Body Emotion Study Data Collection and Visualization Tool

Yuya Jeremy Ong
DS 330: Data Visualization Final Project

Introduction

The state of the art in understanding human emotions and body language has become of interest with the various applications and use cases of building better computer and human interactive systems. This case builds upon the improved algorithms and feature extraction methods that have accelerated the process of better increased the overall accuracy of building such models. However, one the major issues in building such models and understanding the analytical process involved in building the model is understanding and interpreting the data source we have used.

In our work, we have focused on primarily body language and human pose as key features and looked to analyze their relationships to the various psychological states against their physiological behaviors of the human affective state. In particular, we wanted to explore specific cases of human interactions and analyze the interpretation of a third person's perspective on the perceived emotion based on the various features we present to the subject and model a very similar interpretation of how one would perceive the emotion of the person. For this, we have built both a data collection pipeline as well as a very simple interface which allows us to explore in real-time the subject's annotation and ground truth information to uncover and address specific observations on notable areas such as subjective interpretations and variability across annotators. On one hand, this tool can be utilized as a utility for evaluating the ground truth annotations marked across various annotators to evaluate the agreeableness across each participant of the experiment. Conversely, we can also utilize this as another tool during our modeling process to evaluate how our model generates predictions on these annotated dataset and compare the output of the model with that of the human annotators. We believe that such use cases can help to accelerate research processes involved in understanding and debugging individual patterns closely observed by the model which can help us to better understand and interpret the outputs brought out by the model.

Dataset and Data Preprocessing

Dataset

To compile our dataset, we utilized a set of acted films from the AVA (Atomic Video Actions) Dataset [1], which provides a set of approximately 57.6k videos of unconstrained, but emotionally dynamic set of acted segments of actions and human interactions. Much of the content within these films provide a rich and variable set of clips which provides a collection of unique set of emotions from various genders and cultures - which allows for psychologists to further study and analyze these dynamics from alternative angles as well.

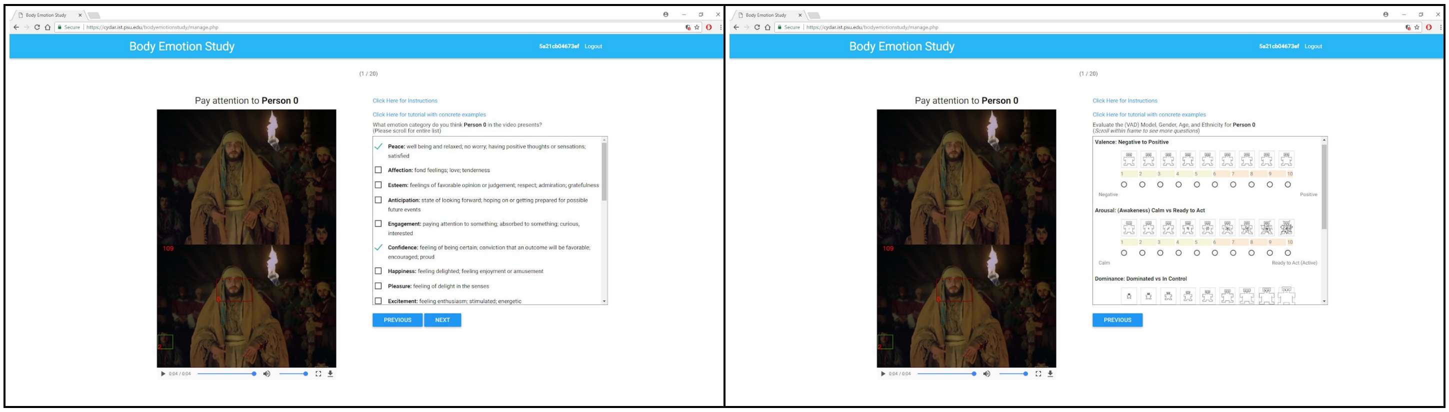
For our video processing pipeline, we first crawled the entire set of films through the provided YouTube video identifiers and downloaded them locally based on the provided metadata from the dataset. Prior to downloading the entire set of clips, we then ran the entire film through a kernel temporal segmentation (KTS) algorithm [2] to crop the clips based on sudden or faded scene changes or intense camera panning which shifts the captured scene dramatically. Prior to segmenting the clips, we then perform a first time based filter to cut down on the number of shorted scenes by reducing the number of clips by their lengths - we limit clips to have an approximate length between 5 to 15 seconds to limit the scope of the type of actions we are observing. After the segmentation and initial filtration process, we then extract the human pose information from the frame using a pose-estimator algorithm from CMU's pose estimation algorithm over each of the frame data [3]. We can further filter out clips that contain a certain consecutive number of frames to filter down the number of clips further only to those that contain human beings within the frame. The human pose estimator will extract key joints and edges which connects the joints per frame. This will be a significant aspect of the processing as much of the annotation information we correlate from this data is collected based on these features. Furthermore, to link these features across each of the frames together, we will be tracking a frame by frame interpolation between the subject of interest using a person tracking algorithm through the use of a Kalman filter based upper body tracking algorithm in conjunction with the Hungarian Algorithm for the temporal assignment of the tracked upper body bounding box features [4]. As a result of this process, we have processed up to 20,283 video samples in our database which is currently up for annotation.

