Recognizing emotions with technology and starting discussions

Research in Emerging Technologies 2017-2018, final paper

Randall Theuns
Amsterdam University of Applied Sciences
Software Engineering
Amsterdam, The Netherlands
randall.theuns@hva.nl

Max Beije
Amsterdam University of Applied Sciences
Business IT & Management
Amsterdam, The Netherlands
max.beije@hva.nl

Frank Portengen
Amsterdam University of Applied Sciences
Business IT & Management
Amsterdam, The Netherlands
frank.portengen@hva.nl

Abstract

According to Waag Society and the Research Group Crossmedia, recent studies have shown that young adults are hard to reach when it comes down to (cultural) heritage. Waag Society is researching how cultural heritage institutions can connect to these groups and how heritage objects can be relevant to (young) people. Both parties believe that a better understanding of the emotions people have, is very important to learn more about the way people value cultural heritage. Therefore, Waag Society and the Research Group Crossmedia asked students of the HvA to design an interactive tool that captures young adults emotions and enables them to discuss these emotions with their peers when looking at (cultural) heritage.

1. Keywords

Facial recognition, prototype, emotions, heritage, discussion, expressions

2. Introduction

During the minor 'Research in Emerging Technologies', we've chosen the project 'Emotions in Heritage' that is commissioned by the Research Group Crossmedia (HvA) and the Waag Society. The Waag Society explores emerging technologies not only related to the internet, but also related to biotechnology and cognitive sciences. Art and culture often plays a central role in our research as well.

Our project involves systematically collecting the different emotions people experience when they perceive (cultural) heritage, and use this information to spark meaningful conversations between different parties by visualizing the different emotions. Cultural heritage could, for example, mean a painting, a building, or even a tradition. Earlier research regarding emotion recognition has already been conducted. However, these researches were mostly conducted from a psychological point of view; what emotions are being expressed and why do young adults express a specific emotion? As such, models to define the emotions already exist. Currently, there is a lack of instrumentation to capture these emotions and allow young adults to openly discuss their emotions regarding heritage.

There are many options available to recognize emotions. For example, facial expression recognition, voice recognition, text recognition, and wireless signals. Initially, we did research regarding these different methods. We found a paper [1] which contained the various recognition methods and also included the accuracy with which they were able to recognize emotions. Using these measurements, we made the choice to limit how our project will recognize emotions to the method with the highest accuracy. The result of this research concluded that facial expression recognition had the highest accuracy at an average of 94.48%

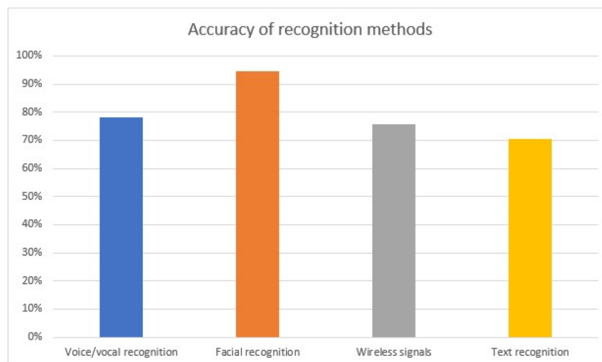

Figure 1: Accuracy of recognition methods

Besides accuracy, two more restraints are defined:

(1) If at all possible, all code and libraries used must be open source.
(2) The prototype has to display the predicted emotions in an interactive way.

Due to these requirements, and our choice to use facial expression recognition, the main research question of this paper is: *Does the visualization of emotions of young people (aged 16-26) by facial expression recognition software, lead to discussion regarding the displayed content?*

By working towards answering this question, we are developing a prototype that collects the relevant information that is required for the goal of the project: starting a meaningful conversation between young adults, regarding a specific subject. Furthermore, if the prototype proves to be useful and provides valid data, it might be used in subsequent research as a tool for collecting data.

3. Related work

Our project, as defined by the Waag Society, is already based on two previous research projects. The first project explores a method to sympathize with other people's emotions around heritage objects. Young people, teachers, and heritage professionals each map their emotions around a certain heritage topic and use that mapping as a starting point for discussion [6]. The goal of this research is very similar to our own, however the means to reach that goal differ, in that our research aims to use technology to read the emotions of the participants, as opposed to mapping them themselves.

