Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

1-2020

EmoCo: Visual analysis of emotion coherence in presentation videos

Haipeng ZENG

Xingbo WANG

Aoyu WU

Yong WANG
Singapore Management University, yongwang@smu.edu.sg

Quan LI

See next page for additional authors

Citation
1

Author
Haipeng ZENG, Xingbo WANG, Aoyu WU, Yong WANG, Quan LI, Alex ENDERT, and Huamin QU

EmoCo: Visual Analysis of Emotion Coherence in Presentation Videos

Haipeng Zeng, Xingbo Wang, Aoyu Wu, Yong Wang, Quan Li, Alex Endert and Huamin Qu

Fig. 1: Our visualization system supports emotion analysis across three modalities (i.e., face, text, and audio) at different levels of details. The video view (a) summarizes each video in the collection and enables quick identification of videos of interest. The channel coherence view (b) shows emotion coherence of the three modalities at the sentence level and provides extracted features for channel exploration. The detail view (c) supports detail exploration for a selected sentence with some highlighted features and transition points. The sentence clustering view (d) provides a summary of the video and reveals the temporal patterns of emotion information. The word view (e) enables efficient quantitative analysis at the word level in the video transcript.

Abstract— Emotions play a key role in human communication and public presentations. Human emotions are usually expressed through multiple modalities. Therefore, exploring multimodal emotions and their coherence is of great value for understanding emotional expressions in presentations and improving presentation skills. However, manually watching and studying presentation videos is often tedious and time-consuming. There is a lack of tool support to help conduct an efficient and in-depth multi-level analysis. Thus, in this paper, we introduce *EmoCo*, an interactive visual analytics system to facilitate efficient analysis of emotion coherence across facial, text, and audio modalities in presentation videos. Our visualization system features a channel coherence view and a sentence clustering view that together enable users to obtain a quick overview of emotion coherence and its temporal evolution. In addition, a detail view and word view enable detailed exploration and comparison from the sentence level and word level, respectively. We thoroughly evaluate the proposed system and visualization techniques through two usage scenarios based on TED Talk videos and interviews with two domain experts. The results demonstrate the effectiveness of our system in gaining insights into emotion coherence in presentations.

Index Terms—Emotion, coherence, video analysis, visual analysis

---

- *H. Zeng, X. Wang, A. Wu, Y. Wang and H. Qu are with Hong Kong University of Science and Technology. Y. Wang is the corresponding author. E-mail: {hzengac, xingbo.wang, awuac, ywangct}@connect.ust.hk, huamin@cse.ust.hk.*
- *Q. Li is with AI Group, WeBank, China. E-mail: forrestli@webank.com.*
- *A. Endert is with Georgia Institute of Technology. E-mail: endert@gatech.edu.*

1 INTRODUCTION

Emotions play an important role in human communication and public speaking. Most recent literature advocates using emotional expressions that can improve audience engagement and lead to successful delivery [12]. As humans express emotions through multiple behavioral modalities, such as facial and vocal changes, emotion coherence across those modalities can have significant effects on the perception and attitudes of the audience [10]. Therefore, exploring multimodal emotions and their coherence can be of great value for understanding emotional expressions in presentations and improving skills. Nevertheless, existing research in multimedia [25, 26, 31] has mainly focused on integrating multimodal features to recognize and analyze the overall emotion in presentations. Thus, they are insufficient for analyzing scenarios with incoherent emotions expressed through each modality, which can occur inadvertently [33, 50] or deliberately (e.g., deadpan

humor). To this end, an analysis tool for systematically exploring and interpreting emotion coherence across behavioral modalities is needed to gain deeper insights into emotional expressions.

Visual analytics have been introduced in emotion analysis to ease the exploration of complex and multidimensional emotion data. Much effort has focused on analyzing emotions from a single modality such as text data [8, 24, 48, 52], and to a much less extent, videos [40] and audios [9]. While their visualization approaches demonstrate success in analyzing the corresponding emotion modality, it is difficult to integrate them for multimodal analysis due to their different time granularities and dynamic variation. In addition, existing systems for multimodal emotion analysis [16, 51] only encode overall statistics, providing scant support for in-depth analysis, such as identifying dynamic changes of emotion coherence and inferring the underlying emotion states (e.g., deadpan) from videos. Moreover, those systems do not account for different levels of details, which may result in overlooking important emotion patterns. In summary, due to the multi-modality and varying granularity of emotional behavior in videos, it is challenging to conduct simultaneous emotion analysis across different modalities and explore the emotion coherence.

To address the above challenges, we work closely with two professional presentation coaches to propose novel and effective visual analytics techniques for analyzing multimodal emotions in presentation videos. Following a user-centered design process, we derive a set of visualization tasks based on the interviews and discussions with our two experts. As a result, we develop *EmoCo*, an interactive visualization system (Fig. 1) to analyze emotion states and coherence derived from face, text, and audio modalities at three levels of details. The channel coherence view summarizes the coherence statistics, and the sentence clustering view provides an overview of dynamic emotion changes at a sentence level. Once sentences of interest are selected, the detail view enables exploration of emotion states and their temporal variations along with supplementary information, such as voice pitches. Rich interactions are provided to facilitate browsing the videos and inferring the emotion states of the speaker. Two usage scenarios with TED Talk videos and expert interviews demonstrate the effectiveness and usefulness of our approach.

In summary, the primary contributions of this paper are as follows:

- We design and implement a prototype system to help users explore and compare the emotion coherence of multiple channels, including the emotions from facial expressions, text and audio, with multiple levels of details.
- We propose novel designs to facilitate easy exploration of emotion coherence, such as the channel coherence view with an augmented Sankey diagram design to support quick exploration of detailed emotion information distribution in each channel, and the sentence clustering view based on clustering to help track the temporal evolution.
- We present two usage scenarios and conduct semi-structured expert interviews to demonstrate the effectiveness of *EmoCo*.

2 RELATED WORK

This section presents three relevant topics, namely, emotion modalities, emotion visualization, and multimedia visual analytics.

2.1 Emotion Modalities

One central tenet of emotion theories is that emotional expressions involve different modalities, such as facial and vocal behavior [10]. Within this framework, emotion coherence among those channels plays an important role in human communication. Many psychological experiments [23, 42, 46] have demonstrated the hindering effect of incoherent expressions on emotion perception and recognition by others. Correspondingly, focusing on more modalities than the basic facial expressions alone can enable the discovery of underlying emotion states [4]. Despite such promising benefits, recent psychological research debates that the coherence across emotion modalities is not necessarily high and surprisingly weak for certain types of emotions [11, 33]. For example, Reisenzein el al. [33] found that facial expressions might not co-occur with experienced surprise and disgust. These ongoing debates have

motivated our research in multimodal emotion analysis. Specifically, our work looks at how to analyze emotions and their coherence derived from face, text, and audio modalities.

In line with the psychological experimental research, research in affective computing has evolved from the traditional uni-modal perspective to a more complex multimodal perspective [29, 41]. A great amount of work [30, 38] has focused on utilizing multimodal features to enhance emotion recognition. This line of work has examined different combinations of feature modalities and has identified those that do not contribute to recognition performance. Some work [32, 43] has employed deep architectures to capture complex relationships among multimodal features; however, they do not explicitly account for their coherence and thus are insufficient in detailed exploration. Our work is different from theirs in two aspects. First, we do not assume that emotions must be coherent on different behavioral modalities, as recent evidence from psychological research claims. Instead, we adopt well-established methods to extract emotions from different modalities and explicitly examine their coherence. Second, we use visual analytics to bring in human expertise in interpreting and analyzing the true emotions from videos, which thus provides a more detailed analysis.

