking the Sankey diagram to reduce visual clutter and dilutes less meaningful information (T9). Fig. 5(C) illustrates such effect.

Interface interaction is enabled for contextualized analysis of a TED Talk (T7). When clicking a Presentation Fingerprinting, detailed information of the corresponding TED Talk is displayed in the Video View.

5.2.4 Video View

The Video View illustrates the original content of a TED Talk, enabling users to examine the interplay of verbal and non-verbal techniques in detail and verify observations in the video space (T7). It consists of four components: a video player, a word cloud, a script viewer, and an elastic timeline. The video player shows the video, title and tag of a TED Talk. The word cloud displays the frequent words whose color represents corresponding rhetorical modes. The transcripts of the currently playing segment are displayed in the script viewer. Besides the textual content, users might selectively highlight words of certain part-of-speech tags such as a conjunction. This could help users to identify the interplay pattern between gestures and semantics.

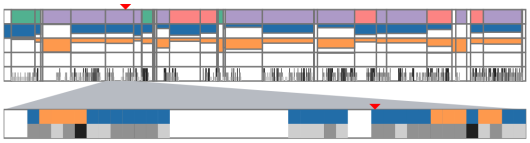


Fig. 10. Elastic Timeline. It consists of two layers: the top layer provides an at-a-glance overview of the presentation techniques; selecting a column unfolds the bottom layer which shows the appearance of non-verbal presentation techniques per half second. In addition, timestamps are provided on both layers to enhance browsing experience.

5.2.5 Elastic Timeline

We provide an elastic timeline to facilitate the browsing and analyzing of the TED Talk (T5). It (Fig. 10) has two layers. At the top layer, the timelines are segmented according to the transcript snippet. The usage of presentation techniques is arranged vertically with rows to reveal the at-a-glance distribution (T1), akin to the Presentation Fingerprinting with two differences. First, three types of body postures are independently represented by three rows in the middle. Within the cells of these rows, bar charts are displayed to show the proportion of corresponding posture during the time interval. This design decision is directly motivated by experts' feedback on our initial design, which is stacked bar charts with a single row. Our experts prefer the former, considering the tradeoff between visualization compactness versus cognitive loads. Second, we adopt the bar chart within the entire row to represent the body gestures. The gesture category and normalized movement are decoded by the height and color of the bar, respectively.

Clicking on the top layer will unfold the bottom layer, which displays the gestures and postures during the selected segment. Each grid corresponds to a half second and a consistent colormap is assigned. A blank grid denotes that any information is non-retrievable at that time.

The elastic timeline is linked with the video player and the script viewer. When clicking any entry, the corresponding frame and text are updated. To optimize the browsing experience, timestamps are provided in the video player, the script viewer and both layers of the elastic timeline.

6 Evaluation

To evaluate our system, we applied it to 146 TED Talks, implemented the final prototype with Flask, VueJS and D3, and ran two studies: a case study with experts and students, to reflect the fulfillment of domain-specific analytical goals and gained insights, and a user study with 16 target users, to demonstrate the capacity of undertaking visualization tasks and gather feedback on the visualization design.

6.1 Case Study

The goal of this case study was to demonstrate how our system assist users in reaching their domain analytical goals (Section 3.2). We conducted a case study with our domain experts (E0-2) whom we collaborated with during the design progress and three undergraduate freshmen (S0-2) who wished to learn presentation techniques. The whole session lasted for around one hour. It started with a demonstration where we explained the visualization design and interaction. We additionally introduced the backgrounds of related presentation techniques for students, since they had not been trained with related knowledge. Subsequently, users were given thirty minutes to openly navigate through the system, during which they were encouraged to speak about their findings and express opinions. In the meantime, their feedback was recorded. Finally, the sessions ended with twenty-minute post-interview discussion on findings and comments.

In this section, we describe how users utilize our system to acquire insights into presentation techniques used in TED Talks. We integrate their findings into a coherent case study.