The second project examines how participation, narratives, digital media, and atmosphere affect museum visitors when visiting an exhibition and how exhibition makers can influence these visitor experiences. One of the research questions is how people are emotionally affected when encountering one of these means []. Just like the first project, our research differs due to our focus on the technological aspect of detecting emotions.

Besides the above two project which were already defined at the start of the project, we've also found research conducted towards facial expression recognition [1]. This research focuses solely on detecting emotions using a camera. We also had the opportunity to try out the software created during this research. Unlike this research, our research also seeks to visualize these predicted emotions in a meaningful way in order to spark a discussion.

Lastly, we found a series of articles that aim to create open source facial expression recognition software [7]. All of the chosen libraries and datasets were open source or easily available. We have chosen these articles as the basis for our own prototype.

4. Material

Before creating the prototype, we were free in choosing *how* we were going to implement the emotion recognition. However, there were a couple restrictions and requirements. First, a very limited time span to create the prototype in. The prototype had to be up and running in roughly 8 weeks so that we could use it to perform our research. Secondly, preferable everything about the prototype had to be open source, as requested by the Waag Society. Because of prior desk-research, combined with the above restrictions, we chose to use facial expression recognition using a camera as the basis for our prototype. This decision was further solidified upon the discovery of several articles regarding open source facial expression software using Python, OpenCV, Dlib, and Facial Landmarks [7].

Aside from the facial expression recognition part of the project, the prototype had to be able to visualize the emotions to allow participants to reflect on their emotions and spark a discussion. The goal of the visualization is to show the user of the prototype their emotions over a certain time span, and allow the user to navigate through this time span. One of our clients, the Waag Society, had already done work regarding this subject, including the classification of emotions [5] which we will use, and a possible model for the visualization. We have used both the classification, as well as the model as a basis for our prototype. The emotions we have used in our prototype are as follows: *sadness, anger, contempt, disgust, fear, happiness, neutral, and surprise.*

In order to support our user research, the prototype has to go through a number of steps. Below, each of these steps is explained in more detail:

(1) Show a video to the research subject, and start filming at the same time. Store the video for analysis.
(2) Once the video is done, analyze the video and store the results of the analysis.
(3) Visualize the generated model on-screen when the user is ready.

The first step, recording and storing the video, is done with OpenCV[1] to allow video capture from the webcam. The prototype records at 10 frames per second (FPS). Next, the video captures are stored locally on disk. Displaying the video that the user of the prototype watches while he is being recorded, is considered part of the visualization, and will be discussed later in this chapter.

The analysis and prediction of emotion from video frames is largely the same as described in the series of articles written by Paul van Gent [7]. We use a trained, linear SVM machine learning model to predict the emotions on someone's face, based on 68 specific facial landmarks. The linear SVM machine learning model was created with scikit-learn[2], an open source machine learning library for Python. In order to train the SVM model, we needed a suitable dataset of images, labeled with the accompanying emotions. For this, we used the Cohn-Kanade and Extended Cohn-Kanade [2, 4] data sets. Each image in the dataset is preprocessed before being used to train the SVM model. This includes detecting a face on each image, cropping and resizing the face to a common size, removing all color, and using OpenCV's CLAHE algorithm. Once the preprocessing and sorting of the images is done, we use Dlib's [3] shape predictor to fetch the 68 specific facial landmarks. These landmarks, together with the label describing the emotion on the face, are then used to train the SVM model. The total dataset used to train the model contains 834 images. Using a larger dataset will likely result in a more accurate, stable, result.

The recorded video from our prototype is analyzed frame by frame. Each frame goes through the same preprocessing steps as the dataset used to train the SVM model. Once preprocessed, the 68 landmarks from each frame are then sent through the SVM model. Since this is a probabilistic model, the percentage chance for each of the 8 emotions is returned as a result. We found that these results could fluctuate fairly rapidly in a short number of frames. In order to get a more stable result, we've used the average of 5 each emotion over a span of 5 frames (half a second). This means that our final prototype predicts the emotions for every half a second of recorded video. Once the predictions are made for the entire video, the results are stored in an Sqlite database to be used later for the visualization. This also allows us to store the results for analysis later in our research.