2.2 Emotion Visualization

Emotion visualization has become a prominent topic of research over the last decade. Most effort has focused on analyzing emotions extracted from text data, such as documents [13], social media posts [17, 52] and online reviews [8, 24, 48]. For instance, Zhao et al. [52] analyzed personal emotion styles by extracting and visualizing emotion information. On a larger scale, Kempter et al. [17] proposed Emotion-Watch which summarizes and visualizes public emotional reactions. Less research has addressed facial and audio emotional expressions, which often involve more rapid evolution and smaller time granularity. Tam et al. [40] utilized parallel coordinates to explore facial expressions in measurement space, supporting the design of analytical algorithms. However, these systems have mainly centered on uni-modal emotions without considering information from other modalities. Our visualization approach integrates and tailors those visualization approaches to the varying granularity.

A few systems have been proposed to assist in emotion analysis from a multimodal perspective. Zadeh et al. [51] utilized histograms to visualize the relationships of sentiment intensity between visual gestures and spoken words in videos. More recently, Hu et al. [16] inferred latent emotion states from text and images of social media posts and visualized their correlations. Nevertheless, their visual approaches only encode overall statistics, lacking an in-depth analysis and visualizations of emotion states and changes at different levels of detail. Differing from them, we propose an interactive visual system to help multi-dimensional and multimodal emotion analysis, which few previous systems have addressed.

2.3 Multimedia Visual Analytics

A number of visual analytics systems have been proposed to assist in video analysis and knowledge discovery. One major challenge is the granularity level, which varies from the video and clip level to the word and frame level. On the one hand, many systems summarize the video content into temporal variables and represent them by line-based [15, 22] or tabular charts [14, 28] to enable analysis at the video level. Hierarchical brushing is often introduced to extend the analysis to finer levels [19]. While these systems support an effective visual exploration of the temporal overview and dynamics, they provide scant support for performing analytics tasks such as cluster analysis. On the other hand, some approaches [21, 34] discard the temporal information and consider short clips or frames to be the basic analysis units. For instance, Renoust et al. [34] unitized a graph design to visualize the concurrence of people within videos. Our method combines the above-mentioned approaches to support visual analytics at different levels of detail. Moreover, we designed a novel sentence clustering view to track the temporal evolution. Modalities pose another challenge on multimedia mining [44, 45]. Exploring the synergy among modalities can reveal higher-level semantic information [7]. While many computational methods have been proposed to support cross-modal analysis,

little research has specifically looked at visualization approaches. Stein et al. [39] proposed a computer vision-based approach to map data abstraction onto the frames of soccer videos. Xie et al. [49] suggested a co-embedding method to project images and associated semantic keywords on a 2D space. Wu and Qu [47] proposed a visual analytics system to explore the concurrence of events in visual and linguistic modalities. However, these systems only capture implicit or simple relationships among different modalities, so they are insufficient to promote in-depth analysis. To address those issues, we propose an *enhanced-Sankey view* to explicitly visualize the coherence across three emotion modalities and offer quick exploration of each modality.

3 DATA AND ANALYTICAL TASKS

In this section, we first describe the data processing procedures and the derived output. Next, we summarize the tasks based on a user-centered design process with two professional presentation coaches.

3.1 Data Processing

We conduct a series of data processing steps to extract emotion information from face, text, and audio modalities. We first apply well-established methods to extract information from each modality independently. Next, we fuse those data together based on their semantic meanings and align them at different levels of time granularity.

Facial Feature Extraction: The Microsoft Azure Face API [1] is employed to perform face detection, grouping, authentication and emotion recognition because of its good performance [18]. We experimentally consider the preponderant face group to be the speaker, and we greedily merge it with other face groups upon facial authentication, because the same speaker might fall into several groups. The output includes a set of emotions (i.e., anger, disgust, fear, happiness, sadness, surprise, contempt and neutral) with confidence values for the speaker in each video frame.

Text Feature Extraction: We adopt the predefined, human-labelled text segments from the TED Talk website as the data input, because each one forms a small semantic unit containing a few sentences with similar emotions. The official evaluation of IBM Watson Tone Analyzer service [2] indicates that it performs well on text emotion analysis [2]. Therefore, we use the Tone Analyzer API to extract emotion tones, including anger, disgust, fear, happiness, sadness, and analytical (neutral). We mark the last tone as "neutral" for consistency.

Audio Feature Extraction: The audios are first segmented in line with the aforementioned transcript segmentation. We use the neural network [35] to filter out audio clips containing laughter, because we observe that they severely affect the emotion recognition results. Next, we compute the Mel Frequency Cepstral Coefficient (MFCC), a feature usually used for audio emotion recognition, from extracted clips. After that, we feed this feature into a baseline model [6], which achieves a 96% accuracy in speech emotion recognition for our testing on the RAVDESS dataset [20]. Finally, there are seven detected emotions, including anger, disgust, fear, happiness, sadness, surprise and neutral.

Multi-modal and Multi-level Fusion: We fuse the extracted emotion data based on their categories and time granularity. For the multi-modal emotion categories, since different emotion recognition models are used for each channel and their emotion categories can be different, we use the union of all the possible categories in each modality, thus resulting in eight emotions in total (i.e., anger, disgust, fear, happiness, sadness, surprise, contempt and neutral). For multi-level fusion, we consider three levels of time granularity (i.e., the sentence level, the word level, and the frame level) based on advice from our two domain experts. In the previous steps, text and audio emotions have already been aligned at the sentence level, while facial emotions have been extracted frame by frame. To conduct sentence-level fusion, we calculate the most frequent facial emotion in each sentence to represent its predominant emotion. For word-level alignment, since the starting and ending times of each word have been detected by using the IBM

---

[1] https://azure.microsoft.com/en-us/services/cognitive-services/face/
[2] https://www.ibm.com/watson/services/tone-analyzer/

Watson Tone Analyzer API, we can easily map the facial, text, and audio emotions to each word based on its detected time period.

3.2 Data Description

We collect 30 TED Talk videos[3] to explore emotion coherence of presentation videos. Each video is about 10 minutes long and of high quality, with more than one million online reviews.

After data processing, each TED Talk is described by: 1) the original video and transcript; 2) facial emotions per frame; 3) text and audio emotions per transcript segment; 4) aligned emotions of face, text, and audio modalities per sentence, per word, and per frame. Emotions of each channel are associated with the confidence values output by corresponding models, and further summarized by the preponderant emotion with the highest confidence.

3.3 Task Analysis

Following a user-centered design process, we worked closely with two coaches, denoted as E1 and E2, from a presentation training company for about four months. Both coaches have more than five years of experience in presentation training. Their current coaching practice is grounded on videotaping presentations to analyze and provide feedback on the performance, which is tedious and time-consuming. Therefore, we iteratively developed and refined our system to assist them with the video analysis based on their feedback. Here, we summarize the distilled visualization tasks according to the granularity level as follows:

Video level exploration aims to summarize the emotions of each video, and provide video context for detailed exploration:

T1 To summarize emotion information in a video. It is necessary to summarize emotion information to offer an overview of the entire video collection, which helps users identify videos of interest and thereby guide effective exploration. The emotion information should include the emotion states of each modality and their coherence to represent the overall pattern.