6.1.1 Obtain an Overview

At the outset, our users wanted to obtain an overview of the presentation techniques used in the TED Talks (G1, G2). E0 decided to explore the overall semantic structure and compare it with the theory in the TED official guide [41], which advocates the story-idea-evidence-idea structure. After selecting all talks in the Projection View, the general picture was unfolded in the Comparison View. Following the streamgraph which represents the temporal distribution of rhetorical modes, E0 immediately observed the large pink stacked area, which indicates the dominant usage of narration. She commented: *"This is in line with the expectation, because storytelling is the golden standard of TED Talks."* She further noted that around 60% of TED Talks started with narration, thereafter exposition gradually increased and overtook narration in the last one third of the presentations. In the meantime, the proportion of argumentation was maintained around 15% throughout the presentation, except for a trough for the prologue and a surge in the coda. She endorsed these results from a professional standpoint: *"This conforms to the paradigm. Presentations usually start with telling stories or citing instances. Then speakers express their opinions in brief, supported by substantial evidence. They emphasize their ideas at the end."*

Afterwards, she shifted attention to the Sankey diagram which showed the interplay of verbal and non-verbal techniques. She first noticed the three bar-sets and found that open postures and expressive hands were the dominant postures and gestures, respectively. Stiff hands accounted for roughly 30%, suggesting that speakers moved their hands 70% of the time. In addition, close postures made up approximately 35% of the body postures, which drew E1's attention. *"Close postures are considered negative. I didn't expect such heavy use"*, he added, *"This deserves in-depth exploration later"*. E0 then inspected the flows and inferred, *"Narration appears to accompany open postures and expressive hands to a greater extent, while stiff hands occur with argumentation more frequently."*

6.1.2 Identify Presentation Styles

To get a deeper understanding of how the usage of presentation techniques vary among different groups (G3), E2 decided to examine clusters in the Projection View and compare the patterns.

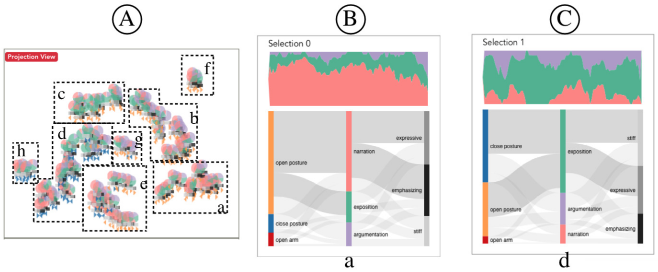


Fig. 11. To identify the presentation styles and compare the patterns. (A) Observe eight clusters in the Projection View; (B) Compare two presentation styles indicated by the concurrence pattern in the Sankey diagram: cluster *a* (story-telling style) has a major concurrence among open posture, narration, and expressive gesture; (C) whereas cluster *d* (scientist style) utilizes close posture, exposition, and stiff gesture.

Fig. 12. Story-teller and Scientist Style. Left: the speaker passionately narrated a story about hair with incessant gestures. Right: the speaker explained the global learning crisis in a sedate manner.

After specifying the feature space in the Control Panel, he noticed eight clusters in the Projection View as depicted in Fig. 11 (A), including five large clusters (*a-e*) and three small groups of outliers (*f-g*). He selected the cluster *a* (Fig. 11 (B) and inspected its Sankey diagram where the top most frequently concurrent tuples were highlighted, whereby he observed the dominant concurrence of open postures, narration, and expressive hands. He clicked a Presentation Fingerprinting with a thick edge, which indicated its typicality. He glanced over the talk and verified that the speaker used a rich set of body language to narrate a vivid story (Figure 12 Left). *"This is the typical storytelling style"*, he concluded.

Consequently, he decided to investigate the difference with other clusters by turning back to the Projection View, where he noted from the glyph design that cluster *d* had a rather distinct colorway. Hence, he counterposed cluster *a* with *d* (Fig. 11 (C)), in which he found heavy usage of close posture, exposition, and stiff hands. Similarly, he skimmed through a video and confirmed that the speaker explained the global learning crisis in a demure manner and only used gestures for emphasizing (Fig. 12 Right). He explained, *"This is the scientist style. They deliver complex information and look unemotional or reserved."*.