The video clips that have now been processed now undergo an annotation process which we have conducted through a rigorous process to ensure that we have a well representative ground truth data for each annotation. To perform our annotation process, we collected our annotators through Amazon's Mechanical Turk website by submitting a HIT for asking each annotators to annotate 20 entities or video clips for every job. To ensure the quality of the data is high, we make sure that the quality of the annotator is very high by having them go through a preliminary process for validating whether the participants are able to analyze emotional content of the video by checking to see their EQ. We do this through a rigorous multiple choice asking them various situational cases of emotional tests to evaluate whether they are able to effectively judge the emotion of a person. The first level of this assessment has helped to remove potential spammers or ineffective annotations from our dataset.

Within each of the annotations, we ask users to first check to see whether our automated pipeline was able to successfully extract and track the human within the clip by verifying whether the clip is "corrupted". We defined the videos as being corrupted whether any of the scene segmentation or feature extraction process failed at any point. This allows us to evaluate whether the clips we have processed are of high quality and allows us to reduce and filter out any invalid data from our dataset. After checking to validate the quality of the clips are validated, we then ask the annotators to mark of a multiple checkbox interface comprised of various discrete emotional states of various states. We presented users with a set of 27 emotions along with a specific description that they can select from and allow us to get some approximated state of the affective state of the user. Then the annotators would then annotate the emotional content based on a VAD (valence, arousal, and dominance) based likert scale which scales from 1 to 10. Collecting this information allows us to analyze the emotional content of the subjective from a much more dimensional aspect. Apart from the VAD based model, we also collected information on the age, gender, and nationality of the subject of interest in the film. This information allows us to also provide much more context around based on the subject perceptions and can help to better model the various phenomena in order to best asses the type of emotions the subject is undergoing within the clip. Both annotation interfaces are shown below in Figure 1.

Finally in the last section of the annotation process, we ask annotators to best select the temporal region of the given clip which best represents the emotional state of the person. In this interface, we present them with two text boxes to enter the starting and ending frame in which the action of the given emotional state is best represented. To make the annotation process of the clip easier, we present them with corresponding frame number annotations for each of the videos along with a bounding box indicating the subject ID. Furthermore, to make this process even accessible we provide the option to the user to slow down the clip based on various speeds to better assess the emotional frames.

In addition to all the information we collect about the annotations, we also collect additional metadata such as the time it took for each of the annotation process, in seconds. Furthermore, we collect additional background information of the subject who is annotating the dataset on things such as age, gender, nationality, and country of residence to also help compare cross-cultural effects on the perception of emotions.


Body Emotion Study
5a21cb04673ef Logout
(1 / 20)
Pay attention to Person 0
109
Click Here for Instructions
Click Here for tutorial with concrete examples
PREVIOUS
NEXT
Body Emotion Study
5a21cb04673ef Logout
(1 / 20)
Pay attention to Person 0
109
Click Here for Instructions
Click Here for tutorial with concrete examples
PREVIOUS

What emotion category do you think **Person 0** in the video presents?
(Please scroll for entire list)

✓ **Peace:** well being and relaxed; no worry; having positive thoughts or sensations; satisfied
□ **Affection:** fond feelings; love; tenderness
□ **Esteem:** feelings of favorable opinion or judgement; respect; admiration; gratefulness
□ **Anticipation:** state of looking forward; hoping on or getting prepared for possible future events
□ **Engagement:** paying attention to something; absorbed to something; curious, interested
✓ **Confidence:** feeling of being certain; conviction that an outcome will be favorable; encouraged; proud
□ **Happiness:** feeling delighted; feeling enjoyment or amusement
□ **Pleasure:** feeling of delight in the senses
□ **Excitement:** feeling enthusiasm; stimulated; energetic

Evaluate the (VAD) Model, Gender, Age, and Ethnicity for **Person 0**
(*Scroll within frame to see more questions*)

Valence: Negative to Positive
1 2 3 4 5 6 7 8 9 10
○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Negative Positive

Arousal: (Awakeness) Calm vs Ready to Act
1 2 3 4 5 6 7 8 9 10
○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Calm Ready to Act (Active)

