Detecting Negative Emotion for Mixed Initiative Visual Analytics

Prateek Panwar
University of Ontario
Institute of Technology
Oshawa, ON, Canada
prateek.panwar@uoit.ca

Christopher Collins
University of Ontario
Institute of Technology
Oshawa, ON, Canada
christopher.collins@uoit.ca

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada
ACM 978-1-4503-5621-3/18/04.
https://doi.org/10.1145/3170427.3188664

Abstract

The paper describes an efficient model to detect negative mind states caused by visual analytics tasks. We have developed a method for collecting data from multiple sensors, including GSR and eye-tracking, and quickly generating labelled training data for the machine learning model. Using this method we have created a dataset from 28 participants carrying out intentionally difficult visualization tasks. We have concluded the paper by a discussing the best performing model, Random Forest, and its future applications for providing just-in-time assistance for visual analytics.

Author Keywords

Data Analysis; Emotion Detection; GSR; Eye Tracking

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Introduction

Data analysis is a challenging task which can become frustrating when dealing with an unfamiliar dataset or a new analysis tool or visualization. Also, working in a real-world scenario with deadlines can increase the cognitive load of the user. As a result, the likelihood of disengagement or making mistakes increases. Therefore, the motivation of this work is to detect these negative emotions so that, in future, we could use this model to provide appropriate interventions to help the analysts on-the-fly. For example, if a user feels confused or frustrated during the task then the system should be able to detect it and provide a suitable help appropriate to the cause of frustration.

Experimental Setup

Room: The user study was conducted in a private room with controlled lighting. The researchers observed the participants from an adjacent room to minimize pressure or distraction due to observation.

Hardware: Shimmer GSR (5Hz) was used to record the skin conductivity and Tobii X60 eye tracker (60Hz) was used for tracking gaze information.

Signals: Skin conductance, pupil diameter and gaze location.

Visualization Tool: Pivot-Slice[14] is an open source toolkit which uses scatter plot-like representation for visualizing multidimensional data. The interface allows users to customize the outputs by applying multiple filters. We have modified the user interface according to our study design (Fig 3).

Tasks: 2 practice and 4 main analytics tasks. Main tasks were designed in increasing level of difficulty.

A user's emotional state can be detected by capturing the data from physiological sensors (e.g. electroencephalogram (EEG), electrocardiogram (ECG) and, skin conductance) or physical sensors (e.g. facial expression, speech, body posture and gaze tracking) or a combination of both. When selecting between various biometric sensor options, we prioritized those which were less intrusive and less likely to cause discomfort. Finally, we decided to use the combination of a galvanic skin response device (GSR) and an eye tracker to measure arousal and valence.

It is likely that in analytics tasks, both the detected features (e.g., gaze scan-paths) and the related emotional intensities will be different than when reading a text, looking at an image, or playing a video game. Therefore, a specific model for this scenario is required. Using two types of sensors, we created a dataset and model from 28 participants performing visual analytics tasks with a visualization interface called PivotSlice [14] (see Figure 3).

Related Work

Past research in affective computing has investigated detecting negative emotions or disengagement such as mind-wandering (MW), stress and frustration. Bixler et al. [1] demonstrated a technique for detecting MW using 3 types of gaze features for a supervised classifier: global, local and contextual. They achieved 72% accuracy in detecting MW during reading. Thus, an eye tracker is an unobtrusive but effective device for detecting mind states. Moreover, a survey by Sharma et al. [11] compared the performance of different signals and algorithms for identifying stress. They compared 13 possible physiological and physical signals and concluded that heart rate variance (HRV) outperforms every other signal for stress detection, followed by EEG then GSR. For our study, we investigated these signals and finalized on GSR as it is very easy to setup and gives the least amount of noise. Many works use a combination of biometric devices to achieve better accuracy, and have proposed on-the-fly models [8, 9].

