EmotionCues: Emotion-Oriented Visual Summarization of Classroom Videos

Haipeng Zeng, Xinhuan Shu, Yanbang Wang, Yong Wang, Liguo Zhang, Ting-Chuen Pong, and Huamin Qu, *Member, IEEE*

Abstract—Analyzing students' emotions from classroom videos can help both teachers and parents quickly know the engagement of students in class. The availability of high-definition cameras creates opportunities to record class scenes. However, watching videos is time-consuming, and it is challenging to gain a quick overview of the emotion distribution and find abnormal emotions. In this paper, we propose *EmotionCues*, a visual analytics system to easily analyze classroom videos from the perspective of emotion summary and detailed analysis, which integrates emotion recognition algorithms with visualizations. It consists of three coordinated views: a summary view depicting the overall emotions and their dynamic evolution, a character view presenting the detailed emotion status of an individual, and a video view enhancing the video analysis with further details. Considering the possible inaccuracy of emotion recognition, we also explore several factors affecting the emotion analysis, such as face size and occlusion. They provide hints for inferring the possible inaccuracy and the corresponding reasons. Two use cases and interviews with end users and domain experts are conducted to show that the proposed system could be useful and effective for analyzing emotions in the classroom videos.

Index Terms—Emotion, classroom videos, visual summarization, visual analytics

1 Introduction

Emotions play a critical role in understanding classroom learning contexts. Prior studies have shown that emotions can influence students' learning behaviors [1], [2], [3], [4], [5]. For example, emotions can affect students' attention, their motivation to learn [6], students' learning strategies and self-regulation [2]. On the other hand, students' emotions can give important hints about their learning status in class. Specifically, students' emotions can reflect their feelings about the course content and indicate their engagement in group discussions. Therefore, it is of great value to explore the emotions of students in a classroom context. This kind of information can help both teachers and parents know about students' learning status and further help teachers improve teaching.

However, it is not easy for teachers to quickly capture and explore students' emotions in the classroom context [4], especially when they need to pay attention to many students at the same time. With the great development of digital technologies, it is possible to videotape student behaviors by using suitable recording equipment [7]. Analyzing these videos provides teachers and parents with great opportunities to tackle the above issues. Inspired by this, we are motivated by a crucial research question: *how can we help teachers and parents better explore students' emotion status and their engagement in class through classroom videos?*

The most straightforward way to digest classroom videos is watching them one by one. However, it is very time-consuming, especially when users need to conduct emotion analysis for every student. Although video summarization techniques have been widely used in analyzing different kinds of videos, such as surveillance videos [8] and sports videos [9], they could not directly be applied to summarize emotions in classroom videos. These general summarization techniques usually extract keyframes from a video based on different criteria [10], such as object recognition and event detection, and do not explore emotion evolution in videos. In addition, without providing a user-friendly interface and smooth interactions to support different analysis requirements, it is hard for users to explore students' behaviors in classroom videos using existing emotion recognition methods [11]. Therefore, an interactive visualization system that supports emotion extraction and visual analysis of students' emotion evolution would be highly valuable for parents and teachers.

It is a nontrivial task to visually summarize and analyze the emotion evolution of a group of people from lengthy and continuously recorded classroom videos due to three major challenges. First, summarizing the emotion evolution of a group of people is challenging. It is not easy to visually summarize multiple persons' emotion trends with little visual clutter and allow users to explore the emotion trend of an individual at the same time. Second, uncovering emotion patterns of each person is difficult, let alone comparing emotion patterns of different people. Various factors have to be taken into consideration, such as the emotion distribution across time, the correlation between different emotions, face size, and occlusion. Third, the results from the computer vision algorithms for emotion recognition are still not perfect [11]. Many factors account for the inaccurate results, such as complex face conditions and head motion occlusion [12]. These inaccurately extracted emotions may mislead users, and it is challenging to deal with these issues. Therefore, when designing a visualization system,

- *H. Zeng, X. Shu, Y. Wang, Y. Wang, T. Pong and H. Qu are with the Hong Kong University of Science and Technology, Hong Kong. E-mail: {hzengac, xshuaa, ywangdr, ywangct, tcpong, huamin}@cse.ust.hk.*
- *L. Zhang is with Harbin Engineering University, China. E-mail: zhangliguo@hrbeu.edu.cn.*

we should inform users of the potential errors and factors that can lead to these errors.

To address these challenges, we have designed and developed an interactive visualization system called *EmotionCues*, to analyze classroom videos with emotion-based cues at various levels of details. Our system contains three major views: the summary view displaying the overall emotions and their dynamic evolution, the character view describing personal characteristics related to emotions, and the video view enhancing the video analysis with further details. Given the possible inaccuracy of emotion recognition, we identify several influencing factors and visually encode them in our proposed system to infer students' situations. We believe these influencing factors play an important role in emotion analysis. For example, a person who is far away from the camera and often has a relatively smaller face size in the video is more likely to be incorrectly recognized. Moreover, for a person who is often occluded by others, the emotion recognition has a higher risk of being incorrect. Thus, we integrate these influencing factors into our analytical workflow. Rich interactions are also provided to enhance our system. To evaluate the usefulness and effectiveness of *EmotionCues*, we present two use cases and collect feedback through interviews with end users and domain experts. The major contributions of this paper are as follows:

- We propose an interactive visualization system to support automatic emotion extraction from classroom videos and multi-level visual summarization from the perspective of emotion.
- We, for the first time, integrate the model uncertainty into emotion analysis for classroom videos and provide visual cues for reasoning about the emotion evolution.
- We conduct two use cases and interviews with end users and domain experts to show the usefulness and effectiveness of our system.

2 Related Work

The related work of this paper can be categorized into three types: emotion analysis in learning scenarios, video visualization, and temporal data visualization.

2.1 Emotion Analysis in Learning Scenarios

Emotion analysis for understanding students' learning status has received considerable attention over the past few years. It is widely believed that emotions have great effects on students' learning and achievement [1], [2], [3]. Firstly, emotions can affect students' motivation and efforts toward their study [6]. Secondly, emotions have a great influence on students' cognitive processes, such as attention, learning and memory [13]. Thirdly, emotions can influence students on choosing their learning strategies and adjusting their level of self-regulation in learning [2]. Therefore, emotions can motivate and guide students' behaviors in learning, which serves as indicators of the status of students' participation in learning. Analyzing and understanding students' emotions can help teachers better organize their teaching and improve students' performance in learning [2].

Various approaches have been proposed to capture and measure students' emotions in learning scenarios [14], which can mainly be divided into two categories: *self-report methods* and *non-self-report methods*. *Self-report methods* allow users to collect emotion data that are not observed directly at relatively low cost. For example, Balaam et al. [4] collected emotion data with the Subtle Stone designed for supporting students' emotional communication in the classroom. However, data collected by *self-report methods* can sometimes be biased or unreliable. Also, *self-report methods* may distract students from the learning content when used during classes. *Non-self-report methods* are proposed to solve this problem, such as detecting learners' emotions in texts with language processing methods [15] and analyzing students' emotions in images with computer vision techniques [16]. In this paper, we apply computer vision techniques to analyze students' emotions in classroom videos.

Visualization can facilitate emotion analysis in learning scenarios. A multitude of studies have been conducted on analyzing and visualizing collected emotion data [17], [18], [19], [20]. However, existing work cannot be directly applied to our target scenarios, i.e., analyzing both an individual's and a group of people's emotion evolution in classroom videos. For example, Hernandez et al. [17] and Srivastava et al. [18] applied computer vision techniques to extract emotions from images or videos. However, they simply provided basic visual representations to describe emotion information without further analysis. Zhao et al. [19] proposed a system named PEARL for understanding personal emotion styles derived from social media by applying text analysis. However, this work mainly focused on analyzing individual emotion evolution without considering a group of people. Aimed at users' emotional reactions in public events, Kempter et al. [20] described EmotionWatch, a tool that visually summarized the emotion evolution of a group of people. However, this tool did not allow users to track and analyze individual emotion evolution. What is more, existing work failed to consider how to handle the inaccuracy of emotion recognition algorithms, which is important in our target scenarios.

In our paper, we focus on analyzing the emotion of both an individual and a group of people in videos. We propose a visualization system to conduct a detailed analysis of the emotion evolution and summarize content in a video from the perspective of emotion. Our system incorporates automatic emotion extraction and provides users with insights into the possible inaccurate results.

2.2 Video Visualization

Video visualization aims at using visualization techniques to provide users with a quick overview of the video content. The important information (e.g., some key features and crucial events) is extracted from the original video and presented through a visual method, helping users easily recognize certain patterns of a video. Borgo et al. [21] conducted a comprehensive survey on the research of this topic. After Daniel and Chen first introduced video visualization [8], it has been successfully applied in many applications such as sports video analysis [22], [23], [24], presentation video analysis [25], surveillance [26], [27] and biology applications [28]. Also, there are some studies focusing on efficiently exploring large video collections, such as Video Lens [29].