He showed great interest in case-based exploration of the remaining three groups of outliers (G4), since he highly valued learning from abnormal or negative examples. He selected cluster *f* and found that emphasizing gestures accounted for more than 50% within these five TED Talks. Guided by the highlighted edge and bar chart of the Presentation Fingerprinting ((Fig. 5 (C) illustrates a similar effect), he soon identified one typical TED Talk. To investigate the case, he clicked on it and browsed the video recording in the Video View. Assisted by the elastic timeline, he easily navigated to video clips with emphasizing gestures. He observed that the speakers adopted an exaggerated set of gestures such as fully spread arms, in the context of expressing strong

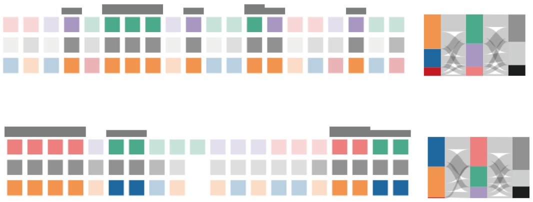


Fig. 13. Presentation Fingerprinting of Prof. Bertozzi (top) and Dr. Wayne (bottom). The top concurrence tuples among presentation techniques are highlighted.

emotions. *"This looks overwhelming and might be offensive. His hands frequently get out of the Strike Zone - the area that your hands should occupy when giving presentations. This is bad"*, he commented. He then examined cluster f in the same way, in which those six speakers kept wringing their hands during more than 70% of the presentation. He criticized such behavior because it conveys nervousness and defensiveness. Finally, he unfolded cluster g which predominately relied on argumentation and said, *"This shows that it is atypical to keep expressing ideas without narration or exposition. Audiences might get bored."*

6.1.3 Compare Cases

The freshman S0 wished to explore the way that biological scientists deliver presentations (G3, G5), since she planned to study biological science. Through searching the term *cell* in the control panel, she found two related speakers - Dr. Wayne and Prof. Bertozzi. She selected their presentations and inspected the Presentation Fingerprinting. From the first row, she noted that both presentations started with narration and ended with exposition. She then shifted attention to the Sankey diagram, where she found major differences: the bar-sets of Prof. Bertozzi had large yellow and green color blocks, which indicated her dominant usage of open postures and exposition. In contrast, Dr. Wayne seemed to adopt open/close postures, narration, and exposition equally. She therefore wondered about their interplay and highlighted the top concurrence tuples. As shown in Fig. 13, Prof. Bertozzi majorly adopted open postures for exposition and argumentation, while Dr. Wayne gave expositions with close postures and told stories with open postures. She then glanced over two presentations in ten minutes and concluded, *"They are of two styles. Bertozzi gave much information with rich body language to engage audiences. Instead, Wayne told a scientific story. She usually held her hands together when explaining things to make her look more serious."*

6.1.4 Deep Investigation

With an interest in resolving doubt about the unexpected heavy use of close postures as observed when obtaining an overview, E1 decided to investigate the cases from a different prospective, supplemented by case investigation (G4, G6). To select the TED Talk with larger use of close postures, he specified the feature space as posture in the control panel. Following the glyphs in the Projection View, he quickly located two clusters with predominant close postures indicated by the blue color. He draw the same conclusions with E2 on the smaller cluster, interpreting them as negative examples where the wringing of their hands communicated nervousness. Consequently, he selected the larger cluster containing 43 TED Talks and observed that close postures made up around 50%. After spending eighteen minutes quickly browsing 20 of such TED Talks with the help of interface interaction and the elastic timelines, he explained, *"Besides four ladies who wear sleeveless dress and adopt this posture to show elegance, most of them just wring their hands for a short rest. This does not meet the standards. Speakers are encouraged to simply put hands on sides for resting."* Further, he added, *"However, speakers in those TED Talks do not look very nervous, since they also adopt rich gestures during the other half time. This finding changes my minds on close postures - they are complemented by gestures."*