T2 To provide video context for the analysis. Our two domain experts suggest that it is still essential to browse original videos for contextualized exploration in addition to summarized information. Due to the complexity of the data, visualizations should support rapid playback and guided navigation of videos in a screen-space-effective and responsive manner.

Sentence level exploration focuses on summarizing emotion coherence of sentences, as well as the detailed information of each sentence:

T3 To summarize emotion coherence across different modalities per sentence. Sentences in each transcript segment form a basic semantic unit with the same text emotions in our model. Presenting their coherence with facial and audio emotions is therefore a vital prerequisite for understanding the emotional expressions in presentations. For instance, do speakers' facial expressions react in conformity with a happy message such as jokes?

T4 To support rapid location of sentences of interest. Our experts are interested in examining how certain emotions are expressed, which demands the ability to rapidly locate sentences with emotions of interest. In addition, they wish to search for sentences with similar emotion expressions in order to comprehend the effects of such behavior on the overall situation.

T5 To display emotion information along with additional features for explanation. Our experts suggest to offer additional information, such as the face images, keywords, and prosodic features to verify and better understand the emotion expressions. This information should be displayed with the emotion information to guide the exploration.

T6 To show the temporal distribution of emotion states and their coherence. The temporal distribution of emotion states and their coherence represents the most detailed and fundamental characteristics. This information should be presented in detail and responsively due to its large scale.

---

[3] https://www.ted.com/talks

Word/frame level exploration shows the emotion of each word/frame, and can reveal changes in how speakers convey their emotions:

T7 To enable the inspection of details of emotion expressions at the word level. At a more detailed level, the experts want to explore whether the emotion expressions are associated with words. For instance, are certain kinds of words likely to be accompanied by changes in facial expressions?

T8 To reveal transition points of emotion behavior. Our experts are interested in exploring transition between emotion states, because they hope to discover interesting patterns. Therefore, it is important to algorithmically extract transition points and suppress irrelevant details to facilitate a focused analysis.

4 SYSTEM OVERVIEW

In this section, we first describe the analytical pipeline, and then introduce each view of the system.
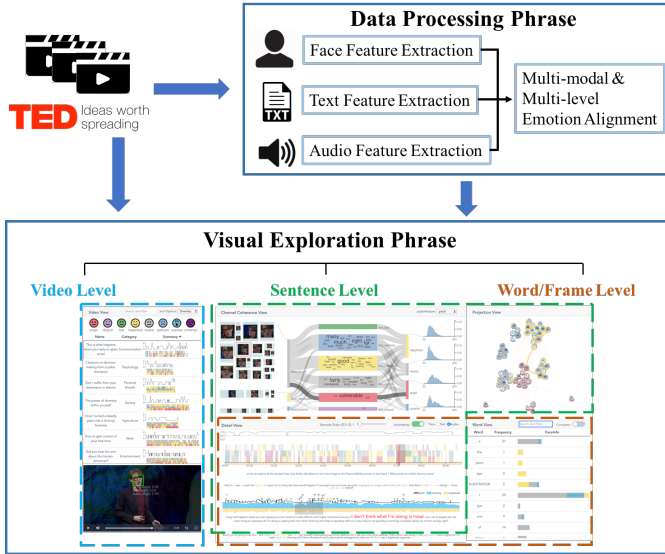

Fig. 2: Our visualization system pipeline for multimodal emotion analysis of presentation videos. In the data processing phase, we utilize well-established methods to extract emotion information from different channels. In the visual exploration phase, five coordinated views are provided to support three-level exploration.

As illustrated in Fig. 2, our system starts from the data processing phase. After the raw video data is collected, some well-established methods are used to extract emotion information from the face, text and audio channels. This extracted data is stored in MongoDB to facilitate smooth exploration. The data processing phase is explained in detail in Section 3.1. In the visual exploration phrase, users can perform three-level exploration with our visualization system. At the video level, users can enjoy a basic overview of each video and select a video of interest for further exploration. Afterward, a summary of emotion coherence based on sentences is provided to help users further explore sentences of interest. Users can then explore some keywords and transition points to further understand the sentences of interest.

Our system has five views (Fig. 1). The video view (Fig. 1a) presents a list of videos that provides a quick overview of the emotion status of the three channels of each video (T1). Users can easily select a video of interest based on their observation for further exploration. The video view presents the selected video at the bottom to help users directly observe the original information about the video (T2). The channel coherence view (Fig. 1b) presents the emotion coherence information of the three channels by using an augmented Sankey diagram design (T3-4). Some corresponding features extracted from different channels are embedded into this view to give some hints on different channels for explanation (T5). The detail view (Fig. 1c) presents detailed information of a selected sentence and its contexts to help users analyze a specific sentence (T6-8). The sentence clustering view (Fig. 1d) reveals the temporal distribution of emotion similarity across three channels at
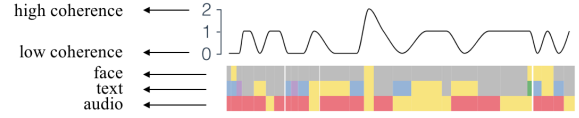

Fig. 3: A design for summarizing the emotion information of the three channels of a video. The line at the top explicitly shows the emotion coherence of the three channels. The bar code chart at the bottom shows more details about the exact emotions of each channel.

the sentence level (T6). The word view (Fig. 1e) provides the frequency of each word in the video transcript and allows users to compare different words with the face information and locate specific words in the sentences of a selected video (T7). Some smooth interactions are also provided in the system. For more details, please refer to Section 5.6.

5 VISUALIZATION DESIGN

Based on the analytical tasks mentioned in Section 3.3, we further summarize a set of design rationales with our collaborators to better design our system, which is shown as follows:

Multi-level Visual Exploration. The mantra "Overview first, zoom and filter, then details on demand" [37] has been widely used in exploring complex data. Thus, to explore the data extracted from videos, we follow this mantra in designing our system. First, we provide the summary information of the video collection to provide users with some hints that can help them identify a video of interest. After selecting a video, users can further explore the emotion coherence at the sentence level. After selecting a sentence of interest, users can drill down to the word/frame level.

Multi-perspective Joint Analysis. To facilitate a detailed analysis of emotion coherence from the three channels in videos, various types of information should be provided. For a better interpretation, the features from these channels are extracted and embedded into the corresponding views. Multiple linked views that show different data perspectives are integrated into our proposed system, and users can combine these views to achieve a multi-perspective joint analysis.

Interactive Pattern Unfolding. Given that the analysis of emotion coherence in presentation videos contains much hidden knowledge, users need to go through a trial-and-error process. Thus, it is helpful for users to interact with the data directly, so they can observe and interpret the results based on knowledge.

The top of Fig. 1a shows the unified color encoding we adopted. For the five common emotion categories (i.e., anger, disgust, fear, happiness, sadness), we mainly borrow colors from Plutchiks emotional wheel [27] and further carefully design the color mapping for the remaining emotions, where user feedback is also considered.