Depending on whether the video frames are used in the visualization or not, video visualization techniques can be roughly divided into two categories [28]: *image-based techniques* and *abstract techniques*. The *image-based techniques* usually select the representative frames of a video to convey the key information. Some work attempts to provide a quick navigation for viewers and still maintain the full video content [30], [31], [32], [33]. Other studies compress video content. For example, Kang et al. [34] introduced a space-time video montage, where the informative parts of the video are selected and merged together. Video skimming techniques generate a short summary of the video by removing the less interesting parts [35]. The *abstract techniques* are employed to visualize the temporal attributes of the video in an indirect way, where the objects in the video may not be directly shown. Daniel and Chen [8] used volume visualization techniques to summarize the video sequences and bent the video volume into a horseshoe shape, which was further extended by considering the motion flow [36]. Duffy et al. [28] designed a glyph to encode the attributes of semens in the video. Meghdadi and Irani [26] visualized the movement trajectory in a 3D cube.

In our work, it is not suitable to apply image-based techniques since we mainly focus on emotion evolution and summary. Directly using images will add to users' burden in tracking the emotion evolution of people in videos. Therefore, we mainly adopt *abstract techniques*, which directly extract people's emotion information from videos, and then visualize the emotion information in an effective way to facilitate analysis. At the same time, due to insufficient performance of recognition algorithms, model uncertainty needs to be taken into consideration. Höferlin et al. [37] proposed a visual analytics method to handle the uncertainty of tracking moving objects in videos, which allows users to provide filtering-based relevant feedback. A lot of research has focused on directly visualizing uncertainty data to raise users' awareness about the uncertainty [38], [39], [40]. In this paper, we adopt both strategies. First, we directly visualize model uncertainty by visually encoding some influencing factors, which gives users some hints about model performance. Then, we also provide a set of interactions for users to correct errors.

Different from prior work on video visualization, this work summarizes video content from the perspective of emotion, which is a field that few studies have focused on before. Also, to the best of our knowledge, the proposed system is the first to integrate the model uncertainty into emotion analysis for classroom videos and provide some visual cues for reasoning.

2.3 Temporal Data Visualization

There have been various approaches for visualizing and analyzing temporal data [41], [42]. Here we discuss the most related visualization techniques for temporal data.

PEARL [19] adopts a themeriver design to show personal emotion style derived from social media. The height of each band represents the proportion of corresponding emotion. Different from this work, we visualize the emotion information derived from videos and provide an emotion overview of a group of people. To visualize a group of people, storyline visualization [43] can intuitively convey relationships among entities over time. Each person is represented as a line. For people with a close relationship, the corresponding lines are drawn closely. However, this kind of visualization will easily cause serious visual clutter as the number of people increases. To better visualize a group of people, aggregation techniques are widely used. For example, TextFlow [44] uses an aggregation flow to show the relationships among topics in the text corpora. Similarly, TelcoFlow [45] adopts an aggregation flow to provide an overview of the collective behaviors of people's movements.

Inspired by prior work, we propose a visual analytics system with the aggregation flow and storyline visualization techniques. Our proposed system provides both an overview of a group of people's emotions and a detailed exploration of a specific person's emotions.

3 Data and Design Requirements

In this section, we first introduce the video data used in the paper and the models employed to process videos. Then, we summarize a set of design requirements based on the discussion with end users.

3.1 Data Description

The video data we use are mainly collected from our collaborating kindergartens. Teachers in the kindergartens use different cameras to shoot videos of children in class. Each video is about 10 minutes long (1.26 G) with a resolution of 1920×1080 and 30 frames per second (FPS). That is, each video consists of nearly 18,000 high-resolution frames with a wealth of details. Also, to further demonstrate the usefulness and effectiveness of our system, we apply the proposed system to videos with similar scenario settings, which are collected from university seminars. For more details, please refer to Section 6.2.

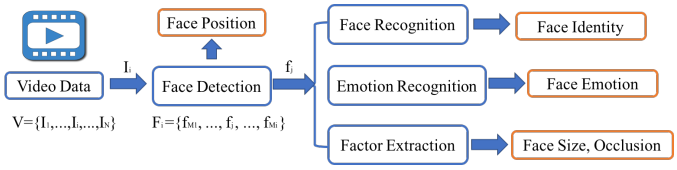


Fig. 1. The workflow of data modeling and processing. Given a video V, face detection is conducted on sampling frames I_i. Then, face recognition, emotion recognition and factor extraction are conducted on detected faces f_j. Finally, we obtain the related information, i.e., face position, identity, emotion, size and occlusion.

3.2 Data Modeling and Processing

Given a video V, we model it as a series of images: $V = \{I_1, I_2, ..., I_i, ..., I_N\}$, where I_i is the i-th frame in V and N represents the number of frames in V. Each video contains a large number of frames, and frames in close proximity are almost the same. Hence, we conduct a sampling method to remove some redundant frames. For example, given 30 frames per second, we can evenly extract three frames in one second by setting the sampling rate to $1/10$, which can largely reduce computation cost without losing too much information. As shown in Fig. 1, when processing a video, we first detect the human faces that appear in the video. Then, we recognize their corresponding identities and facial expressions. We also calculate face size and occlusion information. The whole process is as follows:

Face Detection. We adopt a deep learning model called MTCNN (Multi-task Cascaded Convolutional Networks) [46], a deep convolutional network to predict face and landmark locations, to detect human faces in each sampling frame. It achieves a better performance when compared to other methods, such as OpenCV[1] and dlib library[2]. We can derive the position of each face f_j in the frame I_i, i.e., $F_i = \{f_{M_1}, ..., f_j, ..., f_{M_i}\}$, where M_i is the number of faces in I_i. Specifically, since a face is detected in a rectangular area, we use the position of the rectangle to represent the face. That is, $f_j = [x_j, y_j, w_j, h_j]$, where x_j and y_j are the coordinates of the left top corner, and w_j and h_j are the width and height of the rectangle, respectively.

Face Recognition. Based on the faces detected in each frame, it is necessary to identify each face with its corresponding person. Basically, we compare the input face image with the query face database to find the most similar one. A general approach to face comparison is through image vectorization. Here we adopt facenet [47], a well-established deep learning model for face recognition, which can directly learn a mapping from face images to a compact Euclidean space. The output after applying facenet to each face image (resized to 160×160) is a 128-dimensional vector. We can measure the face similarity based on the distance between face vectors. After that, since one person can only appear once in each frame, we consider face recognition as an assignment problem instead of a classification problem, and adopt the Hungarian method [48], which achieves a better result in our scenario.

Emotion Recognition. Generally, there are two widely used categories of emotion models: *categorical (discrete) emotion states* and *dimensional (continuous) emotion space*. In the categorical emotion models [49], emotions are modeled as a few basic emotion types, such as happiness, sadness, and the like, which are easy for users to understand. In dimensional emotion models, emotions are modeled as points in a continuous n-dimensional space, such as the pleasure-arousal-dominance (PAD) space [50] or the valence-arousal (VA) space [51]. Although dimensional emotion models are more flexible and richer in their descriptive powers, they are often not transparent to people. Therefore, we adopt categorical models due to their intuitiveness and understandability [52]. We fine-tune a CNN model (ResNet-50 [53]) with the FER 2013 dataset [54], a public and widely used dataset for facial expression recognition. The emotions in this dataset are categorized into seven types, i.e., $emotionSet = \{$"$anger$", "$surprise$", "$happiness$", "$neutral$", "$sadness$", "$disgust$", "$fear$"$\}$. The training dataset consists of 28,709 images and the test dataset consists of 3,589 images. The human accuracy of this dataset is $65\% \pm 5\%$ [54], while the training model we use achieves an accuracy of about 71.2%, which is better than the human accuracy. The output of the model is a set S of possible emotions with different probabilities, i.e., $S = \{P_e | \sum P_e = 1, e \in emotionSet\}$, where P_e is the probability of the corresponding emotion in $emotionSet$. Finally, we determine the emotion of the face as one category with the highest probability, i.e., $Emotion = \text{argmax}_{e \in emotionSet} P_e$.

1. https://opencv.org/
2. http://dlib.net/

Exploration of Influencing Factors. Many factors have an influence on the accuracy of emotion recognition [12], [55], such as face size, occlusion, image resolution, and illumination effect. Considering those factors with obvious impacts in our scenarios, here we mainly select face size and occlusion, as these two factors directly reflect the difference of children's status in different time ranges within one video. We propose a straightforward design to visualize face size and occlusion, helping users better understand the model results. Face size can be estimated by the area of rectangles in face detection, and occlusion can be calculated with an existing method [56].

3.3 Design Requirements

Before designing an approach for analyzing students' emotions from classroom videos, the first crucial question to be answered is: *what kind of design requirements are needed by the teachers and parents for better analyzing students' emotions in classroom videos*? Therefore, we worked closely with four end users (two teachers (U1 and U2) and two parents (U3 and U4)) for more than six months by following a user-centered design process. U1 and U2 are from two different public kindergartens and have 6 and 7 years of teaching experience, respectively. U3 is a mother of two kindergarten children. U4 is a father whose daughter has been in a kindergarten for more than one year. We held weekly meetings and also communicated through emails regularly to identify their requirements. By presenting our prototype system to our end users, we gathered their feedback and iteratively refined the specific design requirements based on Shneiderman's mantra (overview first, zoom and filter, then details-on-demand) [57] and Tamara's nested model [58]. The requirements can be summarized as follows:

- **R1** **Obtain the emotion status of all the people in a video.** Given a specific video, users have a great interest in gaining a quick overview of the video content. For example, what is the overall emotion trend as the video progresses? What kind of emotion dominates the video? Compared with checking the original video back and forth, a visual overview would greatly reduce the browsing burden.
- **R2** **Uncover emotion patterns of an individual in a video.** After gaining an overview of the given video, users concentrate on an individual of interest. For example, most parents are concerned about their own children, and they are likely to explore individuals in a video. What is the emotion pattern of a selected person in this video? How do his/her emotions evolve over time?
- **R3** **Compare emotion portraits of different people.** Users would like to explore a person of interest, especially to obtain his/her relative status in a video. Further comparisons between different people empower users to identify abnormal patterns. For example, teachers may worry about a special student in the class, and parents are curious about whether their children behave differently compared to others. Therefore, comparing different people's emotion patterns and measuring their similarity and difference are very valuable for users.

