EmoCo: Visual Analysis of Emotion Coherence in Presentation Videos

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

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

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, 30] or deliberately (e.g., deadpan
Within this framework, emotion coherence among those channels plays 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 [20, 41]. A great amount of work [30, 33] 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 [22, 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. [57] 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] utilized a graph design to visualize the concurrency 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 Qiu [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 is employed to perform face detection, grouping, authentication and emotion recognition based on its good performance [38]. 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 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 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 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 reactions 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.

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 different channels. In the visual exploration phase, five coordinated views are provided to support three-level exploration.

4.1 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 phase, 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. 3). 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 the sentence level (T7). The word view (Fig. 1f) 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” (27) 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. 1 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 Plutchik’s 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 three channels (Fig. 1b), 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.

$$D_{coherence} = \begin{cases} 2, & E_{\text{face}} = E_{\text{text}} & E_{\text{text}} = E_{\text{audio}}, \\ 1, & E_{\text{face}} \neq E_{\text{text}} & E_{\text{text}} \neq E_{\text{audio}} & E_{\text{face}} \neq E_{\text{audio}}, \\ 0, & E_{\text{face}} \neq E_{\text{text}} & E_{\text{text}} \neq E_{\text{audio}} & E_{\text{face}} \neq E_{\text{audio}}, \\ \text{others} & \end{cases}$$.
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. 1) 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 5.5).

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. 2). In Fig. 2(b), 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. 2c). 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 better display the emotion coherence information and its dynamic trend. The final design is shown in Fig 2.

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. 3, this view contains three parts, namely the face channel (Fig. 3[a]), the text channel (Fig. 3[b]) and the audio channel (Fig. 3[c]). 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 emotion, 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 interaction feature allows users to further explore the corresponding emotions. Therefore, we considered a Sankey diagram design (Fig. 3). 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.
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. 7a, 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.

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

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.

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. 8a), 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. 8b). 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 speakers Edward de Bono (2009). 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. 1b). 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. 1b). He selects the sentence to unfold its details. Following the line chart in the middle, Kevin notices fluctuations of the black line and many grey lines, 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 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. 5), 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 expressions. 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. 5). 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. 8a-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. 8a-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. 8a-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. 3a). 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. 5, 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.
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 into 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
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 emotion categories 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 sample frames, 29 text segments, and the corresponding audio segments for the first video, and we tested 1000 sample 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 [13] 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.

**ACKNOWLEDGMENTS**

The authors would like to thank the anonymous reviewers for their valuable comments. This work is partially supported by a grant under Hong Kong ITF UICP scheme (grant number: UIT/142).