Voigt et al. [12] have discussed some of the challenges of data scale and proposed a context-aware recommendation algorithm which leverages online annotations to provide help. Similarly, Gotz et al. [5] generate recommendations in visualization, driven by the user behaviour (implicate signals) and successfully reduced overall task completion times and errors. There are other papers which predicted user's learning curve on-the-fly and provided help accordingly for visualization tasks [2, 7]. Finally, Hung et al. [6] assert the importance of user engagement in information visualization and propose a self-assessment questionnaire.

In initial investigations, we found that eye-tracking alone was susceptible to loss of data in long-duration tasks due to head movement. Based on this and the results of previous studies, we selected GSR and eye tracking as our sensors. The combination of these two devices was likely to provide stable data for the detection of emotional arousal and valence with minimal physical intervention. Most research in the area of emotionally-responsive interfaces target applications such as gaming, intelligent tutoring system and on-screen reading. We build on past work in eye tracking and visualization [4] to build a multi-sensor system to detect negative mind states specifically for visualization tasks.

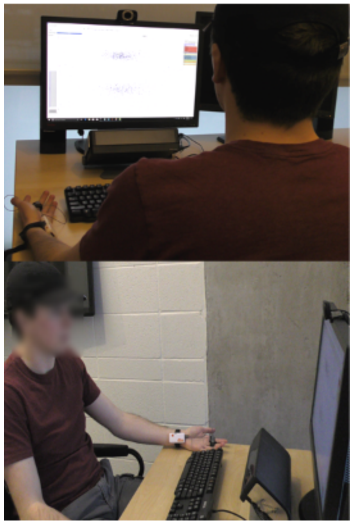


Figure 1: Experimental setup.

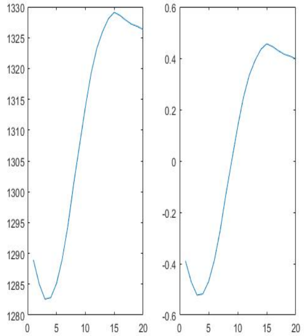


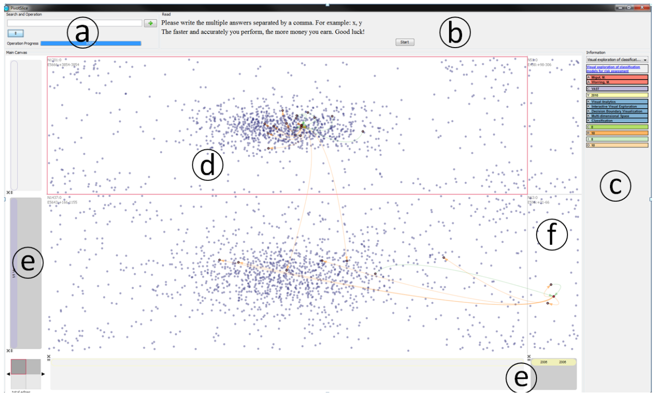
Figure 2: GSR signal. Left: raw data (x); Right: standardized data using $x_{std} = (x - \mu)/\sigma$; μ and σ are the mean and standard deviation of the window.

Data Collection Study

After refining the setup in a pilot study with 7 participants, we recruited 32 university students in the third year of study and higher, from our local faculty of science. Due to hardware issues, we had to discard the data from 4, leaving 28 final participants (26 male, 2 female, mean age 24). We note the imbalanced gender ratio could limit the generalizability, and will be addressed in future. Figure 1 shows the setup. To get the emotional responses closer to the real world scenario, we made the analysis tasks more engaging using gamification. Participants were instructed that their compensation depended on their performance (completion time and the number of errors).

The session was divided into three parts: introduction, performing the tasks, and retrospective think aloud. The introduction consisted of attaching and calibrating the GSR, followed by a brief introduction about the tasks and interface, PivotSlice. The introduction took approximately 13 minutes and allowed participants to get comfortable with the GSR device attached to their wrist. Next, participants completed two practice tasks. This step was essential to reduce measurement of emotional responses due to unfamiliarity with the interface. Also, a reference handout (2 pages) was available throughout the experiment in case the participant forgot any of the interface functions. Finally, participants carried 4 main tasks, in increasing order of complexity.