R4 **Reveal model uncertainty with influencing factors.** Emotion recognition algorithms are not perfect and the accuracy is influenced by multiple factors [12]. Leveraging these factors properly can provide useful cues for inferring underlying patterns. For example, the accuracy of emotion recognition probably decreases, when the algorithm processes a child face image with a small face size in the video or occluded by others. It would also be better to allow users to investigate model accuracy and correct corresponding errors if needed.

R5 **Provide context for video analysis.** The visual analytics system is based on complex recognition models and abstract data representation. Users also want to know the original video context, which helps them understand the analytical results and validate assumptions. For example, what kind of scenario leads to a change of emotions? Do their assumptions about these findings make sense?

4 System Overview

Fig. 2 demonstrates the whole pipeline of our system including the data processing phase and the visual exploration phase. The first phase processes a set of raw videos and extracts emotions using computer vision algorithms. The details are provided in Section 3.2. Then, based on the five major design requirements (Section 3.3), we design an interactive visualization system, which can support the visual analysis of classroom videos at two different granularities: the overall evolution patterns of all people's emotions, and the specific description of a person's emotion evolution.

We implement a web-based system based on the Vue.js front-end framework and the Flask back-end framework. Fig. 3 shows that our system contains three major views, namely, summary view (Fig. 3a-b), character view (Fig. 3c) and video view (Fig. 3d). The summary view contains two parts, namely, the emotion archives (Fig. 3a) on the left-hand side and the emotion flow (Fig. 3b) on the right-hand side. The emotion archives provide the proportion of the overall emotional composition of the people in the video (R1). The emotion flow summarizes the emotion evolution of the people in the selected video (R1). In addition, selecting people in the emotion archives connects corresponding lines to the emotion flow, which can help users compare the emotion evolution of selected people and identify certain interesting patterns (R2-3). The character view provides a visual signature of selected people, which enables users to explore the relationship between emotion information and influencing factors information (R4). Furthermore, this kind of signature can facilitate users to make comparisons between different people (R3). The video view shows the video we are exploring, which presents pieces of evidence for our exploration (R5). It also highlights all the faces detected in the video and provides a series of interactions for correcting inaccurate results caused by the model.

5 Visualization Design

In this section, we first describe three design rationales of our system based on the discussion with our end users (U1-4). Then, we present a set of corresponding visual designs

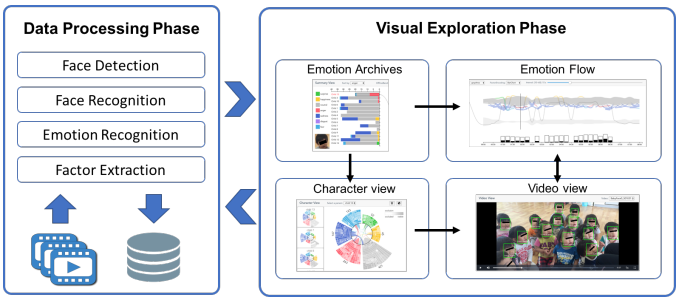


Fig. 2. The visualization system pipeline for emotion-oriented video summarization. After collecting raw videos, we go through a data processing phase (face detection, face recognition, emotion recognition and factor extraction) and store the extracted data into MongoDB. Then, we can smoothly perform the visual exploration phase, including the overall exploration and the detailed exploration.

aiming at the in-depth analysis of emotions in classroom videos. For a unified color encoding, we use the following color scheme, i.e., green for surprise, yellow for happiness, gray for neutral, magenta for anger, blue for sadness, purple for disgust, and cyan for fear, which is consistent with conventions and also recommended by our end users.

5.1 Design Rationales

We followed the three design rationales below to design our proposed system, *EmotionCues*.

Intuitive encoding and design. The target users of our system are mainly teachers and parents, who may lack the fundamental background of data visualization. Simple visual designs with familiar metaphors are easier for them to understand.

Smooth interactions with prompt feedback. The system should support smooth interactions allowing an engaging exploration for users to focus on the emphasized person without distractions. The prompt visual feedback in response to each interactive action can enhance the whole system with a better user experience.

Multi-scale visual exploration. Emotion information extracted from videos covers a rich variety of content. Multi-scale exploration giving both an overview and the details can support a quick and thorough understanding of the video content from the perspective of emotion.

5.2 Summary View

It is important to provide users with an overview of the emotion evolution of individuals (R1). Thus, we design a summary view to provide users with both a static and a dynamic summary of the emotions: the emotion archives (Fig. 3a) to visualize the emotion distribution of individuals (static summary), and the emotion flow (Fig. 3b) to show the dynamic evolution of these emotions (dynamic summary).

Emotion Archives. As shown in Fig. 3a, each horizontal stacked bar represents one person, and the length of each area inside indicates the number of occurrences of the corresponding emotion. This design can provide users with the basic emotion information of all the people in the video and the proportions of each type of emotion they display.

The horizontal stacked bars are initially sorted by name. To better compare different emotion components, this view supports sorting interactions, which allow users to sort

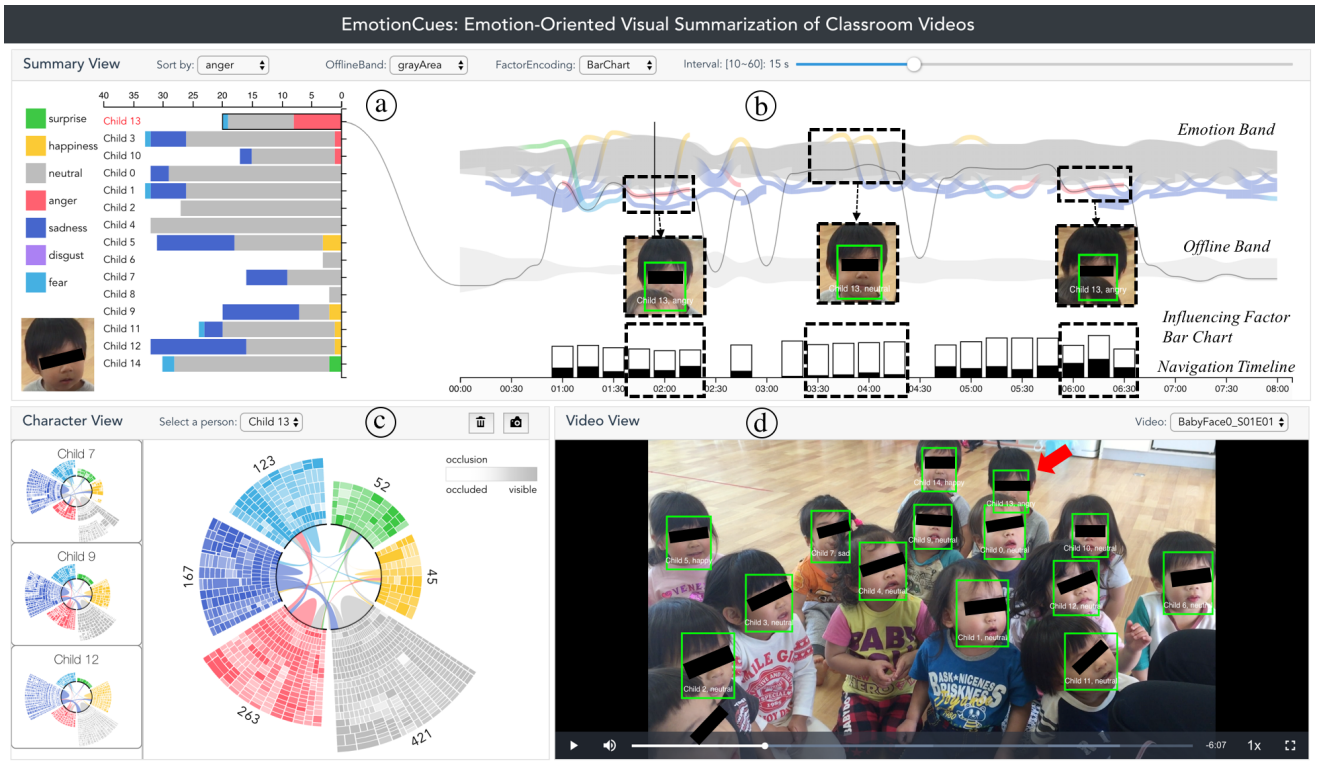


Fig. 3. Our visualization system for emotion-oriented video summarization with three major views. The summary view consists of two parts: the emotion archives (a) display the emotion distribution of each person in the video, and the emotion flow (b) shows the overall and individual evolution. The black dashed rectangles and thumbnails of faces in the emotion flow are added manually for illustration. The character view (c) describes a selected person's emotions with different factors, such as face size and occlusion. The video view (d) provides the original video and highlights the faces detected in each frame. Children's eyes are covered with black rectangles for privacy reasons. The same operation is also done for the subsequent related figures.

horizontal bars according to different attributes, such as the total amount of emotions and the dominant emotions. The default ranking order in the horizontal stacked bar is based on the legend order. Once users sort the horizontal stacked bars by a specific emotion type, bars of this emotion type will align on the right for easy comparison. Two examples are shown in Fig. 7 (sorted by sadness) and Fig. 9a (sorted by happiness). This view supports a selection interaction to further analyze the person of interest. By clicking a horizontal stacked bar, the corresponding person is highlighted with a line connected to the emotion flow for further exploration.