Dominance: Dominated vs In Control

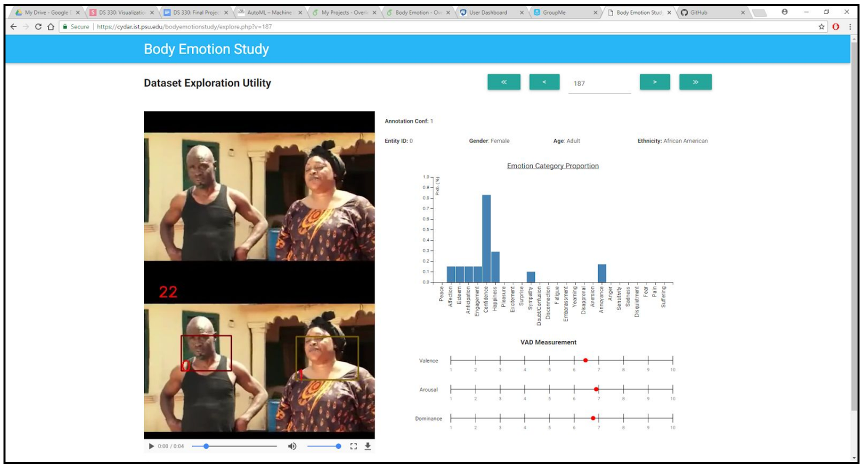
Figure 1: (Right) Discrete emotion selection interface; (Left) VAD emotion selection interface.

Dataset Preprocessing

Prior to plotting the data onto an interface, our objective was to analyze and interpret the data of the ground truth across various annotators. For this task we decided that it would be best to allocate the ground truth data processing as a pre-computed value to reduce the load time on the web based server we would be using to render our results. In this process, we extracted the information which has been persisted from our SQL database and converted to a csv file where we can easily perform processing on the dataset. We then narrowed down the clips which contains at least 5 or more ground truth annotations and conglomerated their results to evaluate several different components of the data. First, we computed their kappa coefficient to evaluate and check to observe the agreeableness across the ground truth annotation to ensure that the annotations are in fact consistent across and agreed upon each of the annotators of the dataset. Prior to that, we also computed the proportion of each of the 27 different emotions annotation value across each of the annotators. In the end of this process, we obtained a 27 dimension vector which comprises of a probability of the annotation of the video containing a specific discrete emotional state. We also computed the average value for each of the dimension of the VAD composition and extracted the mode value for the age, nationality and gender for each of the annotations present in the clip. The entire preprocessing pipeline was built through a python script which took the csv dump from our MySQL database and converted the file to an output file which contained the extracted ground truth statistics from the dataset.

Visualization Design

The key component in our dataset analytics tool aims to help users to check the individual annotations by being able to cross validate and check the annotations from the dataset and evaluate the various components of the dataset in conjunction with the original dataset. The interface we built allowed users to evaluate the clip while simultaneously allows them to view the ground truth dataset generated by the annotation in real-time. The interface we designed was based on a very simple interface where we had the original video with another video that includes the feature extracted pane that includes the frame number and the entity annotation. Besides that, we also provided an interface to visualize the confidence, dispersion or agreeableness of the annotator thoughts on the emotional state of the subject. In this interface it made it very easy for the user to be able to clearly understand how certain actions may be easier to classify by humans, while for some of the footages, some subjects maybe hard to annotate and thus presents of spread of different emotional states. Furthermore, we implement the visualization for the VAD based model by indicating the average location for which each ground truth is located. The user is able to navigate through every single one of the dataset by either entering a specific known video id or by clicking on the arrow button interfaces. The final design of the interface can be seen below in Figure 2.



Annotation Conf: 1

Entity ID: 0 Gender: Female Age: Adult Ethnicity: African American

Figure 2: Dataset Exploratory Tool

Implementation

The entire site, from the data collection process to the annotation visualization process, was all built with a combination of various technologies including PHP, MySQL, HTML, JavaScript, and D3.js. The backend process of handling the various backend form logic and database queries were all process through PHP interfaced through MySQL - this included both read and write operations to and from the database. On the other hand, we utilized the Materialize framework to build the interface structure of the site for the survey elements.

For the implementation of the data evaluation tool, we utilized D3.js to build the bar graph and the VAD analysis tools, which all constructed through simple lines and dynamic point location assignments based on the values of the corresponding features. For the dataset exploration interface, we pulled the ground truth data from a precomputed csv value to reduce the load on the MySQL server. The MySQL server for this particular page only loads the corresponding metadata regarding the sample - including the raw video file name as well as whether the video has been marked as corrupted by any of the annotators.