6.1.5 Post-interview Discussion

Effectiveness. During the post-interview discussion, the experts and students appreciated the effectiveness of our system to reach the domain analytical goals. Through the above case study, our experts gained domain-specific insights and endorsed them from their professional points of view. They commented that these findings generally matched the existing theories in the education field, and further offered statistical evidence for the usage of presentation techniques. They particularly valued the visual analytic system. E2 said, *"The visualization empowers me to quickly identify presentation styles and negative examples, which are not possible with our existing methods."* E1 added, *"it also leads to some new findings. I am surprised by the heavy use of close postures, since we have been teaching students to avoid that. It deserves further discussion in our field."*

Potential Applications. More encouragingly, our experts would like to incorporate the system into their current researching and teaching practices. E0 showed strong interests in statistic-based evidence of presentation techniques. She suggested a wider classification of gestures such as listing or pointing to enhance the system use for research purposes. She also posed new questions about the sequential pattern of gestures, for which she wished our system to render a deeper understanding. E2 saw the chance of utilizing our system to assist in "example and non-example" teaching, since he succeeded to identify several examples and non-examples. He expected us to extend the current pool of presentation videos. Meanwhile, he commented that it would be more powerful if our system could automatically return information about rhetorical trajectory and body language by inputting any presentation video.

6.2 User Study

Our visualization system provides a prospective on analyzing multimodal content in video collections. The goal of this study was to evaluate how our system could assist in the proposed visualization tasks (Section 3.3). We also wanted to gather user feedback on the usefulness of our visual design.

6.2.1 Set-up

We recruited 16 students (9 male, 7 female) between the ages of 19-28 (mean 23.18, std 2.482). They comprised both undergraduate and postgraduate students, with backgrounds in computer science, electrical engineering, humanities, and life science. They all had normal or correct-to-normal vision.

Each study took approximately 45 minutes and was conducted on a 13-inch MacBook Pro Retina machine. The first 15 minutes were used to explain the visualization system, followed by 5 minutes of free exploration. The participants then went through a series of 8 tasks in 15 minutes. Those tasks were designed to utilize all visual components and reflect the whole set of proposed

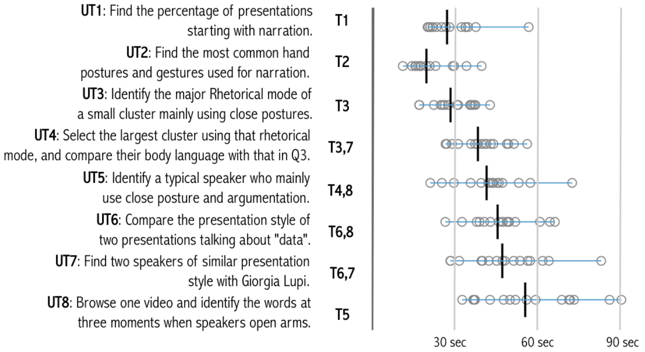


Fig. 14. User tasks and completion time. The middle column indicates corresponding visualization tasks (Table 2). Black bars shows the mean.

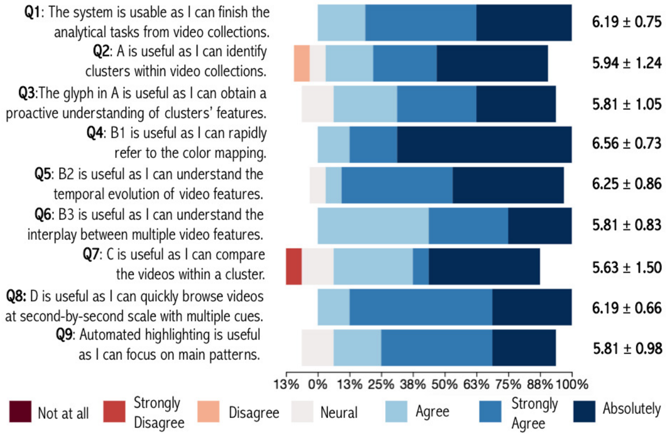


Fig. 15. Description and results of questionnaires. The rightmost column denotes Mean ± SD.

visualization tasks in Section 3.3, except for T9 which did not involve user interaction. Participants were timed from when they started to when they wrote down their answer. The visualization system was reset prior to each task. They were offered another 5 minutes for free exploration before they filled in questionnaires. It ended with a short post interview.