Justification: We considered alternative designs before finally adopting this design. For example, a pie chart for each person encodes the distribution proportion. However, it is difficult for users to recognize the number of occurrences of each emotion in a pie chart. Moreover, it does not support an intuitive comparison among different people. When they are ranked based on the degree of happiness, the pie chart cannot provide direct visual cues. Inspired by Lineup [59], we adopted the horizontal stacked bar design.

Emotion Flow. The emotion flow (Fig. 3b) comprises four major parts, namely, an emotion band, an offline band, an influencing factor bar chart, and a navigation timeline from top to bottom.

The emotion band mainly adopts a flow-based design. The order of the emotion band follows the legend order, i.e., positive emotions at the top, neutral emotion in the middle, and negative emotions at the bottom. The width of each band encodes the number of people at that timestamp. Our end users (U1-4) are all satisfied with this flow design, which reveals the overall emotion evolution clearly.

The offline band shows the number of persons who have appeared in the video but are not identified in the scene at that moment owing to various reasons such as face occlusion and being outside the scene. When many people are not captured, the offline band may look prominent. To get a better visual effect, we allow users to configure different styles of the offline band, such as the gray area and dashed line area. The branches flowing in or out of the offline band are hidden by default since they can cause severe visual clutter. When users brush an area in the offline band, the corresponding flows will show up (Fig. 7f).

The influencing factor bar chart mainly shows the aggregation information of face size and occlusion (R4). We use a bar to encode different influencing factors for each time span. The height of the bar indicates the face size detected in the video (the higher, the larger face size), whereas the black shading area of the bar represents the occlusion degree.

The navigation timeline at the bottom of the emotion flow (Fig. 3b) is connected to the video view and mainly used for navigation. Once users click on the timeline, the video view locates the corresponding frame. As this view aggregates information over a time span, we allow users to customize this span via the aggregation slider. For example, users can set the interval value to 30 seconds, then the view of the aggregation information is based on 30 seconds. We simply adopt the maximum and average strategies to determine the final emotion and calculate influencing factors in this span, respectively. Other aggregation strategies can easily be adopted by our system.

Moreover, we use lines to connect the emotion archives to the corresponding flows. Each line represents a person,

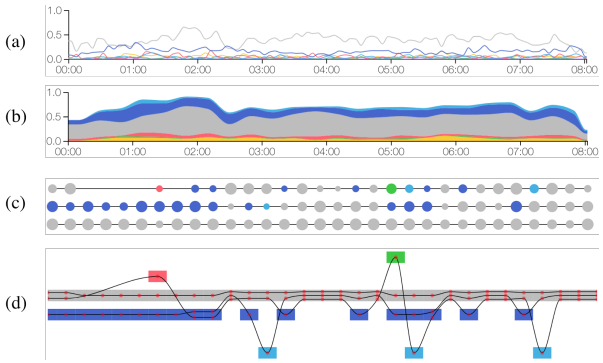


Fig. 4. Alternative designs for emotion flow. (a) A line-based design, where each line represents an emotion. (b) A stream graph-based design, where each layer corresponds to an emotion. (c) A sequence-based design, where each thread shows a person's emotion evolution. (d) A storyline-based design, where each line illustrates a person and emotion categories are regarded as locations.

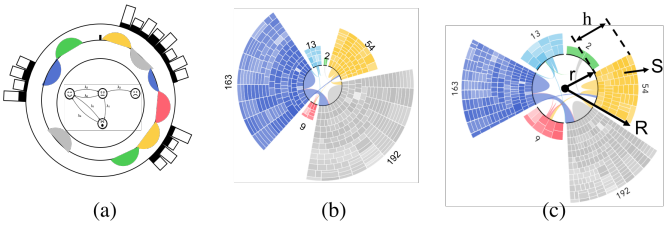


Fig. 5. Portrait glyph design. (a) Time-based circular design: the first circle is used to encode emotion evolution and the second circle is used to encode occlusion information. A node-link graph in the middle is used to show the transitions among different emotions. (b) Emotion category-based design (separation based on each emotion's proportion). (c) Emotion category-based design (equal division): each annular sector represents the number of each emotion, and each slice encodes face information detected in the video (slice size for face size and opacity for occlusion). A chord diagram in the middle shows the transitions among different emotions.

starting from a person selected in the emotion archives, and then connects to his/her emotion flow. Therefore, users can easily track personal emotion evolution (R2) and compare the emotion evolution of different people (R3).

Justification: Before adopting a flow-based design to describe the emotion evolution, we discussed several candidate designs, which are shown in Fig. 4. In the beginning, we considered using lines to show the emotion evolution. However, multiple lines corresponding to various emotions may lead to severe visual clutter (Fig. 4a), which makes it challenging to capture the video content. In order to alleviate the visual clutter, we turned to a stream graph, where different layers represent different emotions. This kind of design (Fig. 4b) directly shows the variety of emotion distribution over time. However, the stream graph design hardly provides hints for the emotion evolution of individuals. To reveal personal emotion evolution, we considered using a sequence-based design. As shown in Fig. 4c, each thread represents a person, and emotions are encoded by different colored dots along the thread. This design can easily track the emotion evolution at a personal level. However, it is hard to discover relationships among different people. Therefore, we considered a storyline-based design, where each line represents a person, and each emotion category can be regarded as a location. As shown in Fig. 4d, we can observe interactions between different people. However, our end user (U3) commented that "It is a little complicated. Since we have video data, it would be better to give a simple summary view and provide some hints for seeking the corresponding frames in the video data". After further discussing with end users (U1-4), we came up with the flow-based design, which shows the overall emotion evolution and individual emotion evolution with interactions. The end users (U1-4) felt satisfied with this design.

5.3 Character View

The character view visualizes the emotion status of a selected person (R2) with a portrait glyph. Comparison between different emotion portraits enables users to identify and compare the characteristics of different people (R3). As shown in Fig. 5, we adopt a tailored donut chart in this design. Each annular sector on the outer part represents an emotion. The area of each annular sector illustrates the amount of the corresponding emotion which appears in the video. We calculate the height of each sector using the following equations:

$$\begin{cases} S = \dfrac{\theta}{360}\pi(R^2 - r^2) \\ R = r + h \end{cases} \Rightarrow h = \sqrt{\frac{360 * S}{\theta * \pi} + r^2} - r$$

where θ, h, and S are the degree, height, and area of the annular sector respectively. Meanwhile, r and R are the inner and outer radius respectively. These annotations are illustrated in Fig. 5c.

Each annular sector is further divided into multiple small slices. The total number of slices within each sector is displayed beside it, as shown in Fig. 5c. Each slice represents information of the selected person at a particular frame, consisting of the emotion, face size, and occlusion. The slice size encodes the face size, and its opacity encodes the occlusion extent. Low opacity of the slice indicates that recognizing an emotion at that time is difficult. That is, the corresponding person is severely occluded and the recognition result is likely to be incorrect. For example, a person with a larger and unobstructed face at that time corresponds to a larger and visible slice in our system. Moreover, the contour-based treemap [60] sheds light on our layout to place these slices. For each annular sector area, slices are arranged by time, in a clockwise and outward direction.

In the center of the portrait, we adopt a chord design to encode transitions among different emotions. The chord thickness represents the frequency of the emotion transfer. For example, a person whose emotion switches from neutral to happiness will often have a thick gray chord from the neutral sector (gray) to the happiness sector (yellow).

We can easily observe detailed emotion information and influencing factors for the person of interest (R4) with such a tailored donut chart design. Also, the comparison of the basic emotion information between different people (R3) becomes easy with the screen snapshot function. If users want to explore details, they can click the snapshot of interest for further exploration. Snapshot examples are demonstrated on the left-hand side of the character view (Fig. 3c).

Justification: Before determining an emotion category-based design to describe detailed emotion and influencing factor information, we considered several alternative designs. First, bar charts were our initial design. However, they require much space and it is not easy to integrate other

(a) (b)

Fig. 6. An interactive method for users to correct an inaccurate label for the highlighted girl. The left-hand side (a) shows the interface for correcting face recognition results, while the right-hand side (b) shows the interface for correcting emotion recognition results.

designs into bar charts. Thus, we preferred a circular design, which is space efficient and easy to integrate with other designs like the chord we used in Fig. 5c. We also considered encoding temporal information around the circle (Fig. 5a). Nevertheless, it is difficult to show more details and provide a quick overview for comparison. Then, we further explored the category-based design with a chord diagram. For the circular information encoding, we compared the designs between equal division (Fig. 5c) and separation based on the proportions of different emotions (Fig. 5b). That is, for the equal division design, we assigned the same angle to each emotion. As for the latter, we divided the whole circle on a proportional basis according to each emotion's appearance. However, when an emotion accounts for a tiny proportion, it is difficult to inspect it clearly in this design. After discussing with our end users (U1-4), they all appreciated the equal division design (Fig. 5c), which clearly reveals the relationships between emotions and influencing factors. Therefore, we chose the equal division (Fig. 5c) as our final design.