[1] https://opencv.org
[2] https://scikit-learn.org
[3] https://dlib.net

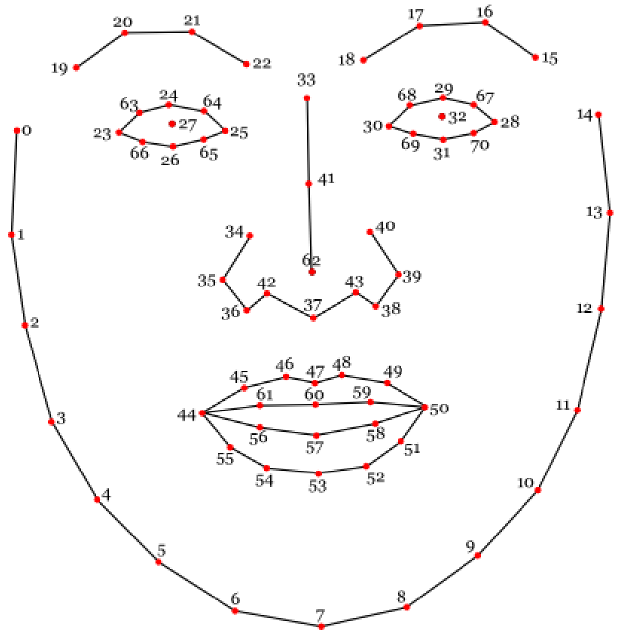

Figure 2: 68 facial landmarks

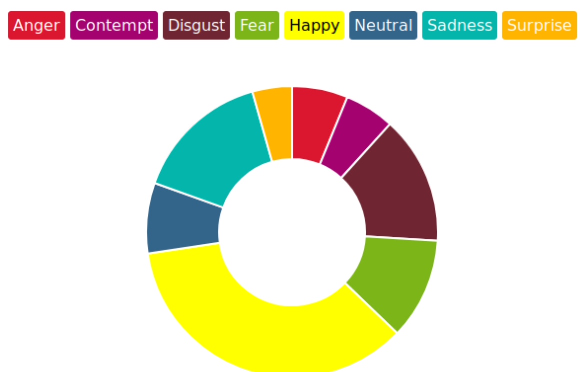

Figure 3: Visualized model as generated by the prototype

The visualization itself is a web application. A Flask[4] webserver is used to allow API access to the rest of the prototype. HTML, CSS and Javascript are then used to create the user interface for the prototype, including the visualization of the model. Initially, we wanted to use Qt[5], a C++ UI library, but due to a lack of time and expertise on our side, we decided to opt for something more familiar, which in this case resulted in a web application. The user interface consists of four pages. One to display the video (and start the recording), one to manage the sessions with our research participants, one to create a new session, and lastly, one to show the visualization. The visualization page uses doughnut charts for each recorded participants. The charts are generated using Chart.js[6], a Javascript library for creating a multitude of charts. The predicted emotions are displayed half a second at a time, with a time line slider to allow the participant to scroll through the predictions throughout the video.

[4]http://flask.pocoo.org

[5]https://qt.io

5. METHOD

After thoroughly testing the prototype on every aspect, working towards answering the main question was imminent. *Does the visualization of emotions of young people (aged 16-26) by facial expression recognition software, lead to discussion regarding the displayed content?*

The research that we performed was qualitative research. Our prototype was built for testing purposes, and not as an end-product. Therefore, we greatly value any feedback and improvement suggestions for our prototype from our research participants, as this can be used in future development. Because we are researching different emotional reactions of users, we're looking for very specific results. This means that our research design is very specific as well. The research design has the following requirements:

- Users of the prototype have to be in a closed-off room to prevent any distractions which can affect test results.
- Users of the prototype will not be given any information regarding the purpose or operation of the prototype, since this can affect their emotional reactions and therefore the test results.
- Users will be asked to use the prototype one-by-one.
- Users of the prototype will be asked a set of questions which will be focused on accuracy and improvement of the prototype.
- Each video shown by the prototype has a maximum length of one minute due to technical restrictions.