6.2.2 *Task Completion*

We represent the description of each user task and completion time in Fig. 14. All participants completed any task within 91 seconds, while all tasks were finished in less than 1 minute on average. We observed small variations for UT1-4, which were close-ended questions with a fixed answer. This suggests that all participants could understand the task and complete them rapidly. One outlier existed in UT1, with which a participant explained that she had difficulties to start.

We noted large variations starting from UT5, which was mostly due to their increased difficulties. Solving UT5-8 would require more user interactions and observation among visual components, and the analytical methods and answers were not necessarily unique. Participants exhibited varying response times in both coming up with solutions and utilizing the system.

6.2.3 *Questionnaire*

The questionnaires were designed to evaluate the usefulness of our system and gather explicit feedback on the visual design (Fig. 5). They were created under the guidance suggested by Rossi et al. [42] to ensure reasonableness, appropriateness, feasibility, and stakeholder engagement. We adopted a 7-likert scale and calculated the mean score and standard deviation.

We provide the description and results of each questionnaire in Fig. 15. As a whole, all participants agreed that our system is usable for analytical tasks on video collections (Q1). For visual components (Q2-9), our system received an average rating of 5.96 out of 7, which was very encouraging. Specifically, our participants were very satisfied with the legend (Q4), the themeriver (Q5), and the Video View (Q8). The Projection View was broadly appreciated (Q2-3), although we received one criticism that it should take larger screen space.

Participants reported less satisfaction with the Video View, especially the Sankey Diagram (Q6) and the Presentation Fingerprinting (Q7). In the post-questionnaire interview, they explained that they were unfamiliar with the Sankey Diagram, thus it took more time to comprehend. They mentioned that they would prefer more daily-use charts. For the presentation fingerprinting, most participants reported that they mainly used them to verify their findings obtained in the aggregate window, and were less interested in examining the details of each presentation.

7 Discussion

Our work addresses two main aspects: a domain-specific aspect for visual analysis of presentation techniques, and a system aspect for visualization interfaces for multimodal content in video collections. In addition, we discuss our lessons learned and implications of multimedia visual analytic system design.

7.1 Visual Analysis of Presentation Techniques

Inspired by expert's current practices, our work explores the verbal and non-verbal presentation techniques in TED Talks. Our work has completed domain analytical goals, and the findings conform to and further extend existing knowledge. Domain experts appreciate our system and see the opportunity to utilize it in their research and teaching practices. However, our analysis has several limitations.

Research scope. Presentation is a complex undertaking which requires the combined harmony of verbal and non-verbal actions as well as other devices. Our work only explores three important presentation techniques suggested by our experts. Based on our experience, we outline the following aspects for follow-up studies: facial expression, use of space, tones, vocal emphasis, and eye contact. However, such analysis presents challenges on integrating reliable computer vision and natural language processing techniques, as well as dealing with uncertainty. In addition, exploring presentation videos under other scenarios could lead to a more comprehensive analysis.

Accuracy. Our system could not address the accuracy issue of extracted data, which might be detrimental to the analytical process. Although we adopted the state-of-the-art method, it could not achieve 100% accuracy. We therefore outline two future works: 1) to annotate a corpus to study the ground truth of presentation techniques used in TED talks and evaluate it with our system; and 2) to encode the uncertainty and thereby inform users of potential inaccuracy, which needs further study in those communities to obtain the numeric confidence scores.

7.2 Visual Analysis of Multimedia Content

Our approach offers new opportunities which are not yet available in existing visualization systems for multimedia collections. We integrate a video browse interface into a visual analytic framework which guides analysis from left ro right through three interactively coordinated views in a more screen-space-effective and easy-to-distinguish manner. Each view includes appropriate visual components to summarize and represent the data at different levels of detail, supporting inter-, intra-cluster, and within-video analysis. We also provide interface linking and faceted search to prompt analytical effectiveness.