5.4 Video View

We provide the original video for users to explore in the video view. Users can play the video at slow, normal, or fast speeds. When users pause the video, the corresponding faces in each frame are highlighted. Users can also pick out the parts of interest for further exploration based on their observation from the emotion flow. This view is mainly used for providing evidence for users (R5). When users explore other views and find something interesting, they can link to corresponding frames in the video view. Accurately extracting information from a video is a challenge. Therefore, we provide an interactive way for users to correct this inaccuracy (Fig. 6). When users identify a wrongly labeled person, they can click on the person and select the correct label. The face label will be automatically updated in the database. Similarly, users can correct the emotion information. With this method, users are allowed to interactively correct any inaccurate information caused by models.

5.5 Interactions

Based on the design rationales we identified, our system should support simple but effective interactions, such as click, brush and hover, to facilitate users in exploring video data. Here, we summarize the interactions provided in our proposed system.

Linking. The three views in the system are linked together. For example, when users click on the timeline below the emotion flow in the summary view, the video view displays a corresponding frame and vice versa. When playing the video, the black vertical line in the emotion flow will move forward at the same pace. Once users select a person of interest, the corresponding portrait is shown in the character view. Since each slice represents a face extracted from the video, an interaction is provided to seek the corresponding face by clicking a slice in the portrait glyph. This interaction can help users obtain evidence from the video to verify their findings.

Filtering. We adopt a set of filtering interactions to improve the scalability of the system and alleviate the visual clutter. Users can select people of interest, then the corresponding lines show how the selected people's emotions evolve in the emotion flow. One emotion can be dominant in some videos, and users are allowed to filter out some emotions and focus on the emotion of interest.

Configuration. The system enables users to make configuration changes. For example, users can set the time span used in the summary view by adjusting the slider bar at the top. The corresponding views are then updated. Also, users are allowed to configure different styles of the offline band to have a better visual effect.

6 Use Cases

In this section, we present two use cases to evaluate our proposed system, *EmotionCues*. In the first case, a teacher used *EmotionCues* to explore kindergarten videos. In the second case, a professor applied *EmotionCues* to analyze university seminar videos. The first case directly targets kindergarten scenarios, while the second case further examines the effectiveness when our system is applied to other similar scenarios. Neither the teacher nor the professor was involved in designing *EmotionCues*. After introducing the system, we asked users to freely explore corresponding videos with our system in a think-aloud protocol. The main tasks of their exploration are: 1) find out the overall emotion status and engagement of the students; 2) explore some students of interest; 3) compare the emotion distribution and evolution of different students. We recorded their exploration process and findings with their consent. Finally, we interviewed the users and collected their feedback. Please note that the videos used in the following cases have been authorized for use for research purposes. We covered the eyes with black rectangles to avoid revealing their identities, as shown in Figs. 3, 6, 7, 8 and 9.

6.1 Case One: A kindergarten classroom video

Here we report the case when a teacher was exploring a kindergarten classroom video. The video lasts about 10 minutes. Fifteen children are sitting on the ground and listening to a story told by a caregiver in the video. The teacher has about 4 years of teaching experience. Her duty is to teach a class of 20 children. She told us that she needs to pay attention to the children's status, for example, whether they are happy and whether every student is engaged in her class. Manually reviewing the classroom videos is often

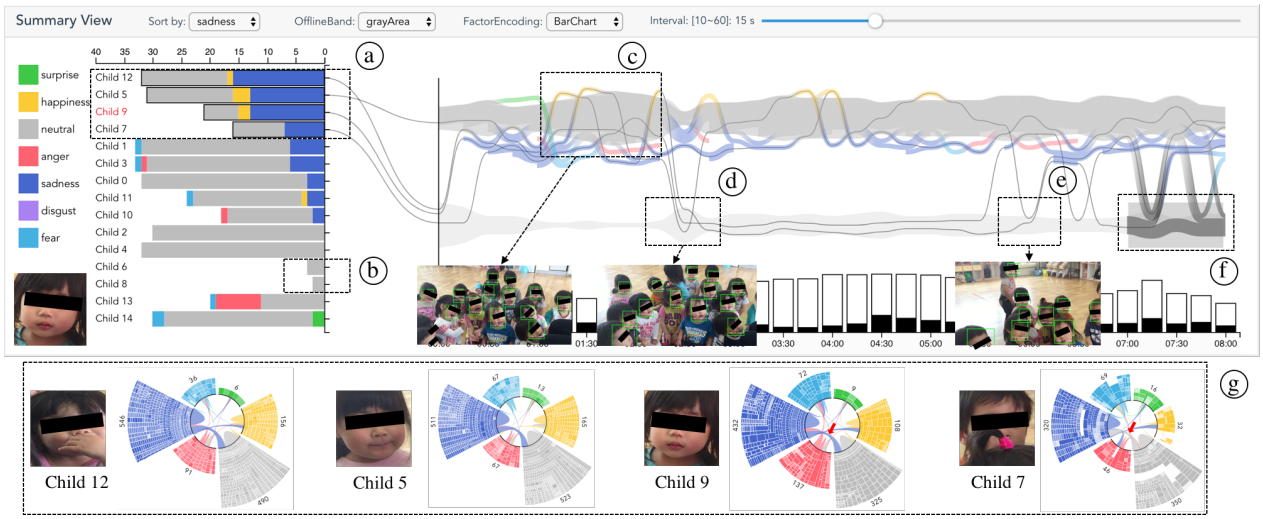


Fig. 7. Findings in Case One. (a) The top four children are listed by sadness sorting. (b) Two children with short stacked bars. (c) A happy moment of children is located by observation of the emotion flow. (d) Child 9 leaves the camera. (e) Child 9 comes back. (f) Brush interaction on the offline band. (g) Four portrait glyphs of the four children. The black dashed rectangles and thumbnails in the summary view are added manually for illustration.

very time-consuming, so she showed great interest in trying *EmotionCues* on exploring the classroom videos.

The first goal of the teacher was to observe the overall emotion status of all the children and the emotion evolution of each child in the video. Such kind of information can help her know the children's feelings about her teaching during the class. When she was analyzing the video with *EmotionCues*, she chose to look at the summary view first and the emotion archives directly showed the emotion distribution of children (Fig. 7). She found the dominant emotion was neutral, followed by sadness, happiness, and anger (R1). The same distribution was also observed in the emotion flow (R1). She could easily find some happy moments in the video, where the children were imitating the caregiver's actions (Fig. 7c). Then she observed that the stacked bars of Children 6 and 8 were very short (Fig. 7b). After checking the emotion flow and the video view, she quickly discovered that the two children's faces were often totally occluded by others and sometimes moved off the screen (R5). Further, she found an interesting student, Child 13, as his major emotion was anger (Fig. 3a), which was different from the other students. She clicked the name of Child 13 in the emotion archives (Fig. 3a) to further explore the details (R2). She found that it was mainly due to the partial occlusion of Child 13's face, which is clearly indicated by the high percentage of the black shading area of the influencing factor bar chart (Fig. 3b). This kind of visualization of the potential uncertainty effectively informed her of the potentially inaccurate emotion recognition results (R4).

Apart from exploring the overall emotion distribution, the kindergarten teacher also wanted to explore the emotion status of the children of her interest (R2) and compare the emotions of different children (R3). Initially, she felt confused and asked why sadness was so prevalent. She sorted the children by sadness and selected the "top four saddest" children (Children 12, 5, 9 and 7) in the emotion archives (Fig. 7a). Then she checked the snapshots of their portrait glyphs in the character view (Fig. 7g) and the corresponding frames in the video view (R5). These children were indeed recognized as being sad in some frames when they were actually listening to the teacher carefully. In addition, she further compared these four children. She observed that both Children 9 and 7 have shorter stacked bars compared with Children 12 and 5 (R3). Similarly, the portraits of Children 12 and 5 were denser than the portraits of Children 9 and 7, as shown in Fig. 7g. After checking the original video through linking interactions, she easily found that this was because the camera did not capture Children 9 and 7 in some moments (R5). Further, she noticed that the portraits of Children 9 and 7 told an entirely different story. The portrait of Child 7 had more shallow slices, which indicated that he was very often occluded by others. While the portrait of Child 9 indicated she did not have too much occlusion. By checking the video view, she found that Child 9 left the camera scene for a period of time (Fig. 7d-e), while the place where Child 7 sit was easily occluded by others (R4). She further noticed that the thick chords in the center showed that Child 9 was more likely to have switched from sadness to neutral, while Child 7 was more likely to have switched from neutral to sadness. These two children had different feelings in the class, and the teacher said that this was an interesting finding and worth further checking with the children.