5.1 Video View

Description: As shown in Fig. 1a, the video view is divided into three parts. The top part of this view is the legend, which presents our adopted color scheme that helps users know which color is associated with each emotion included in our system. The middle part presents a list of the video information. There are three columns, namely, name, category and summary, and each row provides these three types of information for each video. The first two columns, which indicate the names of videos and their corresponding categories, are easily understandable. The summary column uses a line and a bar code chart to show the coherence of the information of the three channels (Fig. 3), which provides users with a quick overview that helps them in selecting a video of interest (T1). The line is used to explicitly show the degree of emotion coherence among the three channels. A higher value of the line corresponds to a higher coherence, as shown in Equation 1. Specifically, "2" means that emotions of the three channels are all the same, while "0" means that emotions of these channels are all different.

$$D_{coherence} = \begin{cases} 2, & E_{face} = E_{text} \ \& \ E_{text} = E_{audio}, \\ 0, & E_{face} \neq E_{text} \ \& \ E_{text} \neq E_{audio} \ \& \ E_{face} \neq E_{audio}, \\ 1, & others \end{cases}$$
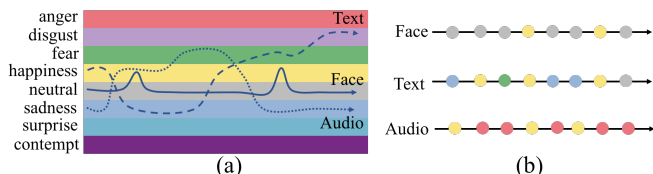
(1)

Fig. 4: Two alternative designs for summarizing emotion information from the three channels. (a) Three different curves represent the three channels. The position of each curve indicates the current emotion, as the background has eight different-colored emotion bands. (b) Three straight lines represent three channels. Each color dot represents an emotion at one moment.
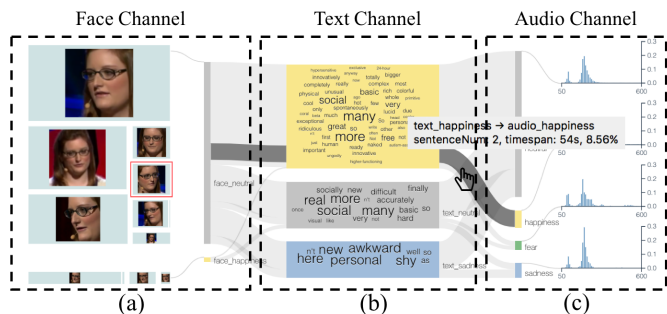

Fig. 5: An augmented Sankey diagram design for summarizing emotion coherence from three channels as well as for providing extracted features for explanations. Each node represents one type of emotion and each link represents a collection of sentences with certain emotions shared by two channels, either the face and text channels or the text and audio channels. (a) A treemap-based design to show a quick overview of representative detected faces in the video. (b) A word cloud design to highlight some important words, which gives users some hints about the corresponding context. (c) A histogram design to show the audio feature distribution for different emotions.

where $D_{coherence}$ indicates the degree of coherence, and E_{face}, E_{text} and E_{audio} indicate the emotion types in the corresponding channels.

We also include a bar code chart to show the emotion information of the three channels, with the X-axis representing the length of the video and the Y-axis representing the permutation of the face, text, and audio channels. The color of each rectangle represents the emotions in the three channels. Users can search or filter the video list by typing some keywords in the search function. They can also sort these videos based on specific criteria, such as the coherence, diversity and percentage of one type of emotion. After users click on a row, the video of interest will be selected.

After a video of interest is selected, the original video is presented in the bottom part of the video view (Fig. 1a) to allow users to explore detailed information (T2). Although the extracted information from the video is informative, referring to the original video can sometimes provide a better explanation. In this view, users are allowed to play the video at a slow, normal, or fast speed. After a video is paused, the detected faces are highlighted by green rectangles and the detected emotion information from the three channels will be shown at the same time. When exploring other views, users can easily seek to the corresponding frames by using some provided interactions (Section 5.6).

Justification: Our end users emphasized the need for a quick summary of each video. Originally, we used a scatter plot to visualize a list of videos. Each dot in this scatter plot represented a video, and clusters represented similar videos. However, our end users would like to have more information, such as a quick summary of each video. Thus, they preferred to use a list to show video information. To better show the emotion coherence overview of each video, we considered some alternative designs (Fig. 4). In Fig. 4a, the eight different-colored bands in the background represent different emotion categories. The emotion information of each channel was represented by three different curves, allowed us to see how emotions change in each channel. However, this kind of design had obvious visual clutter, so we came up with another design (Fig. 4b). We used a straight line to represent each channel, with each color dot indicating an emotion at a specific moment. This design allowed users to easily observe how emotions change on different channels, but it did not make efficient use of space. Therefore, we decided to use a more compact design: a three-row bar code chart. Furthermore, our end users suggested we add a line chart to explicitly show the coherence information and its dynamic trend. The final design is shown in Fig. 3.

5.2 Channel Coherence View

Description: To show the connection between the three channels in the selected video, as well as showing some features extracted from the corresponding channels, we come up with an augmented Sankey diagram design. As shown in Fig. 5, this view contains three parts, namely the face channel (Fig. 5a) on the left-hand side, the text channel (Fig. 5b) in the center and the audio channel (Fig. 5c) on the right-hand side. First, we adopt a Sankey diagram design [36] to visualize the connection among the face, text, and audio channels (T3). The emotion information is detected based on each sentence in the videos. In this way, each node in the Sankey diagram represents one type of emotions, and each link represents a collection of sentences with the emotions between two channels, either face and text channels or text and audio channels. The

height of each link represents the total duration of the corresponding sentences. Hence, these links can give users some relevant information on how a speaker conveys his emotions from different channels when he utters these sentences. For example, a link from the left-hand neutral node to the middle happiness node shows that the speaker is talking about something happy while keeping a neutral face, while a link from the middle sadness node to the right-hand neutral node indicates that the speaker is saying something sad in a neutral voice. We add a hover interaction feature to better illustrate the connection among these channels. In Fig. 5, when users hover over a link between the middle and right-hand nodes, the corresponding link between the left and middle nodes will also be highlighted, thereby highlighting the emotion connection between the three channels.

To provide more information from these channels, we embed features from these channels into the Sankey diagram (T5). For each node (face emotion) in the face channel, we adopt a treemap-based design to present an overview of the detected faces, since the data structure is intrinsically hierarchical (i.e., each node contains several links, each link contains several sentences and each sentence contains many face images). Each rectangle in treemap represents a cluster (a link), whereas the size of the rectangle represents the number of faces in a specific cluster. As shown in Fig. 5, the corresponding rectangle area of the link (neutral face, happiness text, and neutral audio) is highlighted. Then, we overlay a representative image on each rectangle. Currently, the representative image for each cluster means the image nearest the center point of the cluster. Other strategies can be easily adopted. For text information, we embed a word cloud into the middle nodes. After considering their frequency and sentiment, we calculate the importance of each word. Thus, a word cloud is used to show important words in the corresponding sentences and provide users with some context. For audio information, we use histograms to visualize the average distribution of corresponding sentences. Users can configure different audio features, including pitch, intensity, and amplitude, and then formulate the corresponding histograms.

Justification: To better illustrate the emotion coherence information of different channels, we considered some alternative designs (Fig. 6). At first, we came up with a chord diagram-based design (Fig. 6a), where each channel is represented by an arc, and the links between different arcs represent their connections. Using this design, we could observe emotion coherence information of different channels. However, this kind of design had serious visual clutter in the middle and was not space efficient for embedding extracted features. Therefore, we considered a Sankey diagram design (Fig. 6b). When we presented our prototype to our end users, they commented that it would better to add more information. They also faced difficulty in understanding the connections between different channels. To address this problem, we developed an augmented Sankey diagram design with some interactions, which was favorably received by our end users.
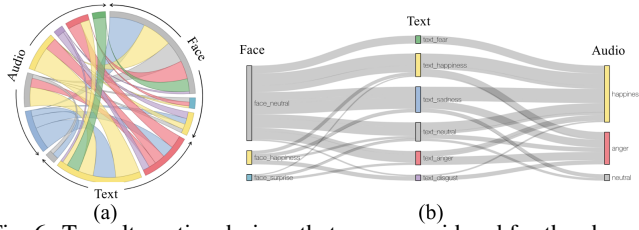
5

Fig. 6: Two alternative designs that were considered for the channel coherence view. (a) A chord diagram design for showing emotion connection of different channels. Each channel is represented by an arc, and each connection is represented by a link. (b) A normal Sankey diagram design for showing emotion connection between different channels. Emotions in each channel are represented by nodes; their connections are represented by links.