Generalizability. While our work targets at presentation videos, we speculate that our approach would assist in similar video analysis in other domains. During our design process, we have abstracted visualization tasks, which cover a wider category than the baseline system and guide our design. The visualization framework could be adopted to similar analysis in other domains. One example is to explore the film styles reflected by scenes and scripts. We could extract the scene-setting (e.g. urban, natural) and the events (e.g. dialogue, tussle) from videos, as well as speaking styles (e.g. romantic, trendy) from scripts. Consequently, these data could be used and transferred with our system by some modifications on visual components such as the glyph.

Limitations. We identified several limitations of the visual design. First, the Presentation Fingerprinting does not add significantly to the analytic process as a whole, since users would prefer to use it as a verification and navigation tool. We plan to simplify the Sankey diagram component by fading out or removing the links, and enhance the whole functionality by encoding additional information such as the sequential pattern of body gestures, which could support more visualization tasks. Second, while we have adopted methods to reduce visual clutter in the Projection View, the overlapping among glyphs is still unavoidable with the increasing number of videos. Allowing for advanced focus-plus-context techniques, hierarchical clustering, and automated aggregation so that each glyph represents a cluster might ameliorate this problem. Third, our system only supports side-by-side comparison between two clusters. We plan to integrate more approaches such as a difference view which directly encode differences among multiple clusters. This will pose new challenges on the screen space usage and layout. Finally, our system utilizes different visualization to encode temporal information to support particular visual tasks for each standalone view, which might place a cognitive burden on users. To tackle this problem, we can design a new visualization for temporal data to achieve higher consistence among the system.

7.3 Design Implication

We discuss our lessons learned about designing multimedia visual analytic systems.

From exploration to query. We observed that the experts' analytical tasks transited gradually from exploration-based to query-based in the analytic progress. In the beginning, the task was to explore and gain a brief understanding of the multimedia collection. As their understanding got deepen, they conducted more querying tasks such as to find particular videos and even to examine frames. This is different from visual analytic in classic database where the analyst does not necessarily query and examine a single data item. We suspect that the integration of exploration and query is critical for multimedia visual analytic system. This poses challenges on designing visualization components, which could serve as both analysis and navigation tools and encode data consistently at both collective and individual level.

Raw multimedia vs. extracted features. Another difference of multimedia visual analytics is that human are already expert at perceiving and analyzing the raw data - the images and videos. However, human understanding is not machine-readable. While feature extraction is necessary for automatic analysis, it carries less meaning to human and leaves their expertise wasted. Such dissonance could impede the visual analysis process. As such, we feel that it is essential to extract high-level semantic features which renders intuitive understanding. In addition, while our approach visualizes both the raw multimedia and extracted features separately yet interactively, we believe that designing a more integrated visualization warrants further research.

8 Conclusion

In this work, we present a design study for developing a visual analytic system, which empowers users to explore the verbal and non-verbal techniques in TED Presentations. It enables interactive analysis of multimodal data and mainly focuses on three presentation techniques, *i.e.*, rhetorical modes, body postures and gestures. Through an iterative design progress with domain experts, we characterize the domain-specific analytical goals and apply the-state- of-art methods to extract related techniques. Moreover, we derive a set of visualization tasks that guide our design. Two studies are conducted to evaluate our system. An in-depth case study with domain experts and general users demonstrates the effectiveness of our approach in achieving analytical goals. The findings accord with and further supplement the existing theories. Through a user study, we demonstrate the capacity of undertaking visualization tasks for visual analysis of video collections. We validate the usefulness and show that our system could support more analytical tasks compared with the baseline system.

In future work, we intend to acquire a wider understanding of presentation techniques by (a) integrating advanced algorithms to extract additional features and improve the accuracy; and (b) enhancing the visual design to assist more analytical tasks such as sequential mining. We also plan to evaluate our system with videos in other presentation scenarios and application domains, moving towards large-scale, multimodal multimedia analytics.

Acknowledgments

We would like to thank Siwei Fu and Yong Wang for proofreading the paper, all the study participants and the anonymous reviewers for their valuable comments. This work is partly supported by grant ITF UIT/138.