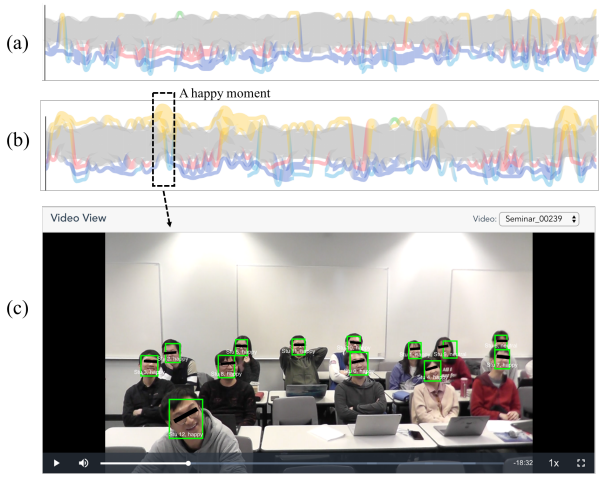


Fig. 8. Two different events of the same group seminar. (a) An academic presentation. (b) An internship experience sharing. (c) One happy moment of the internship experience sharing is highlighted.

Overall, she felt that it was easy to use *EmotionCues* and it can effectively help her explore the emotion status of children and track their evolutions. She highly appreciated the summary function and smooth interactions, which can save her much time. However, she commented that emotion recognition was not so accurate. She was glad that we have considered the model uncertainty by visualizing the influencing factors, which could give her hints about the inaccurate results. For those portrait glyphs, she commented that though they looked complicated at first glance, it was actually easy for her to understand after the introduction.

6.2 Case Two: Seminar videos

In this case, we worked with a professor in a university who needs to weekly organize group seminars with students in a research lab. She was very interested in using the system to monitor and improve the quality of the seminars. Previously, she often needed to talk to students and collect their feedback about each seminar. However, our video-based approach and system could make the whole process more accurate and efficient. She explored two videos which record two events in a seminar: one video is about an academic talk where a postgraduate student presents his research idea and seeks feedback from the audience, and the other video is an experience sharing event where another postgraduate student shares his 4-month overseas internship experience. The two videos are 32 and 24 minutes, respectively. Thirteen students appear in the videos.

The first goal of the professor was to get a quick summary of students' emotions in the seminar events, which could directly indicate the audience's feedback about the seminar. Previously, she got feedback from students by talking to them, which, however, was a bit subjective and time-consuming. When using *EmotionCues* to analyze the two seminar videos, she was able to get a quick and insightful emotion summary from the summary view. As shown in Fig. 8, it was easy to distinguish the differences between the two seminar events (Fig. 8 a-b). For the academic presentation, more students tended to be neutral, sad or angry (R1). At first glance, she was a little confused with the sadness and anger emotion. Then with the help of linking interactions, she checked the corresponding frames in the video view (R5). She found that when the students looked serious or confused, the backend emotion recognition tended to recognize their emotion as sadness or anger. Though these labels were not technically correct, they reflected the overall emotion status of students when the students were focusing on the seminar and thinking about the research questions posed by the presenter. For experience sharing, the professor quickly found that there was less sadness and anger, and the students looked happier, which was quite different from the academic presentation event (R1). This was mainly because the second seminar presenter shared many interesting findings and stories during his overseas internship. A screenshot of the happy moment during the experience sharing event was shown in Fig. 8c.

The professor was also quite interested in exploring the engagement of students in the seminar, as their emotion could indicate whether they were engaged in the seminar events or not. She selected the video of the experience sharing event for further exploration. It seemed that most of the students were engaged in this activity, since the majority of them had long stacked bar charts and showed more positive emotions (Fig. 9). As shown in Fig. 9a, after sorting the emotion by happiness, the professor easily found that Student 4 had most happiness. She further clicked Student 4 in the emotion archives to check the details. The major emotion of Student 4 was happiness, followed by neutral and sadness (R2). Since the experience sharing was full of joy, the professor was curious about why Student 4 showed such sadness. She further checked the video view and found that Student 4 was sometimes felt confused about the content, which looked like sadness in terms of facial expression.

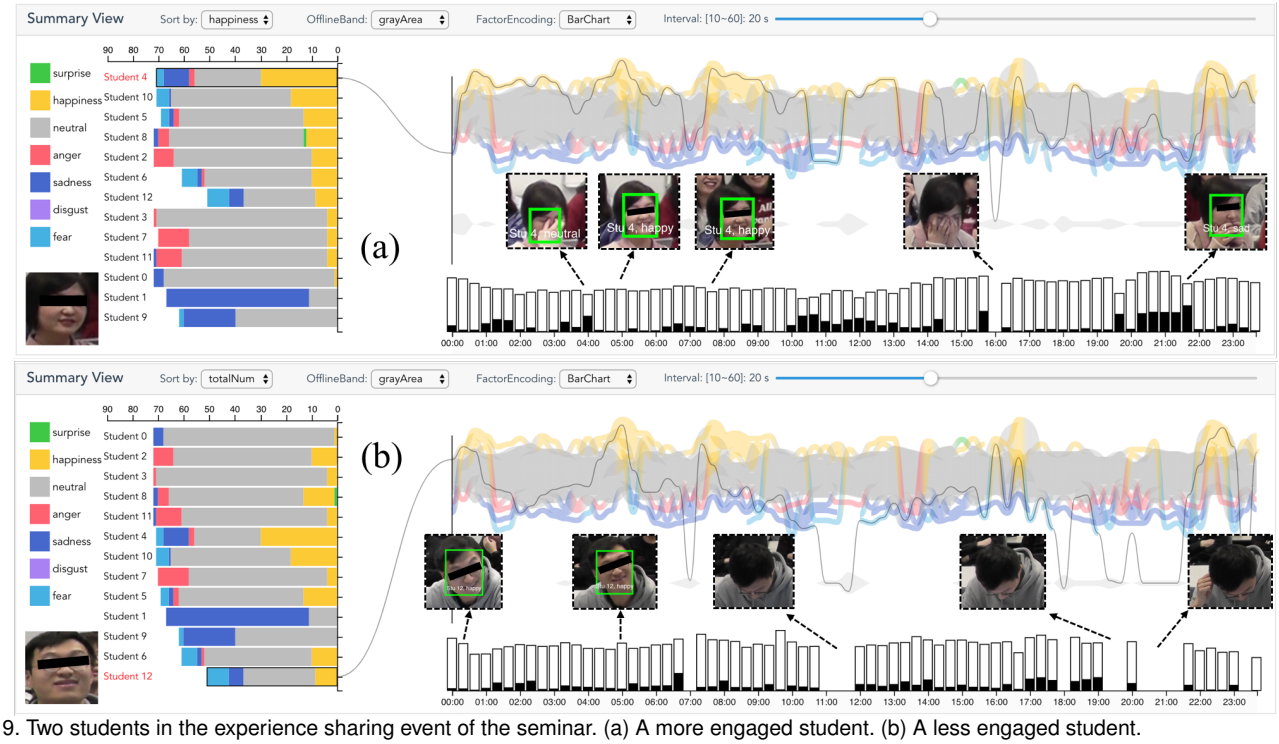


Fig. 9. Two students in the experience sharing event of the seminar. (a) A more engaged student. (b) A less engaged student.

In other moments, the influencing factor bar chart had a higher occlusion value. This was because Student 4 put her hands in front of her, which caused occlusion (R4). Sometimes, the girl could not even be detected. The professor also explored another interesting student, Student 12, as his stacked emotion bar (Fig. 9b) was much shorter than others, indicating that he appeared in much fewer frames of the whole video (R2). After checking the details in the emotion flow view and video view, the professor quickly understood the reason for this: Student 12 seemed distracted and he was always bowed down and looking at his notebook (R5). This made his face not able to be detected (Fig. 9b).

Overall, the professor said that she was excited to use *EmotionCues* for analyzing the seminar videos and confirmed the usefulness of *EmotionCues*. She also pointed out some limitations of *EmotionCues*. For example, some emotion labels need to be interpreted in context. It would be more helpful if *EmotionCues* can directly visualize the engagement of students in a classroom.

7 Interviews and Feedback

To further evaluate the usefulness and effectiveness of our system, we also conducted semi-structured interviews. We interviewed both end users and domain experts to extensively evaluate our system. The end users are our four aforementioned collaborating end users (U1-4), while domain experts are three experts (E1-3) with background knowledge about our target scenarios. Specifically, E1 is a CEO of a company and has worked closely with kindergartens for more than five years. E2 is a university professor, who has been working in education and childcare for more than 20 years. E3 is a university professor and his research interests mainly focus on the engagement of classes and online education. None of the experts have been involved in the design of *EmotionCues*. Each interview lasted for about an hour. We first briefly introduced *EmotionCues*. Then we went through an example to explain the functions and visual encoding of *EmotionCues*. After that, we allowed them to freely explore our system in a think-aloud protocol. Considering the key steps of exploring videos with *EmotionCues*, we designed the following tasks for their free exploration:

- Describe the overall emotion distribution and emotion evolution in the video.
- Select a person of interest and explore how his/her emotions evolve over time.
- Find a person who is most different from the person you selected in terms of emotion distribution and emotion evolution and explore the major reasons for the differences.
- Correct some inaccurate recognition results through interactions with *EmotionCues*.

During the process, we conducted audio-recording with their permission, in order to accurately collect their feedback. Their feedback is summarized as follows:

Feedback from end users. They were satisfied with the visual designs of the system, since it was straightforward and met their requirements. They also expressed that the system was easy for them to use. Further, they all agreed that this system can help them better analyze students' emotion status and engagement from both collective and individual perspectives. As teachers, U1 and U2 needed to pay attention to the whole class, so they liked to use the summary view to observe the overall emotion status from the summary view. They highly appreciated the design of the offline band to represent some students who were not captured by the cameras. What is more, both U1 and U2 encouraged us to capture more in-class behaviors of students, such as focus and distraction. As parents, U3 and U4 paid more attention to exploring children's portraits and comparing different children in the character view. At first glance, they thought the portrait glyphs may be a bit complicated. However, after using *EmotionCues* for a while, they felt that it was easy to understand and use. Also, the snapshot function was useful for comparing the emotion distributions of different children.