5.3 Detail View

Description: As shown in Fig. 1c, the detail view consists of two main parts. The bar code chart at the top shows face emotions at frame level and text and audio emotions at sentence level, which provides a more detailed summary than the corresponding bar codes in the video view (Fig. 1a). Users are allowed to adjust the scale of the bar code and scroll along the timeline to explore the details. Due to the imperfect accuracy of emotion recognition, the confidence scores of each channel are encoded as a line chart to give hints about the potential inaccuracy. Also, users can interactively choose to turn on/off the hint or switch to different channels. Once users have selected a node or a link in the channel coherence view (Fig. 1b), those selected sentences will be highlighted in the bar code chart. Then users are allowed to select a sentence of interest for further exploration. The corresponding sentence context will be shown at the bottom part of the detail view. Specifically, the sentence being explored is shown in the middle, and the two previous sentences and two following sentences are also shown to provide more context. Three audio features for the selected sentence, i.e., pitch, intensity, and amplitude, are explicitly visualized as a line chart and a theme river, which reveals temporal changes of audio features for the selected sentence. When users brush on part of the sentence, corresponding words will be highlighted. Furthermore, to better visualize the changes of face emotions, we use two inverted right triangles to represent each transition point. The left one represents the emotion before the change, the right one represents the emotion after the change. To avoid visual clutter, dashed lines are used to indicate the location of the transition. Additionally, when transitions happen, corresponding words are also highlighted with colors according to the changes of the facial emotions.

5.4 Sentence Clustering View

To explore how a speaker changes his strategy in conveying emotion over time on different channels, we need to visualize the temporal distribution of the emotion coherence information of different channels. As shown in Fig. 7b, inspired by the time-curved design [5], we project the emotion information of each sentence as a glyph point on a 2D plane by using the t-SNE projection algorithm, where the vector is constructed as Equation 2. Points are linked with curves by following time order. To show the information of each sentence more clearly, we design a pie chart-based glyph. Three equally divided sectors of a circle are used to encode emotion information of the face, text and audio channels. To be specific, the top left shows text emotion, the top right shows face emotion, and the bottom shows audio emotion. Color is used to represent the type of emotion, and radius is used to represent the emotion probability (certainty). The larger the radius, the higher the emotion probability. To show temporal information of these sentences, both color and sentence ID in the middle of a glyph are used to represent time order. A lighter color means an earlier time, while a darker color means a later time.

$$Vector = [Pr(E_{face}), Pr(E_{text}), Pr(E_{audio})] \qquad (2)$$

where $Pr(\cdot)$ indicates the detection probability for each emotion in the corresponding emotion category and $E_{(\cdot)}$ indicates one type of emotion in different channels.
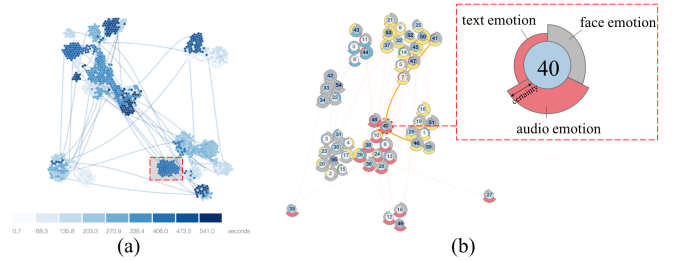

Fig. 7: Visual designs for the sentence clustering view. (a) A frame-based projection without glyph design. (b) A sentence-based projection with glyph design showing emotion information of the three channels, as well as time information.

Justification: Originally, we used a frame-based projection, which projects emotion information from a frame level. As shown in Fig. 7a, there were too many points and not enough clear information. By observing that those clusters (e.g., the one selected with a red dashed box) of points are almost from the same sentence, our end users commented that there was no need to drill into the frame level in this view. It would be better to explore at the sentence level. Then we considered sentence-based projection, which has a better visual effect. As shown in Fig. 7b, based on our end users' suggestions, we further embedded glyphs to show both emotion and time information. Our end users were satisfied with this design.

5.5 Word View

Our end users expressed that they would like to further conduct word-level exploration, especially the frequency of the words used and corresponding emotions when uttering these words. In this view (Fig. 8c), we provide detailed information for each word used in the video. Three attributes are shown, namely word, frequency and face information. For each row, the word column directly shows the word used in the video; the frequency column indicates how many times each word is used in the video; and the face information column visualizes the duration of saying this word and the emotion percentage of face emotion by using a stacked bar chart. The length of each component in a stacked bar chart indicates the duration of the expressed type of emotion. For those faces do not be detected, we use dashed areas to represent them (Fig. 8c). For focusing on detected emotions, users are allowed to hide these dashed areas by turning off the switch button. Furthermore, users are allowed to sort the word view by specific criteria, such as frequency, as well as by using a keyword search.

5.6 Interactions

Our system *EmoCo* supports various interactions, empowering users with strong visual analytical abilities. The five views provided in the system are linked together. Here, we summarize the interactions adopted in our proposed system.

Clicking Once users click a video of interest in the video view, the video will be selected and other views will be updated accordingly. In the channel coherence view, when users click nodes or links of interest, corresponding sentences will be selected and highlighted in the detail view. In the detail view, users can click specific sentences to explore their context information. Similarly, users can click a word in the word view to highlight sentences in the detail view. Furthermore, users are allowed to click the timeline in the detail view for seeking corresponding places in the video.

Brushing When users brush the bar code in the detail view to select corresponding sentences, then corresponding sentences will be highlighted in the sentence clustering view. Conversely, when users brush some points in the sentence clustering view, the corresponding sentences will be highlighted in the bar code in the detail view. In addition, once users select a sentence, they are allowed to brush an area of the selected sentence and identify its words.

Searching and Sorting In the video view and detail view, to allow users to quickly discover a row of interest, we add searching and sorting interactions. Users are allowed to search by some keywords and sort the list by a specific criterion.

6 USAGE SCENARIO

In this section, we describe two usage scenarios to demonstrate the effectiveness and usefulness of *EmoCo* to accomplish the visualization tasks in Section 3.3 and discover insights.

6.1 How to be emotional

In this scenario, we describe how Kevin, a professional presentation coach, can find examples for teaching his students to express emotions more effectively. His teaching is based on the book *Talk like TED* by the keynote speaker Carmine Gallo [12], where the author attributes the best presentations to be emotional. To strengthen his teaching, Kevin would like to find more examples with considerable emotional expressions. However, it is time-consuming to browse the large video collections and identify representative clips. Therefore, he refers to *EmoCo* to explore the videos and find evidence for teaching.

After loading the video collection, Kevin directly notes the video list in the video view (Fig. 1a). He wishes to find the video with the most emotions. Thus, he sorts the videos by diversity of emotions, whereby the video entitled *This is what happens when you reply to spam email* appears at the top. He observes many colors in the corresponding bar code chart, which denotes that this presentation contains diverse emotions (T1). He also notes the frequent fluctuation of its line chart, which indicates that the speaker's emotion coherence varies a lot. As such, he considers this presentation to be representative of good emotional expressions, so he clicks it for further exploration.