We needed to determine ground truth moments of negative emotion for training a classifier. However, asking participants to label 25 minutes of video would be tedious and error-prone. In pilot testing we developed a simplified classifier to detect time points in which the GSR signal shows significant change — these points became the candidate times for labelling emotion in a retrospective think aloud. Participants were shown video from the top 7 candidate time points and asked to narrate what was happening at that time, and specifically what their emotional state was. When a emotion was mentioned, all data windows in which the time point appears, and any other peaks in the data stream within 1 minute of the selected time point, were labelled with the emotion. The expansion of the labelled time period to include any adjacent peaks was used to increase the amount of accurately labelled data without requiring additional effort from the participant.



Please write the multiple answers separated by a comma. For example: x, y
The faster and accurately you perform, the more money you earn. Good luck!

Figure 3: PivotSlice interface [14]. (a) Search Panel, (b) Task Panel, (c) Information Panel , (d) Unfiltered data region, (e) Filter axes, (f) Filtered data region.

While we acknowledge that not annotating full dataset may affect performance, our hypothesis was that our method would capture sufficient data samples for appropriately recognizing significant emotional reactions. We have discussed the potential impacts of partial data labelling in a later section. As we are interested in negative emotions, we dropped all positive emotion events from our dataset.

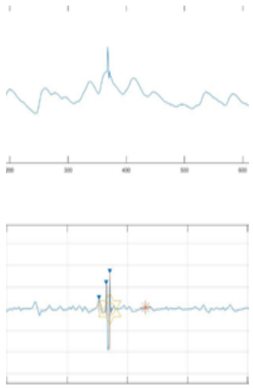


Figure 4: Top: Standardized GSR data; Bottom: peak detected in the signal using Lagrangian interpolation.

R	Feature
1	Std GSR (Lagrangian)
2	Mean fixation duration
3	Mean PD (Lagrangian)
4	Std PD (Lagrangian)
5	Std GSR (zscore)
6	No. of peaks PD (zscore)
7	Mean GSR (Lagrangian)

Table 1: Rank (R) of the final features.

Preprocessing and Feature Extraction

First, high-frequency noise from the GSR and pupil data were removed with a low-pass filter. Next, samples flagged by the Tobii SDK as low confidence (due to blinking and head movement) were counted (for future processing), then replaced using linear interpolation of adjacent values to generate the final dataset for the feature extraction.

Training and testing datasets were created by sampling feature values using a moving window across the data streams from the sensors. We tested our model with 5s and 10s window sizes, also, with different overlap values ranging from 50% to 75%. Finally, the best accuracy we achieved with a 10s window with 60% overlap. 21 features were extracted from the dataset: 8 from the pupil size, 5 from gaze location, and 8 from the GSR data. Details about each feature are in the following subsection.

Pupil Dilation (PD) and Skin Conductance (GSR)

Raw data is not generalizable as every individual has a different baseline and therefore cannot be used for training. We used a standardization method by calculating the z-score to scale the baseline without losing the nature of the data (Figure 2). This technique is an alternative to normalization and works best in sliding windows. After scaling the data, total 4 features from each signal were calculated — mean, standard deviation, number of peaks and sum of the height of the peaks. The data is linear and any change in the baseline would affect these values.

Next, we calculated 7-point second-order Lagrangian interpolation for finding the points where the signals changed relative to the baseline [13] (Figure 4):

$$g''[n] = 2000 * \frac{2g[n+3]+g[n+2]-2g[n+1]-2g[n]-2g[n-1]+g[n-2]+2g[n-3]}{20h^2}$$

Here, $g[n]$ represents the value at index n in the standardized list and h is the sampling frequency of the device. There are several other algorithms that can be used for the same such as wavelet transformation, Fourier transformation and CUSUM algorithm. Again, the 4 features; mean, standard deviation, number of peaks and sum of the height of the peaks; were calculated.