References

[1] T. Currie, "Ten tips to help you gesture like a genius during your next presentation," May 2015. [Online]. Available: https://podiumconsulting.com/the-podium-project/2015/5/7/ten-tips-to-help-you-gesture-like-a-genius-during-your-next-presentation/

[2] P. Khoury, "5 "talking with your hands" rules charismatic leaders use," March 2017. [Online]. Available: https://magneticspeaking.com/hand-gestures-and-talking-with-your-hands-presenting/

[3] C. R. Sugimoto, M. Thelwall, V. Larivière, A. Tsou, P. Mongeon, and B. Macaluso, "Scientists popularizing science: Characteristics and impact of TED talk presenters," *PLoS ONE*, vol. 8, no. 4, p. e93609, 2013.

[4] A. Tsou, M. Thelwall, P. Mongeon, and C. R. Sugimoto, "A community of curious souls: An analysis of commenting behavior on TED talks videos," *PLoS ONE*, vol. 9, no. 4, p. e62403, 2014.

[5] D. Taibi, S. Chawla, S. Dietze, I. Marenzi, and B. Fetahu, "Exploring TED talks as linked data for education," *British Journal of Educational Technology*, vol. 46, no. 5, pp. 1092–1096, 2015.

[6] C. R. Sugimoto and M. Thelwall, "Scholars on soap boxes: Science communication and dissemination in TED videos," *Journal of the Association for Information Science and Technology*, vol. 64, no. 4, pp. 663–674, 2013.

[7] N. Pappas and A. Popescu-Belis, "Sentiment analysis of user comments for one-class collaborative filtering over TED talks," in *Proceedings of the 36th international ACM SIGIR conference on Research and Development in Information Retrieval*, 2013, pp. 773–776.

[8] T. Tsai, "Are you TED talk material? comparing prosody in professors and TED speakers," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, 2015.

[9] D. Kravvaris and K. L. Kermanidis, "Speakers' language characteristics analysis of online educational videos," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 2014, pp. 60–69.

[10] M. Stein, H. Janetzko, A. Lamprecht, T. Breitkreutz, P. Zimmermann, B. Goldlucke, T. Schreck, G. Andrienko, M. Grossniklaus, and D. A. Keim, "Bring it to the pitch: Combining video and movement data to enhance team sport analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 13–22, 2017.

[11] C. Perin, R. Vuillemot, and J.-D. Fekete, "Soccerstories: A kick-off for visual soccer analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2506–15, 2013.

[12] A. Pavel, D. B. Goldman, B. Hartmann, and M. Agrawala, "Sceneskim: Searching and browsing movies using synchronized captions, scripts and plot summaries," in *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, 2015, pp. 181–190.

[13] K. Kurzhals, M. John, F. Heimerl, P. Kuznecov, and D. Weiskopf, "Visual movie analytics," *IEEE Transactions on Multimedia*, vol. 18, no. 11, pp. 2149–2160, 2016.

[14] M. Höferlin, B. Höferlin, G. Heidemann, and D. Weiskopf, "Interactive schematic summaries for faceted exploration of surveillance video," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 908–920, 2013.

[15] B. Duffy, J. Thiyagalingam, S. Walton, D. J. Smith, A. Trefethen, J. C. Kirkman-Brown, E. A. Gaffney, and M. Chen, "Glyph-based video visualization for semen analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 8, pp. 980–993, 2015.

[16] J. Adcock, M. Cooper, L. Denoue, H. Pirsiavash, and L. A. Rowe, "Talkminer: A lecture webcast search engine," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 241–250.

[17] J. Matejka, T. Grossman, and G. Fitzmaurice, "Video lens: Rapid playback and exploration of large video collections and associated metadata," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software & Technology*, 2014, pp. 541–550.

[18] M. I. Tanveer, S. Samrose, R. A. Baten, and M. E. Hoque, "Awe the audience: How the narrative trajectories affect audience perception in public speaking," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 24:1–24:12.

[19] C. Gallo, *Talk Like TED*. Emerald Group Publishing Limited, 2014.

[20] S. Wei, W. Dong, F. Ruiji, L. Lizhen, L. Ting, and H. Guoping, "Discourse mode identification in essays," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 112–122.