Feedback from domain experts. All the experts appreciated the idea of summarizing classroom videos from an emotion perspective. Both E1 and E2 highly appreciated the idea that revealing the model uncertainty with some influencing factors, which can give users some useful hints about the model performance and children's situations. E3 would like to analyze the engagement and teamwork of students in a flipped classroom with such a system. They gave some further suggestions for improvements. For example, they encouraged us to consider other information in the videos, such as sound and movement. E3 recommended a search function for quickly finding students of interest, as well as an interaction to automatically cluster similar portraits for comparison. E2 suggested improving the layout of the lines in the emotion flow, so that visual clutter can be well reduced when multiple children are selected.

8 Discussion and Limitations

The core research question we focus on in this paper is to explore how visualization techniques can be used to help teachers and parents better analyze students' emotion status and their engagement in classroom videos. By working closely with the end users, we compiled a list of detailed user requirements, which further guides the design of our visualization system, *EmotionCues*. It combines automatic algorithms with effective visual designs, where the automatic algorithms extract emotion data from classroom videos and the visualization interface enables interactive exploration. Two use cases and interviews with both end users and domain experts provide support for the usefulness and effectiveness of our proposed system. However, there are still some issues that need further discussion and clarification.

Privacy Issue. We need to carefully consider and balance the tradeoff between positive use cases and abuse of classroom videos. One of the potential issues of analyzing classroom videos is privacy protection. For example, there have been some discussions on whether the video cameras should be installed in the classroom[3], which is beyond the scope of this paper. Though we gain the permission from the kindergartens and parents for using the kindergarten videos for research purposes, we have also tried our best to protect the privacy. For example, we have tried not to reveal the identity information of each child, as shown in Figs. 3,

3. http://www.debate.org/opinions/should-classrooms-have-video-cameras

6, and 7. In real applications, the privacy protection should be achieved through strict control on the access to the video data. For example, each parent can only be granted to access the video and the corresponding extracted information of their own children.

Model Performance. *EmotionCues* is built on the existing algorithms of face detection, face recognition and emotion recognition, and employs their results for further visual analysis. Thus, the performance of the algorithms has an influence on the effectiveness and usability of *EmotionCues*. Since most datasets mainly contain non-children and non-infant images, we further fine-tune the model with an emotion dataset of children (1000 images in total), to improve the model in recognizing children's emotions. The dataset is collected and labeled by the collaborating kindergartens. Each image is labeled by three teachers and the emotion for each face is annotated as the emotion tag labeled by most teachers. In the fine-tuning procedure, we augment the dataset by using some widely-used data augmentation techniques, e.g., rotation, shifting and scaling. We remove the last fully-connected layer (output layer), run the pre-trained model as a fixed feature extractor and then use the resulting features to train a new classifier with the small dataset. Also, before finally choosing the current algorithms, we sampled frames from the current video datasets and manually label them to verify their performance. Specifically, we evenly sample 100 frames from each video introduced in Sections 6.1 and 6.2, where about 10$\sim$15 persons appear in each frame. For face detection, our results show that our approach can correctly detect 1040 faces out of 1084 faces (an accuracy of 95.9%) in the kindergarten video and correctly detects 1188 faces out of 1234 faces (an accuracy of 96.3%) in the seminar video. For emotion recognition, our approach can achieve an accuracy of 64.8% and 68.5% in the kindergarten and seminar videos, respectively. With some more advanced models available, the current models can easily be replaced by other advanced models.

Emotion Recognition. We have carefully considered the situations that the emotions of students are not easily captured or recognized correctly. These situations are inevitable in real life. For example, in some classroom settings, such as participatory exercise and game activities, it is not easy to capture students' emotions correctly. Therefore, we have designed the offline band to represent those students' emotions which are not captured and provided some interactions to correct inaccurate labels. Further, we have integrated some factors that may adversely affect our emotion recognition to give users more hints on students' situations. Our evaluation results demonstrate the usefulness of these visual hints. However, how many details about the model uncertainty are "just right" for different users to trust different results is beyond the focus of this paper. What is more, we are also aware that children's facial expressions can provide parents and teachers with hints on the children's learning and emotion status, but it may not always accurately indicate their real thoughts. Better measurements that consider more factors could be explored in the future. For example, it would be interesting to integrate other data channels, such as audio data and motion/gesture information, which can provide users with more details.

Scalability. In general, there are two scalability issues. First, the number of faces in the videos we showcased is about 15. Too many faces in a single video tend to cause inevitable visual clutter due to limited screen space, especially in the emotion flow view. We plan to explore more interactions besides the current filtering. Second, longer videos tend to generate a more complicated emotion flow graph. We may split the longer videos into several parts via user interaction or computer vision techniques so that users can concentrate on those parts with potential patterns, e.g., wide mood swings. Also, it might be a good idea to adopt a semantic sampling method to extract emotions from a video instead of even sampling.

Generality. It is better to involve more end users to further improve and evaluate our system. In addition, *EmotionCues* is proposed for analyzing emotions in classroom videos, but it is not limited to classroom videos and can be extended to other similar application scenarios, such as analyzing the emotions of characters in a movie, the players' emotions in a contest, and the patients' situations in psychological settings like group or individual therapies. However, it may not be suitable for those general videos where few people appear or it is hard to capture people's emotions due to low resolution.

9 Conclusion and Future Work

We have presented *EmotionCues*, a visual analytics system for effectively summarizing emotions in a given classroom video. To tackle the challenge of unperfected recognition algorithms, a novel visual encoding of emotion portraits, incorporating several important factors for revealing the model uncertainty, is proposed to provide an insightful interpretation of classroom videos. In addition, some well-established techniques, such as stacked bar charts and flow-based graphs, are extended and integrated into our system, supporting visual analysis of students' emotion evolution from both collective and individual perspectives. A rich set of interactions is also provided for flexible visual exploration. Two use cases and interviews with the end users and domain experts show that the system enables users to explore students' emotions in classroom videos.

There are multiple venues for future work. First, we plan to incorporate more influencing factors into our system to help better understand the video content from the perspective of emotion. For example, the head pose including yaw, roll, and pitch angles will be helpful for understanding a student's emotion status. In addition, we would like to further improve the proposed system by better integrating it with more advanced emotion analysis algorithms. Furthermore, we plan to extend our work to other kinds of videos such as films and presentation videos.

Acknowledgment

The authors would like to thank the kindergartens for granting the use of the videos in this paper, and the end users and domain experts helping with the interviews, as well as the anonymous reviewers for their valuable comments. This research partially is supported by ITF grant UIT/138 and the Theme-based Research Scheme of the Hong Kong Research Grants Council under grant T44-707/16-N.

REFERENCES

[1] D. K. Meyer and J. C. Turner, "Discovering emotion in classroom motivation research," *Educational Psychologist*, vol. 37, no. 2, pp. 107–114, 2002.

[2] R. Pekrun, "Emotions and learning," Educational Practices Series, International Academy of Education and International Bureau of Education, 2014, [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000227679. [Accessed Jan. 1, 2020].

[3] F. Zhang, P. Markopoulos, and T. Bekker, "The role of children's emotions during design-based learning activity: A case study at a dutch high school," in *International Conference on Computer Supported Education*, 2018.

[4] M. Balaam, G. Fitzpatrick, J. Good, and R. Luckin, "Exploring affective technologies for the classroom with the subtle stone," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 1623–1632.

[5] S. DMello, T. Jackson, S. Craig, B. Morgan, P. Chipman, H. White, N. Person, B. Kort, R. el Kaliouby, R. Picard *et al.*, "Autotutor detects and responds to learners affective and cognitive states," in *Workshop on Emotional and Cognitive Issues at the International Conference on Intelligent Tutoring Systems*, 2008, pp. 306–308.

[6] E. Skinner, J. Pitzer, and H. Brule, "The role of emotion in engagement, coping, and the development of motivational resilience," *International Handbook of Emotions in Education*, pp. 331–347, 2014.

[7] D. Kilburn, "Methods for recording video in the classroom: producing single and multi-camera videos for research into teaching and learning," 2014, National Center for Research Methods Working Paper. [Online]. Available: http://eprints.ncrm.ac.uk/3599. [Accessed Jan. 1, 2020].

[8] G. Daniel and M. Chen, "Video visualization," in *Proceedings of IEEE Visualization*, 2003, p. 54.

[9] Y. S. Khan and S. Pawar, "Video summarization: survey on event detection and summarization in soccer videos," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 11, pp. 256–259, 2015.

[10] M. Ajmal, M. H. Ashraf, M. Shakir, Y. Abbas, and F. A. Shah, "Video summarization: techniques and classification," in *Proceedings of International Conference on Computer Vision and Graphics*, 2012, pp. 1–13.

[11] S. Li and W. Deng, "Deep facial expression recognition: A survey," 2018, *arXiv:1804.08348*. [Online]. Available: https://arxiv.org/abs/1804.08348. [Accessed Jan. 1, 2020].