Gaze Location Information

We used the gaze location information to calculate the focus point, called a fixation, and the distance between two focus points is a saccade. Just as prior work on reading targeted sequential fixations as indicative of that task [1], we target features which may be useful for visualization. For example, too many saccades and fewer fixations may indicate visual searching related to confusion or frustration. We applied the method of Olsson [10] to calculate fixations and saccades between two consecutive sliding inner windows.

$$m_{before}(n) = \left[\frac{1}{r}\sum_{k=1}^{r} s_x(n-k), \frac{1}{r}\sum_{k=1}^{r} s_y(n-k)\right]$$

$$m_{after}(n) = \left[\frac{1}{r}\sum_{k=1}^{r} s_x(n+k), \frac{1}{r}\sum_{k=1}^{r} s_y(n+k)\right]$$

where s_x represents gaze location in x axis and s_y is gaze location in y axis in a window, n is the sample of interest, and r is the inner window size (5s). The distance d was calculated at every sample in the eye tracking data by,

d(n) = $\sqrt{(m_{after}(n) - m_{before}(n)).(m_{after}(n) - m_{before}(n))^T}$

where d is the gaze distance in window n. All distance values which were more than the standard deviation of distances were marked as saccades. Fixations were calculated by finding the median of the samples between 2 saccades. After calculation of the fixations and saccades, we derived 4 features: number of fixations, mean saccade length, mean fixation duration and standard deviation of

Models	Recall score
KNN	55%
SVM	67%
RF	88%

Table 2: Recall score of the event class with different classification models.

$$\begin{bmatrix}171 & 25\\ 1 & 44\end{bmatrix} \quad \begin{bmatrix}137 & 45\\ 0 & 28\end{bmatrix}$$

$$\begin{bmatrix}199 & 39\\ 2 & 25\end{bmatrix} \quad \begin{bmatrix}236 & 16\\ 2 & 20\end{bmatrix}$$

Table 3: Confusion matrices of RF classification from 4 different participants. 1st column represents non-event and 2nd column is negative response.

Feature	Non-Event	Event
σ_{GSR}	0.066	1.787
	0.403	1.636
	0.228	1.045
MFD	98.166	30.947
	53.545	21.071
	117.60	34.764

Table 4: Samples of noticeable changes in standard deviation of GSR and mean fixation duration (MFD) from non-event to event class in case of frustration in different participants.

fixation points from the centroid of fixations in the window (indicating the extent of the area of interest).

The final feature was the total number of samples where the eye tracking failed (number of blinking and head movement samples). A high value for this feature might indicate that the participant is not looking at the screen.

We wanted our model to be fast, as we will use this model in the future to classify states on-the-fly. To reduce the feature set size, we used a brute force technique to test increasingly large feature combinations. After a certain number of combinations, the accuracy remained approximately constant. All remaining features were dropped. The ranking of the 7 final features is shown in Table 1.

Classification Model and Results

We tested our dataset with several supervised classification models using n-fold cross-validation where n is the total number of participants. The classification models were built in python using the *scikit-learn* library — k nearest neighbour (KNN), support vector machine (SVM) and random forest (RF). We also tested the dataset with neural networks, but it failed due to lack of enough training data.

As most of the time people are not exhibiting a strong negative emotion, our event and non-event classes were unbalanced (1:12), which can affect classifier accuracy. To solve this rare event detection problem, we generated artificial data for balancing the classes using the Synthetic Minority Over-sampling Technique (SMOTE) [3]. We only applied SMOTE algorithm to adjust the training data and kept the original unbalanced data for testing. Since the test data was similarly imbalanced, rather than simple accuracy (which would be high even for a naive classifier), we report the confusion matrices and negative emotion recall scores.

We achieved the best recall score (high detection of true negative emotion states), 88%, using a random forest classifier with 1,000 trees and max depth 4. Table 2 shows the recall scores of all the classifiers we tested. Table 3 shows 4 confusion matrices from 4 random participants. Using the first example from the table, 171 are the non-event (true negative), 25 are false positive, 1 is false negative and 44 represent negative emotions (true positive). The higher false positive rate indicates that positive events were missed in training data as they were not in the top 7 events discussed in the think-aloud session. As we envision an application providing a variety of types of assistance, providing unwanted help (false positive) is less problematic, if designed well, than missing potential moments of frustration and confusion (false negative).