5.1. Participants

The participants are aged between 16-26 years old, and were scouted at the Amsterdam University of Applied Sciences. The participants were invited into a reserved, closed-off classroom, located at the Wibautstraat, where they were able to test the prototype. Our goal was to invite at least 20 participants. Unfortunately, due to technical difficulties and time restrictions, we've only managed to get 13 participants.

5.2. Content

The video had to provoke a measurable reaction from the subjects. Therefore, it is important that we show content that is doing just that. Research conducted by the Journal of Social Psychology [3] has shown that the strongest emotional reaction by people is caused by showing pictures of traumatic events. However, we wanted our subjects to show multiple emotions, and not just the ones that are associated with traumatic events, like *fear* and *disgust*. Therefore, we decided to show the subjects content that is, in any way, provoking.

[6]http://chartjs.org

We have selected three videos which have been edited to last roughly one minute:

(1) Roast of Giel Beelen (Peter Pannekoek roast)
(2) Asian longnecks: Why does this culture consider this beauty?
(3) Failarmy, best fails of 2017

6. Results

As described in section 5, we have conducted prototype testing sessions with 13 participants. For each of these participants, we have two pieces of data:

- Predicted emotions during the shown video, created by the prototype
- Results from the interview we had with each participant during the testing

Furthermore, for every of the three available videos, we have expectations regarding the emotions at certain times of the video. Using these pieces of data, we are interested in the following information from analyzing the results:

(1) Do the participants agree with the emotions predicted by the prototype?
(2) Do the predicted emotions coincide with the expectations we had for certain videos?

The first questions can be answered by cross-checking the data that the prototype generated, against the answers from the participants during the interview. The second question can be verified by checking the expected emotions at certain times in a video, against the average of all the participants who watched that video. This will make it clear whether the expected emotion is also the dominant emotion at that time. Finally, we can combine questions 1 and 2. If both results are positive; that is, the predicted emotions by the prototype match up with our expectations, and are valid according to the participant, then we can conclude that it's highly likely to be accurate.

6.1. Test results

We present a table with the results of the interview. The table gives an overview of the answers given by the participants regarding the accuracy of the prototype. Furthermore, several line charts are given which show the average emotions of all participants from each video, and the expected emotions as a dashed vertical line.

The analyzing of the data generated by the prototype was done using some custom code. This analyzer tool was created using Python and the matplotlib[7] library to generate the charts. The analysis tool is available as well: https://github.com/drtheuns/minor_riet_analysis

6.2. Expected emotions

For each video we had a number of expectations. These differ from general expectation (e.g. a video will be mostly neutral), or very specific expectations (e.g. at 9 seconds in video "Failarmy" we expect happiness). Here we enumerate and explain our expectations.

For the first video, *Failarmy*, we expected 5 different moments where the research participant watching the video would laugh. The reason for this being that this video consists of short humorous clips. These 5 moments are points in the video with climaxes to the clips. Laughter translates to happiness in our prototype, therefore, there are 5 moments where we expect to see happiness in the charts. As mentioned above, these are visible in figure 4 as the dashed vertical lines.

For the second video, *Neck*, we expected the majority of the video to be a mix between neutral and disgust. We expect these emotions in particular because we expect people to not be deeply touched by the beauty traditions of other cultures. Disgust, in particular, we expect due to not understanding or acknowledging the culture depicted in the video as valid. Throughout the video, there are two moments in particular where we expect a spike of disgust. The first is between 23 and 25 seconds. At this moment in the video, a child is first shown wearing the neck coils. This, combined with the narrator mentioning that adding and maintaining these coils is "a painful process", cause us to believe that there will be a spike in disgust. Similarly, we expect another spike between 46 and 49 seconds, due to very comparable reasons; the video shows more coils being added to a child, and the narrator mentions that "pain is a requirement". These two areas are marked on figure 5 as blue colored areas.

For the last video, *Roast*, we expected 6 different points of laughter – happiness – throughout the video. As this was another comedy video, the expected points were moments in which a joke had just been told. Much like *Failarmy*, these points can be seen on figure 6 as vertical lines.