To understand overall emotional expressions (T3), he shifts attention to the Sankey diagram in the channel coherence view (Fig. 1b). He immediately notices that the three Sankey bars have very different color distributions, and the Sankey links between the same color account for only a small portion of widths. Those suggest that the emotional expressions are incoherent across each modality.

He decides to explore each modality for detailed understanding. He starts with the leftmost Sankey bar set and finds the predominant grey color, which indicates the most neutral facial expressions. Similarly, he observes a few happy and surprised facial expressions. Following the face thumbnails to the left, he finds that the speaker has rich facial movements (T5). For example, the speaker tends to raise the corner of his mouth with happy facial expressions, while his mouth tends to open with surprise. As such, Kevin deems facial recognition reliable. In contrast to the leftmost bar set, Kevin observes more emotions, including fear, neutral, happiness, anger, sadness, and disgust, from other two bar sets. He then inspects the histogram to its right, where he finds that anger and surprise tend to yield higher pitches. He considers these results to be reasonable based on his experience.

Next, Kevin decides to inspect detailed expressions with anger, an unusual emotion in presentations. By examining and comparing Sankey links passing red nodes (anger), he identifies the largest link, which connects anger in text and audio modalities with neutral facial expressions. Upon clicking that link, one corresponding sentence is highlighted in the bar code view (Fig. 1c). He selects the sentence to unfold its details. Following the line chart in the middle, Kevin notices fluctuations of the black line and many glyphs, which denote rapid evolution of voice pitches and facial expressions (T8). By browsing the video clip, Kevin understands that the speaker expresses an angry message that replying to scam emails is not mean (T4). He emotes and performs theatrical facial and audio expressions, which render his presentation engaging. Next, he returns back to the bar code view to analyze its context. He notes that both the previous and next sentences have different emotions from the current sentence. Kevin is quite curious. How can the speaker convey various emotions within such a short period?

He observes a gap between those two sentences, and further finds that the bar code tends to be discontinuous (Fig. 1c). Similarly, he notices large distances between two consecutive sentences in the sentence clustering view (Fig. 7b), which indicates rapid changes of the emotions (T6). Interestingly, he finds that the facial modality behaves quite differently from the other two. Facial information usually does not accompany with text and audio information, and vice versa. To find out what happens in the video, Kevin quickly navigates to those discontinuous parts in the bar code (T2). Finally, he finds that the ab-sence of text and audio information is likely the speaker's presentation style. The speaker usually pauses for a while to wait for the audience's reaction, which is a kind of audience interaction strategy.

Overall, Kevin considers this video to demonstrate the emoting presentation style, which is a good example for his teaching. The speaker adopts a rich set of emotions and tends to be incoherent in a theatrical manner, which renders his presentation infectious and engaging. Upon changing emotions, the speaker often pauses for a while for the audience to react and interact with him.

6.2 How to tell jokes

In this scenario, Edward, another presentation coach, wants to teach a student to incorporate humor in presentations. As the student mainly adopts neutral facial expressions, Edward would like to find examples where joke-telling is accompanied by neutral facial expressions to promote personalized learning.

After loading the video collection, Edward sorts them by the percentage of neutral emotions in descending order. By comparing the bar codes in the *Summary* column (Fig. 8a), he finds that the video named "How I learned to communicate my inner life with Asperger's" contains preponderant yellow grids in the middle row, which implies predominant happy emotions in the text modality. Thus, he feels interested in this video and clicks it to inspect details in other views.

From the channel coherence view (Fig. 5), he first observes few emotions in each channel where the neutral expressions predominate. As highlighted in the darker Sankey link in Fig. 5, the speaker tends to deliver happy messages with neutral facial and audio emotions. Since Edward wants to find examples of telling jokes, he clicks on the Sankey link between happy text and audio emotions. Corresponding sentences (Glyph 1 and 27) are highlighted in the sentence clustering view (Fig. 8b). To find out other sentences with a similar way of emotion expression, he simply brushes the nearby area of the highlighted glyphs (T6) to locate them in the bar code (Fig. 8d). He then would like to explore how the speaker delivers those happy messages in detail.

For further exploration, he clicks some of these sentences, and then observes the context or seeks back to the original video (T2). After examining these sentences, he finds that the speaker indeed tells some jokes with a certain presentation style. For example, for the first sentence in the bar code (Fig. 8d-1); the content is shown below. The pitch line tends to be flat (T5), and there are almost no face transition points (T8), which indicates that the speaker does not have many audio changes or face changes when saying this sentence. The speaker in this sentence tells the audience that she is a very visual thinker and not good at language, just like a beta version of Google Translate. After hearing this, the audience laugh. The speaker smiles at the end. As for the second sentence (Fig. 8d-2), the speaker tells the audience that she refused to shower due to hypersensitivity and now she assures her hygiene routine is up to standards. At that moment, the audience laugh and she smiles again. As for the third sentence (Fig. 8d-3), the speaker tells the audience that she loves lucid dreaming, because she can do whatever she wants. Then the speaker throws an example, "I'm making out with Brad Pitt and Angelina is totally cool with it." The audience feel it is very funny and laugh. The speaker also grins. Finally, Edward realizes that this is her presentation style in this video. The speaker tells something funny or ridiculous without showing too many emotions. In addition, Edward wants to check the words that the speaker uses in the video, which may give him more hints (T7). So he directly sorts the words by frequency in the word view (Fig. 8c). He finds that most of the words the speaker uses are some general words, such as "you", "this", and "have". Interestingly, he finds that even when the speaker says the bad word "autism", her facial expressions are neutral, as shown in Fig. 8c, which corresponds to the previous findings. The speaker does not show too many emotions in the face and audio channels most of the time. From her facial expressions, the audience may feel that the presentation is dry. However, by combing emotion in the other two channels, she makes her presentation very interesting.

Overall, Edward thinks this is a good example for his student to learn from. He thinks the presentation style of this video is deadpan humor, a form of comedic delivery to contrast with the ridiculousness of the subject matter without expressing too many emotions.

Fig. 8: A deadpan humor presentation style. The speaker lacks emotions most of the time when saying something interesting. (a) Select a video with many neutral facial emotions and happy text content. (b) Click the link with happy text and audio emotions in the channel coherence view (Fig. 5), then two sentences represented by the Glyph 1 and Glyph 27 are highlighted. Further, brush an area to find out similar sentences with the highlighted sentences. (c) Sort the word view by frequency to find out which words the speaker tends to use. (d) Some sentences are highlighted with red dashed rectangles in the bar code due to the brush interaction in (b). In particular, we examine three example sentences with neutral facial emotions at first and happy emotions at the end. The detailed context of the first example sentence is shown.

7 EXPERT INTERVIEW

To further evaluate our system, we conducted semi-structured interviews with our aforementioned collaborating domain experts (E1, E2). Both E1 and E2 were familiar with basic visualization techniques, such as bar charts and curves. The interviews, which were separated in two sessions, were guided by the set of questions shown in Table 1. The experts were allowed to provide open-ended answers. Each interview session lasted about an hour. After we introduced them to the functions, visual encodings, and basic views of our system, the experts were allowed to freely explore our system in a think-aloud protocol for twenty minutes. In general, we received positive feedback from the experts towards *EmoCo*. Both experts appreciated the idea of leveraging visual analytics techniques to support an interactive exploration of presentation videos with novel encodings.