[12] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenges," in *Proceedings of ACM International Conference on Multimodal Interaction*, 2016, pp. 427–432.

[13] C. M. Tyng, H. U. Amin, M. N. Saad, and A. S. Malik, "The influences of emotion on learning and memory," *Frontiers in Psychology*, vol. 8, p. 1454, 2017.

[14] B. Rienties and B. A. Rivers, "Measuring and understanding learner emotions: Evidence and prospects," *Learning Analytics Review*, vol. 1, pp. 1–28, 2014.

[15] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," University of Texas at Austin, Austin, TX, Tech. Rep., 2015, [Online]. Available: https://eprints.lancs.ac.uk/id/eprint/134191. [Accessed Jan. 1, 2020].

[16] N. J. Butko, G. Theocharous, M. Philipose, and J. R. Movellan, "Automated facial affect analysis for one-on-one tutoring applications," in *Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 382–387.

[17] J. Hernandez, M. E. Hoque, W. Drevo, and R. W. Picard, "Mood meter: counting smiles in the wild," in *Proceedings of the ACM Conference on Ubiquitous Computing*, 2012, pp. 301–310.

[18] R. Srivastava, S. Yan, T. Sim, and S. Roy, "Recognizing emotions of characters in movies," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 993–996.

[19] 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 IEEE Conference on Visual Analytics Science and Technology*, 2014, pp. 203–212.

[20] R. Kempter, V. Sintsova, C. Musat, and P. Pu, "Emotionwatch: Visualizing fine-grained emotions in event-related tweets," in *Proceedings of International AAAI Conference on Weblogs and Social Media*, 2014, pp. 236–245.

[21] R. Borgo, M. Chen, B. Daubney, E. Grundy, G. Heidemann, B. Höferlin, M. Höferlin, H. Jänicke, D. Weiskopf, and X. Xie, "A survey on video-based graphics and video visualization," in *Proceedings of Eurographics (STARs)*, 2011, pp. 1–23.

[22] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Transactions on Graphics*, vol. 27, no. 3, p. 16, 2008.

[23] M. L. Parry, P. A. Legg, D. H. Chung, I. W. Griffiths, and M. Chen, "Hierarchical event selection for video storyboards with a case study on snooker video visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 1747–1756, 2011.

[24] M. Stein, H. Janetzko, A. Lamprecht, T. Breitkreutz, P. Zimmermann, B. Goldlcke, T. Schreck, G. Andrienko, M. Grossniklaus, and D. A. Keim, "Bring it to the pitch: Combining video and movement data to enhance team sport analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 13–22, 2018.

[25] H. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, and H. Qu, "Emoco: Visual analysis of emotion coherence in presentation videos," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 927–937, 2020.

[26] 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*, vol. 19, no. 12, pp. 2119–2128, 2013.

[27] A. Girgensohn, D. Kimber, J. Vaughan, T. Yang, F. Shipman, T. Turner, E. Rieffel, L. Wilcox, F. Chen, and T. Dunnigan, "Dots: support for effective video surveillance," in *Proceedings of ACM International Conference on Multimedia*, 2007, pp. 423–432.

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

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

[30] G. Ramos and R. Balakrishnan, "Fluid interaction techniques for the control and annotation of digital video," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, 2003, pp. 105–114.

[31] K. Schoeffmann and L. Boeszoermenyi, "Video browsing using interactive navigation summaries," in *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, 2009, pp. 243–248.

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

[33] J. Matejka, T. Grossman, and G. Fitzmaurice, "Swifter: improved online video scrubbing," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 1159–1168.

[34] H.-W. Kang, Y. Matsushita, X. Tang, and X.-Q. Chen, "Space-time video montage," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1331–1338.

[35] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, p. 3, 2007.

[36] M. Chen, R. Botchen, R. Hashim, D. Weiskopf, T. Ertl, and I. Thornton, "Visual signatures in video visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, 2006.

[37] M. Höferlin, B. Höferlin, D. Weiskopf, and G. Heidemann, "Uncertainty-aware video visual analytics of tracked moving objects," *Journal of Spatial Information Science*, vol. 2011, no. 2, pp. 87–117, 2011.

[38] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, "The role of uncertainty, awareness, and trust in visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 240–249, 2016.

[39] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz, "Overview and state-of-the-art of uncertainty visualization," in *Scientific Visualization*. Springer, 2014, pp. 3–27.

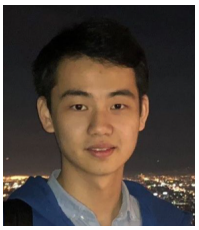
[40] C. D. Correa, Y.-H. Chan, and K.-L. Ma, "A framework for uncertainty-aware visual analytics," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2009, pp. 51–58.

[41] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski,

"Visualizing time-oriented data – A systematic view," *Computers & Graphics*, vol. 31, no. 3, pp. 401–409, 2007.
[42] G. Andrienko, N. Andrienko, U. Demsar, D. Dransch, J. Dykes, S. I. Fabrikant, M. Jern, M.-J. Kraak, H. Schumann, and C. Tominski, "Space, time and visual analytics," *International Journal of Geographical Information Science*, vol. 24, no. 10, pp. 1577–1600, 2010.
[43] S. Liu, Y. Wu, E. Wei, M. Liu, and Y. Liu, "Storyflow: Tracking the evolution of stories," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2436–2445, 2013.
[44] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "Textflow: Towards better understanding of evolving topics in text," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2412–2421, 2011.
[45] Y. Zheng, W. Wu, H. Zeng, N. Cao, H. Qu, M. Yuan, J. Zeng, and L. M. Ni, "Telcoflow: Visual exploration of collective behaviors based on telco data," in *Proceedings of IEEE International Conference on Big Data*, 2016, pp. 843–852.
[46] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[47] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.
[48] J. Dutta and S. C. Pal, "A note on hungarian method for solving assignment problem," *Journal of Information and Optimization Sciences*, vol. 36, no. 5, pp. 451–459, 2015.
[49] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
[50] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.
[51] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2009, pp. 566–569.
[52] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.
[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[54] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proceedings of International Conference on Neural Information Processing*, 2013, pp. 117–124.
[55] L. J. Wells, S. M. Gillespie, and P. Rotshtein, "Identification of emotional facial expressions: effects of expression, intensity, and sex on eye gaze," *PloS One*, vol. 11, no. 12, p. e0168307, 2016.
[56] K. Yuen and M. M. Trivedi, "An occluded stacked hourglass approach to facial landmark localization and occlusion estimation," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 4, pp. 321–331, 2017.
[57] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings of IEEE Symposium on Visual Languages*, 1996, pp. 336–343.
[58] T. Munzner, "A nested process model for visualization design and validation," *IEEE Transactions on Visualization & Computer Graphics*, no. 6, pp. 921–928, 2009.
[59] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, "Lineup: Visual analysis of multi-attribute rankings," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2277–2286, 2013.
[60] W. Wu, J. Xu, H. Zeng, Y. Zheng, H. Qu, B. Ni, M. Yuan, and L. M. Ni, "Telcovis: Visual exploration of co-occurrence in urban human mobility based on telco data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 935–944, 2016.

Haipeng Zeng is currently a Ph.D. candidate in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST). He obtained a B.S. in Mathematics from Sun Yat-Sen University,China in 2014. His research interests include data visualization, visual analytics and video analysis.

Xinhuan Shu is currently a Ph.D. student in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST). She received her B.E. degree in Computer Science and Technology form Zhejiang University, China in 2017. Her research interests include information visualization, visual analytics, and sports visualization.

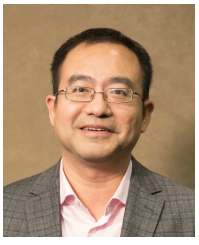


Yanbang Wang is an undergraduate in Department of Computer Science and Engineering at Hong Kong University of Science and Technology, doubly majoring in computer science and mathematics. His current research interest is in big data mining for knowledge extraction and analytics, data visualization, and machine learning and its application to real-world problems.

Yong Wang is currently a postdoctoral fellow in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology, where he obtained his Ph.D. in 2018. He received his B.E. degree in Automation from Harbin Institute of Technology, China in 2011 and M.E. degree in Pattern Recognition and Intelligent system from Huazhong University of Science and Technology, China in 2014. His research interests include data visualization, visual analytics and image processing.

Liguo Zhang is an associate professor at the College of Computer Science and Technology, Harbin Engineering University. He received his B.S. in Automation from University of Science and Technology Beijing (USTB), China in 2004, and Ph.D. degree in Computer Science and Technology from University of Chinese Academy of Sciences (UCAS), China in 2014. His main research interests are image processing and computer vision.

Ting-Chuen Pong is a professor of the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST). He received his Ph.D. in Computer Science from Virginia Polytechnic Institute and State University in 1984. His research interests include computer vision, multimedia computing and IT in Education. For more information, please visit http://www.cse.ust.hk/faculty/tcpong/.



Huamin Qu is a professor in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. His main research interests are in visualization and computer graphics, with focuses on urban informatics, social network analysis, e-learning, and text visualization. He obtained a B.S. in Mathematics from Xi'an Jiaotong University, China, an M.S. and a Ph.D. in Computer Science from the Stony Brook University. For more information, please visit http://www.huamin.org.