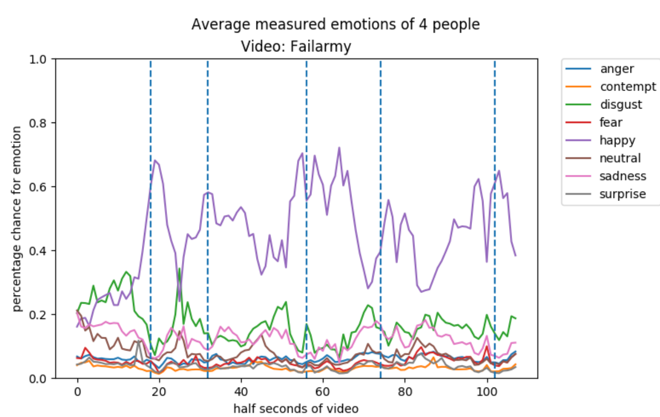

Figure 4: Average emotions of participants with video "Failarmy"

[7]https://matplotlib.org

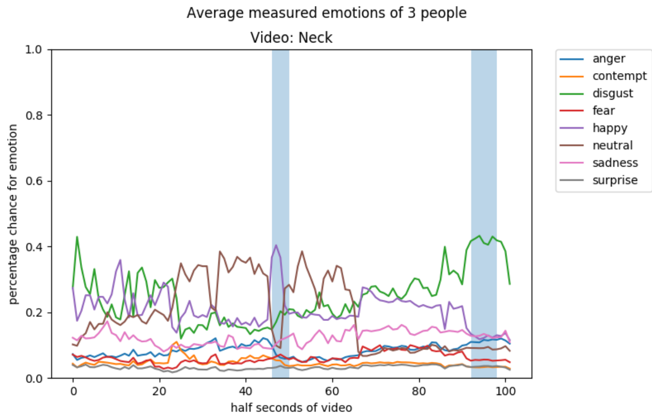

Figure 5: Average emotions of participants with video "Neck"

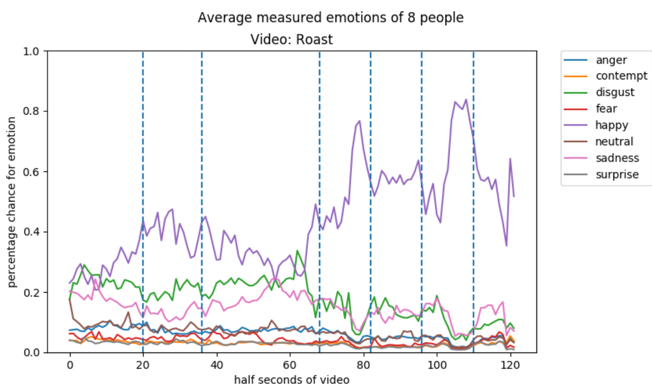

Figure 6: Average emotions of participants with video "Roast"

Based on the above charts regarding the measured emotions versus the expected emotions, in most cases the expectations seems to have come true. In figure 5, however, the expectation of disgust between 23 and 25 seconds seems to have been wrong, as happiness seems to be the most dominant emotion. Towards the end of *Neck*, there is also an increase in chance for happiness, alongside disgust.

Another interesting moment can be seen in figure 4 around 12.5 seconds into the video (x=25). Disgust takes over for a short moment in between two moments where happiness had been the emotion with the highest percent chance.

7. Discussion

In section 6 we defined two questions that, when answered, can give a more conclusive image of the results. Here we will explore our interpretation of the data and answer these questions.

Our first question was: *Do the participants agree with the emotions predicted by the prototype?*. Out of the 13 research participants, X agreed that the emotions depicted by the prototype, were correct with how they felt. Therefore, we can conclude that yes, the participants mostly agreed with the results from the prototype. One emotion often stood out as possibly wrong, namely *disgust*. Many people also agreed, however, that this might be because of they way they look normally. This could mean that either people often look with disgust by default, or that neutral and disgust were not properly distinguished from each other by the prototype. This could be solved by either letting the participants review the video of themselves to see if they agree with the judgment of the prototype, or by improving the dataset used to train the SVM model, thereby increasing the accuracy.