Interactive Visual Design. Both experts confirmed that the system is well designed and follows a three-level exploration hierarchy. The experts were satisfied with the system's views, especially the video, channel coherence, detail, and word views, sharing that "they are easy to follow and understand." As for the sentence clustering view, E1 commented, "It is much better than the original frame-based projection.

It is easy to understand, but it might not be easy for a novice to use it." E1 added, "Actually, it takes me some time to digest the information presented in the sentence clustering view, but after practicing more, I find it very useful for finding those similar moments in a presentation video." Both experts commented that they could easily find a video of interest by just navigating through the concise summary of each video presented in the video view. E1 shared, "The quick summary provides me with visual hints about the video style, which greatly help me find the video I am interested in." Meanwhile, E2 also mentioned that the interactions in the video view, including searching and sorting, were very useful when exploring the presentation videos. Both experts appreciated the ability of our system to identify sentences and words of interest. E2 commented that the current system provides detailed information which facilitates detecting abnormal emotion coherence and emotion changes. "Usually, I tend to pay attention to those unexpected sentences, such as saying something sad with happy emotion, and I would double check whether it is the real case or caused by some problematic emotion detection. These views are very helpful for me to do this checking." E1 was more interested in those emotion transition points. "These transition points usually indicate different contents in a talk. The word view shows the keywords in the context, allowing the

Table 1: Questions for user interviews.

#	Aim	Question
Q1	Visual Design	It is easy/hard to learn to read the video view? why?
Q2	Visual Design	Is it easy/hard to learn to read the channel coherence view? Why?
Q3	Visual Design	It is easy/hard to learn to read the detail view? why?
Q4	Visual Design	It is easy/hard to learn to read the sentence clustering view? why?
Q5	Visual Design	It is easy/hard to learn to read the word view? why?
Q6	Interaction Design	It is easy/hard to find a video of interest for further exploration? Why?
Q7	Interaction Design	It is easy/hard to identify sentences/words of interest? Why?
Q8	Interaction Design	It is easy/hard to find similar presentation styles in a video? Why?
Q9	General	Which part of the visual interface do you think can be further improved? How?
Q10	General	Do you think the system is informative for exploring presentation videos?

speaker to understand how to improve his presentation skills by using the appropriate words."

Applicability and Improvements. Both domain experts expressed their interest in applying *EmoCo* to deal with practical problems in their daily work. Previously, to help people improve their presentation skills, they would ask the speakers to conduct multiple practice talks and record them for later analysis. This process is time-consuming and cannot provide any quantitative analysis. E1 stated that "*EmoCo* is the first computer-aided visualization system for emotion analysis and presentation training that I have ever seen, and it can definitely help me analyze those presentation practice videos and train speakers with clear visual evidence." E2 especially appreciated the human-in-the-loop analysis by sharing that "the cooperative process with *EmoCo* is of great benefit in emotion analysis, as the system provides quantitative measures on the level of emotion coherence, and we can determine whether it makes sense or not." Our collaboration with the domain experts also suggested directions for future studies. First, E1 suggested that it would be beneficial to show the performance of the speaker in real-time with a score in the detail view. E2 pointed out that for facial expressions, he identified a few unusual emotions that may result from the lip movements or facial expressions when saying a specific word. Yet, this is mainly due to the accuracy performance of the underlying facial expression algorithms and is beyond the main contributions of our visual analytics system. He also mentioned that it might be helpful to understand the sentence clustering view by adding semantic labels to those clusters. In addition, the word view could be further improved by adding functions for analyzing weak words that are widely used in the start or end of presentations.

8 DISCUSSION

In this section, we identify limitations of our system and propose directions for future study.

Emotion Coherence. Since we use different methods to extract emotions from three modalities (i.e., face, text and audio) and their emotion categories vary, we choose the union of all possible emotion categories as the final emotion categories for our further calculation of emotion coherence. It works well for most videos, since the common categories (e.g., anger, happiness, sadness, and neutral) from different modalities are often the major types of emotions in the three modalities in most videos. But it can also bring some negative effects on the coherence calculation in certain situations. For example, contempt only exists in the face channel, so it will be always considered as incoherent with the other two modalities. Meanwhile, the current methods of extracting emotions can easily be replaced with other more advanced methods with the same emotion categories in the future. What's more, it is still an open question whether emotions should be kept coherent or not during a presentation. In most cases, emotion coherence should be determined based on the specific topic, the content delivered by the speaker and the adopted presentation style. Motivated by this observation, we designed *EmoCo* to help users understand a speaker's coherent or incoherent emotions in a presentation. However, our system still requires users' manual inspection on the presentation video data. In the future, we plan to provide quantitative measures of emotion coherence and integrate applicable advanced algorithms to facilitate this exploration process.

Emotion Recognition. We adopt well-established methods to extract emotions from different channels, which can achieve high accuracy. When working on this paper, we conducted preliminary evaluations of the emotion recognition accuracy of the three channels on the two videos used in Section 6. We tested 1000 sampled frames, 29 text segments, and the corresponding audio segments for the first video, and we tested 1000 sampled frames, 55 text segments and the corresponding audio segments for the second video. The accuracy of the first video on the face, text and audio channels was 85.4%, 79.3% and 89.6%, respectively, and the accuracy of the second video on the face, text and audio channels was 95.8%, 87.3% and 81.8%, respectively. Therefore, the emotion recognition methods we used are acceptable in our system. Meanwhile, we leveraged the confidence scores and designed a line chart in the detail view to give users hints about possible inaccuracy of the emotion recognition. In addition, the current emotion recognition algorithms can be easily replaced, when more advanced emotion recognition algorithms are available. Furthermore, our work currently considers only eight emotion categories, which may not always satisfy the requirements of different users and tasks. Therefore, we plan to explore more emotion categories to capture more subtle emotions of speakers during presentations.

Generalizability. Our system *EmoCo* is proposed for analyzing presentation videos. Two usage scenarios on TED talk videos with predefined text segments are used to demonstrate the effectiveness of our system. However, *EmoCo* is not limited to TED talk videos and can be easily extended to other presentation videos by adopting automatic transcription techniques [1, 3] to extract the text segments from the presentation videos.

9 CONCLUSION

In this paper, we propose *EmoCo*, an interactive visual analytics system to analyze emotion coherence across different behavioral modalities in presentation videos. Our system comprises of five linked views, which allow users to conduct in-depth exploration of emotions in three levels of detail (i.e., video, sentence and word levels). It integrates well-established visualization techniques and novel designs to support visual analysis of videos. In particular, we propose an augmented Sankey diagram design for analyzing emotion coherence and the clustering-based projection design for tracking the temporal evolution, facilitating the exploration of multimodal emotions and their relationships within a video. Two usage scenarios based on TED Talk videos and interviews with two domain experts demonstrate that our system can enable efficient and insightful analysis.

In the future, we plan to extend our system to support analysis of additional modalities, such as hand gestures. Moreover, we plan to incorporate advanced data mining techniques to enhance the analysis. In addition, with the capability of the proposed system in analyzing emotion coherence, it would also be interesting to further explore whether our system can be applied to detailed performance analysis of emotion recognition algorithms to further improve their accuracy.

REFERENCES

[1] Cloud speech–to–text. https://cloud.google.com/speech-to-text/. Accessed: 2019-06-18.

[2] Measuring the quality of the ibm waston tone analyzer api. https://cloud.ibm.com/docs/services/tone-analyzer?topic=tone-analyzer-ssbts#smqs. Accessed: 2019-06-18.