Discussion and Work-in-Progress

To understand our model, we investigated some frustration events in detail. 23 participants reported frustration in the final task, which was designed to be difficult. We found some noticeable changes in feature values during these periods, which contributed to our model detecting negative states. Table 4 shows some σ_{GSR} signals and mean fixation durations during transitions from non-event (regular) to event (negative emotion) states for the final task. While these features alone were not sufficient for classifier performance, they provide some insight into the feature changes during frustration events: participants arousal level increases along with the decrease in the mean fixation duration.

This work builds the foundation for an on-the-fly system to identify negative emotions. In ongoing work, we are using this information to explore the design space of appropriate interventions to ease frustration in visualization tasks. For example, gaze information can be used to guess the source of frustration, such as the interface widgets or some subset of data items. Interventions may range from providing helpful hints about interface components to suggesting new subsets of data to explore. We aim to create a mixed-initiative interaction mechanism to help to the user and prevent disengagement due to frustration.

Acknowledgements

This work was supported by grants from NSERC.

REFERENCES

1. Robert Bixler and Sidney D'Mello. 2014. Toward fully automated person-independent detection of mind wandering. In *Proc. Int. Conf. on User Modeling, Adaptation, and Personalization*. 37–48.

2. Eli T Brown, Alvitta Ottley, Helen Zhao, Quan Lin, Richard Souvenir, Alex Endert, and Remco Chang. 2014. Finding Waldo: Learning about users from their interactions. *IEEE Trans. Vis. Comput. Graphics* 20, 12 (2014), 1663–1672.

3. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *J. Artificial Intelligence Research* 16 (2002), 321–357.

4. Cristina Conati, Enamul Hoque, Dereck Toker, and Ben Steichen. 2013. When to Adapt: Detecting User's Confusion During Visualization Processing. In *UMAP Workshops*.

5. David Gotz and Zhen Wen. 2009. Behavior-driven visualization recommendation. In *Proc. ACM Int. Conf. on Intelligent User Interfaces*. 315–324.

6. Ya-Hsin Hung and Paul Parsons. 2017. Assessing user engagement in information visualization. In *Proc. CHI Conf. Extended Abstracts*. 1708–1717.

7. Sébastien Lallé, Dereck Toker, Cristina Conati, and Giuseppe Carenini. 2015. Prediction of users' learning curves for adaptation while using an information visualization. In *Proc. ACM Int. Conf. on Intelligent User Interfaces*. 357–368.

8. Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. 2015. Combining eye movements and EEG to enhance emotion recognition. In *IJCAI*. 1170–1176.

9. Judi McCuaig, Mike Pearlstein, and Andrew Judd. 2010. Detecting learner frustration: Towards mainstream use cases. In *Proc. Int. Conf. on Intelligent Tutoring Systems*. 21–30.

10. Pontus Olsson. *Real-time and offline filters for eye tracking*. Master's thesis. KTH Electrical Engineering.

11. Nandita Sharma and Tom Gedeon. 2012. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer Methods and Programs in Biomedicine* 108, 3 (2012), 1287–1301.

12. Martin Voigt, Stefan Pietschmann, Lars Grammel, and Klaus Meißner. 2012. Context-aware recommendation of visualization components. In *Proc. Int. Conf. on Info., Process, and Knowledge Management*. 101–109.

13. Jing Zhai, Armando B Barreto, Craig Chin, and Chao Li. 2005. Realization of stress detection using psychophysiological signals for improvement of human-computer interactions. In *Proc. IEEE SoutheastCon*. 415–420.

14. Jian Zhao, Christopher Collins, Fanny Chevalier, and Ravin Balakrishnan. 2013. Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE Trans. Vis. Comput. Graphics* 19, 12 (2013), 2080–2089.