[21] L. Rozakis, *The Complete Idiot's Guide to Grammar and Style*. Penguin, 2003.

[22] J. R. Zhang, "Upper body gestures in lecture videos: indexing and correlating to pedagogical significance," in *Proceedings of the 20th ACM International Conference on Multimedia*, 2012, pp. 1389–1392.

[23] S. Okada and K. Otsuka, "Recognizing words from gestures: Discovering gesture descriptors associated with spoken utterances," in *Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition*, 2017, pp. 430–437.

[24] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the 2017 IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1302–1310.

[25] J. McGregor and S. Tan, "What to do with your hands when speaking in public," November 2015. [Online]. Available: https://www.washingtonpost.com/news/on-leadership/wp/2015/11/17/what-to-do-with-your-hands-when-speaking-in-public/?utm_term=.e26c92c801d0

[26] A. Biswas, A. Gandhi, and O. Deshmukh, "Mmtoc: A multimodal method for table of content creation in educational videos," in *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 621–630.

[27] A. Pavel, C. Reed, B. Hartmann, and M. Agrawala, "Video digests: a browsable, skimmable format for informational lecture videos," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, 2014, pp. 573–582.

[28] A. Haubold and J. R. Kender, "Augmented segmentation and visualization for presentation videos," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, 2005, pp. 51–60.

[29] J. Kim, P. J. Guo, C. J. Cai, S.-W. Li, K. Z. Gajos, and R. C. Miller, "Data-driven interaction techniques for improving navigation of educational videos," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, 2014, pp. 563–572.

[30] S. Ortiz, "Lostalgic." [Online]. Available: http://intuitionanalytics.com/other/lostalgic/

[31] N. W. Kim, B. Bach, H. Im, S. Schriber, M. Gross, and H. Pfister, "Visualizing nonlinear narratives with story curves," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 595–604, 2018.

[32] K. Schöffmann and W. Bailer, "Video browser showdown," *SIGMultimedia Rec.*, vol. 4, no. 2, pp. 1–2, Jul. 2012. [Online]. Available: http://doi.acm.org/10.1145/2350204.2350205

[33] K. Higuchi, R. Yonetani, and Y. Sato, "Egoscanning: Quickly scanning first-person videos with egocentric elastic timelines," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 6536–6546.

[34] V. V. Edwards, "You are contagious," https://www.youtube.com/watch?v=cef35Fk7YD8.

[35] J. Kitzinger and H. Hauser, "Visualization and visual analysis of multifaceted scientific data: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 3, pp. 495–513, 2013.

[36] "Face landmarks," https://www.faceplusplus.com/landmarks/, Face++.

[37] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[38] M. Harrower and C. A. Brewer, "Colorbrewer.org: An online tool for selecting colour schemes for maps," *The Cartographic Journal*, vol. 40, no. 1, pp. 27–37, 2003.

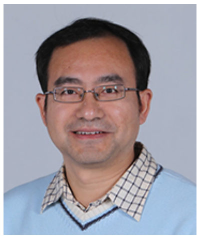
[39] G. Ciotti, "The psychology of color in marketing and branding," March 2018. [Online]. Available: https://www.helpscout.net/blog/psychology-of-color/

[40] M. H. Loorak, C. Perin, C. Collins, and S. Carpendale, "Exploring the possibilities of embedding heterogeneous data attributes in familiar visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 581–590, 2017.

[41] "TEDx speaker guide," https://storage.ted.com/tedx/manuals/tedx_speaker_guide.pdf/, August 2018.

[42] P. H. Rossi, M. W. Lipsey, and H. E. Freeman, *Evaluation: A Systematic Approach*, 7th ed. California: SAGE Publications, 2003.

Aoyu Wu received his BEng degree in Electronic Engineering and Computer Science from the Hong Kong University of Science and Technology (HKUST) in 2017. He is currently a MPhil student at HKUST. His research interests include data visualization and human-computer interaction.



Huamin Qu is a full professor in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. His main research interests are in visualization and human-computer interaction, with focuses on urban informatics, social network analysis, e-learning, text visualization, and explainable artificial intelligence.