[3] Speechmatics. https://www.speechmatics.com/. Accessed: 2019-06-18.

[4] H. Aviezer, Y. Trope, and A. Todorov. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111):1225–1229, 2012.

[5] B. Bach, C. Shi, N. Heulot, T. Madhyastha, T. Grabowski, and P. Dragicevic. Time curves: Folding time to visualize patterns of temporal evolution in data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):559–568, 2016.

[6] P. Barros and S. Wermter. Developing crossmodal expression recognition based on a deep neural model. *Adaptive Behavior*, 24(5):373–396, 2016.

[7] C. A. Bhatt and M. S. Kankanhalli. Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications*, 51(1):35–76, 2011.

[8] C. Chen, F. Ibekwe-SanJuan, E. SanJuan, and C. Weaver. Visual analysis of conflicting opinions. In *Proceedings of the IEEE Symposium On Visual Analytics Science and Technology*, pages 59–66. IEEE, 2006.

[9] C.-H. Chen, M.-F. Weng, S.-K. Jeng, and Y.-Y. Chuang. Emotion-based music visualization using photos. In *Proceedings of the International Conference on Multimedia Modeling*, pages 358–368. Springer, 2008.

[10] C. Darwin and K. Lorenz. *The Expression of the Emotions in Man and Animals*. Phoenix Books. University of Chicago Press, 1965.

[11] J.-M. Fernández-Dols and C. Crivelli. Emotion and expression: Naturalistic studies. *Emotion Review*, 5(1):24–29, 2013.

[12] C. Gallo. *Talk like TED: the 9 public-speaking secrets of the world's top minds*. St. Martin's Press, 2014.

[13] M. L. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner. User-directed sentiment analysis: Visualizing the affective content of documents. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 23–30. Association for Computational Linguistics, 2006.

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

[15] M. Hoeferlin, B. Hoeferlin, G. Heidemann, and D. Weiskopf. Interactive schematic summaries for faceted exploration of surveillance video. *IEEE Transactions on Multimedia*, 15(4):908–920, 2013.

[16] A. Hu and S. Flaxman. Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 350–358. ACM, 2018.

[17] R. Kempter, V. Sintsova, C. Musat, and P. Pu. Emotionwatch: Visualizing fine-grained emotions in event-related tweets. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2014.

[18] S. R. Khanal, J. Barroso, N. Lopes, J. Sampaio, and V. Filipe. Performance analysis of microsoft's and google's emotion recognition api using pose-invariant faces. In *Proceedings of the International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*, pages 172–178. ACM, 2018.

[19] K. Kurzhals, M. John, F. Heimerl, P. Kuznecov, and D. Weiskopf. Visual movie analytics. *IEEE Transactions on Multimedia*, 18(11):2149–2160, 2016.

[20] S. R. Livingstone and F. A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35, 05 2018.

[21] J. Matejka, T. Grossman, and G. Fitzmaurice. Video lens: rapid playback and exploration of large video collections and associated metadata. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*, pages 541–550. ACM, 2014.

[22] A. H. Meghdadi and P. Irani. Interactive exploration of surveillance video through action shot summarization and trajectory visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2119–2128, 2013.

[23] V. I. Müller, U. Habel, B. Derntl, F. Schneider, K. Zilles, B. I. Turetsky, and S. B. Eickhoff. Incongruence effects in crossmodal emotional integration. *Neuroimage*, 54(3):2257–2266, 2011.

[24] D. Oelke, M. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L.-E. Haug, and H. Janetzko. Visual opinion analysis of customer feedback data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 187–194. IEEE, 2009.

[25] T. Pfister and P. Robinson. Speech emotion classification and public speaking skill assessment. In *Proceedings of International Workshop on Human Behavior Understanding*, pages 151–162. Springer, 2010.

[26] T. Pfister and P. Robinson. Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. *IEEE Transactions on Affective Computing*, 2(2):66–78, 2011.

[27] R. Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.

[28] D. Ponceleon and A. Dieberger. Hierarchical brushing in a collection of video data. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 8–pp. IEEE, 2001.

[29] S. Poria, E. Cambria, R. Bajpai, and A. Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[30] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang. Towards an intelligent framework for multimodal affective data analysis. *Neural Networks*, 63:104–116, 2015.

[31] V. Ramanarayanan, C. W. Leong, L. Chen, G. Feng, and D. Suendermann-Oeft. Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In *Proceedings of the ACM on International Conference on Multimodal Interaction*, pages 23–30. ACM, 2015.

[32] H. Ranganathan, S. Chakraborty, and S. Panchanathan. Multimodal emotion recognition using deep learning architectures. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1–9. IEEE, 2016.

[33] R. Reisenzein, M. Studtmann, and G. Horstmann. Coherence between emotion and facial expression: Evidence from laboratory experiments. *Emotion Review*, 5(1):16–23, 2013.

[34] B. Renoust, D.-D. Le, and S. Satoh. Visual analytics of political networks from face-tracking of news video. *IEEE Transactions on Multimedia*, 18(11):2184–2195, 2016.

[35] K. Ryokai, E. Durán López, N. Howell, J. Gillick, and D. Bamman. Capturing, representing, and interacting with laughter. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, page 358. ACM, 2018.

[36] M. Schmidt. The sankey diagram in energy and material flow management: Part i: History. *Journal of Industrial Ecology*, 12(1):82–94, 2008.

[37] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The Craft of Information Visualization*, pages 364–371. Elsevier, 2003.

[38] M. Soleymani, M. Pantic, and T. Pun. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3(2):211–223, 2012.

[39] M. Stein, H. Janetzko, A. Lamprecht, T. Breitkreutz, P. Zimmermann, B. Goldlücke, 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*, 24(1):13–22, 2018.

[40] G. K. Tam, H. Fang, A. J. Aubrey, P. W. Grant, P. L. Rosin, D. Marshall, and M. Chen. Visualization of time-series data in parameter space for understanding facial dynamics. In *Proceedings of the Computer Graphics Forum*, volume 30, pages 901–910. Wiley Online Library, 2011.

[41] J. Tao and T. Tan. Affective computing: A review. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pages 981–995. Springer, 2005.

[42] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze. Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots. *International Journal of Social Robotics*, pages 1–19, 2019.

[43] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017.

[44] V. Vijayakumar and R. Nedunchezhian. A study on video data mining. *International Journal of Multimedia Information Retrieval*, 1(3):153–172,

2012.

[45] S. Vijayarani and A. Sakila. Multimedia mining research-an overview. *International Journal of Computer Graphics & Animation*, 5(1):69, 2015.

[46] M. Weisbuch, N. Ambady, A. L. Clarke, S. Achor, and J. V.-V. Weele. On being consistent: The role of verbal–nonverbal consistency in first impressions. *Basic and Applied Social Psychology*, 32(3):261–268, 2010.

[47] A. Wu and H. Qu. Multimodal analysis of video collections: Visual exploration of presentation techniques in ted talks. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[48] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. Opinionseer: interactive visualization of hotel customer feedback. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1109–1118, 2010.

[49] X. Xie, X. Cai, J. Zhou, N. Cao, and Y. Wu. A semantic-based method for visualizing large image collections. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[50] P.-w. Yeh, E. Geangu, and V. Reid. Coherent emotional perception from body expressions and the voice. *Neuropsychologia*, 91:99–108, 2016.

[51] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[52] J. Zhao, L. Gou, F. Wang, and M. Zhou. Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pages 203–212. IEEE, 2014.