Our second question, *Do the predicted emotions coincide with the expectations we had for certain videos?*, can largely be answered by looking at the charts. Here we can see that in most cases, the expectations and predicted emotions by the prototype match. Only in 1 instance was the expected emotion wrong from the measured emotions. This was the expectation of disgust around 23 to 25 seconds in the *Neck* video. One possible explanation for this discrepancy could be the small amount of people that have watched the video (3 people). It is possible that a single participant skewed the results because they attribute for a third of the average. In all cases, however, the relatively small number of participants could mean that there is not enough data to give a conclusive answer.

If we combine the positive answers for both of the above questions, we get a fairly clear image of the accuracy and usability of the prototype. Although further improvements are necessary, for example a larger dataset, the prototype works sufficiently well to be used in starting a discussion. Out of 13 participants, X agreed that the visualization, especially in group settings, is capable of helping in starting a discussion regarding their emotions on a specific subject.

8. Conclusion

Research question What have we done / defined User research Conclusion on the prototype Some tips for subsequent research

At the beginning of this research we paper, we defined a research question to be answered. This question was: *Does the visualization of emotions of young people (aged 16-26) by facial expression recognition software, lead to discussion regarding the displayed content?*. In order to answer this question, we first defined how we were going to capture emotions using technology. Once we reduced the scope to facial expression recognition, we could continue with deciding how to visualize these captured emotions. We defined the emotions to *sadness, anger, contempt, disgust, fear, happiness, neutral, and surprise*, and chose a doughnut chart as model for the visualization, based on prior research by the Waag Society. We then created a prototype using open source tools and libraries, capable of predicting emotions from video and visualizing them according to the aforementioned model. Finally, we used the created prototype to conduct our user research.

We tested the prototype with 13 different subjects while interviewing them. The subsequent data analysis concluded that the prototype works largely as expected. Our expectations and the results from the prototype were mostly correct, and X out of the 13 subjects agreed with the emotions recognized by the prototype. The user research and data analysis also showed us the imperfections of the prototype. This includes a number of unexpected emotions measured by the prototype. This could be explained by too small of a dataset used to train the machine learning model responsible for predicting the emotions. This is, however, something that would have to be researched in more detail in further research. We recommend further studies to focus on improving the visualization according to the feedback given to us by our research subjects. Furthermore, increasing the dataset for the machine learning, and performing the user research on a larger group of subjects. Additionally, the stability and speed of the prototype could also be significantly improved, perhaps by switching over from Python to C++, and processing the video while it is still being recorded.

9. Acknowledgment

This research was supported by the Amsterdam University of Applied Sciences, the Waag Society research institute, and Research Group Crossmedia. We thank Bernadette Schrandt for all the assistance with the research, including guidance and feedback. We thank Lodewijk Loos and Douwe-Sjoerd Boschman from Waag Society for the assistance during the creation of the prototype. We thank Wouter Meys for setting up the minor and advice with the research. Lastly, we would like to thank all the participants who helped us test our final prototype.

References

[1] MJ Den Uyl and H Van Kuilenburg. 2005. The FaceReader: Online facial expression recognition. In *Proceedings of measuring behavior*, Vol. 30. 589–590.

[2] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. 2000. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 46–53.

[3] Mark R Leary, Kate J Diebels, Katrina P Jongman-Sereno, and Xuan Duong Fernandez. 2015. Why Seemingly Trivial Events Sometimes Evoke Strong Emotional Reactions: The Role of Social Exchange Rule Violations. *The Journal of social psychology* 155, 6 (2015), 559–575.

[4] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 94–101.

[5] Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion* 1, 3-31 (1980), 4.

[6] Jasmijn Rana, M. Willemsen, and H. C. Dibbits. 2017. Moved by the tears of others: emotion networking in the heritage sphere. *International Journal of Heritage Studies* 23, 10 (2017), 977–988. https://doi.org/10.1080/13527258.2017.1362581 arXiv:https://doi.org/10.1080/13527258.2017.1362581

[7] P. van Gent. 2016. Emotion Recognition Using Facial Landmarks, Python, Dlib and OpenCV. (2016).