Result

We have been successful in collecting our data, as we have been able to collect data from nearly 9836 annotators, producing nearly 65,000 annotations - where nearly we have approximately have close 13,000 ground truth data in total. With this data, we have been able to successfully visualize all of the data and allow users to quickly navigate the dataset. From this dataset visualization tool we have been able to successfully identify clips which have strong correlations or even certain variability in their annotations. Furthermore, we have found that this tool also helped to assess the quality of the annotations by helping us to identify strange anomalies within the dataset - such as from annotations which were compromised by malicious spammers who have sent repeated requests through injected values to our online survey. This has helped to combat against any users who maliciously attempt to gain more monetary funds from our surveys compensation.

However, one suggestion we had in the implementation of our survey system was the use in reducing the number of the scaling of our likert scale from 10 to 7. The justification for this was due to the extended value of the scaling makes it very hard for the annotator to be able to quantify the subjective value into a quantitative value - which would ultimately reduce the test-retest reliability of the annotation. This has been well studied according to a study done by Preston et. al in investigating the most optimal value for the number of points to use in a Likert scale.

Conclusion

In short we have implemented a data collection process which involved web mining, video data-preprocessing, and annotation by Amazon Mechanical Turkers. We then took this data and process for its ground truth annotation as precomputed values to have it then rendered on the front end through a very simple interface to visualize the various feature components of our dataset. This has been successfully been shown to help researchers to study the dataset and understand the non-trivial perceptive nature of the dataset as well as the variability of the dataset. Furthermore, to improve the interface, we would introduce other various statistical analysis to help bridge the semantic connections between the discrete and continuous values. To help understand and contextualize our dataset better, it would also be of interest to provide recommendations of similar footages based on similar annotations to be able to comparse the qualitative aspects of the actions side by side and get a better sense for the types of qualities the annotators have picked up from the video annotations.

References

1. Gu, C., Sun, C., Vijayanarasimhan, S., Pantofaru, C., Ross, D. A., Toderici, G., ... & Malik, J. (2017). AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. *arXiv preprint arXiv:1705.08421*.
2. Potapov, D., Douze, M., Harchaoui, Z., & Schmid, C. (2014, September). Category-specific video summarization. In *European conference on computer vision* (pp. 540-555). Springer, Cham.
3. Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2016). Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*.
4. Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016, September). Simple online and realtime tracking. In *Image Processing (ICIP), 2016 IEEE International Conference on*(pp. 3464-3468). IEEE.
5. Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, *104*(1), 1-15.

Appendix

Project URL:

- Data Collection Survey: https://cydar.ist.psu.edu/bodyemotionstudy/
- Data Visualization Tool: https://cydar.ist.psu.edu/bodyemotionstudy/explore.php

Project Source - Javascript (D3.js) Component
(Note: The parameters in this particular Javascript are dynamically generated by the PHP code)

```
// Video ID Search Event Handler
        function goto_id(e) {
            if (e.keyCode == 13) {
                var v_id = document.getElementById("vid_id").value;
                window.location.href = 'explore.php?v=' + v_id;
            }
        }

        // Emotion Category Graphing
        if (document.getElementById("cat_graph")) {
            // Setup Data
            var emo_words = ["Peace","Affection","Esteem","Anticipation","Engagement",
"Confidence","Happiness","Pleasure","Excitement","Surprise",
                             "Sympathy","Doubt/Confusion","Disconnection","Fatigue",
"Embarassment","Yearning","Disapproval","Aversion",
                             "Annoyance","Anger","Sensitivity","Sadness","Disquietment","Fear","Pain","Suffering"];
            var values = [0,0.15,0.15,0.15,0.15,0.83,0.29,0,0,0,0.1,0,0,0,0,0,0,0,0.17,0,0,0,0,0,0,0];

            var data = [];
            for (var i = 0; i < 26; i++) data.push({emotion: emo_words[i], value: values[i] });

            // Initialize Plot
            var margin = {top: 40, right: 20, bottom: 120, left: 40},
                width = 600 - margin.left - margin.right,
                height = 400 - margin.top - margin.bottom;

            var x = d3.scale.ordinal().rangeRoundBands([0, width], .05);
            var y = d3.scale.linear().range([height, 0]);
            var xAxis = d3.svg.axis().scale(x).orient("bottom");
            var yAxis = d3.svg.axis().scale(y).orient("left").ticks(10);
            var svg = d3.select("#cat_graph").attr("width", width + margin.left + margin.right).attr("height", height +
margin.top + margin.bottom)
                        .append("g").attr("transform", "translate(" + margin.left + "," + margin.top + ")");

            x.domain(data.map(function(d) { return d.emotion; }));
            y.domain([0, 1]);

            svg.append("g").attr("class", "x axis").attr("transform", "translate(0," + height +
")").call(xAxis).selectAll("text").style("font-size", "12px")
                .style("text-anchor", "end").attr("dx", "-.8em").attr("dy", "-.55em").attr("transform", "rotate(-90)" );
            svg.append("g").attr("class", "y axis").call(yAxis).append("text").attr("transform", "rotate(-90)").attr("y",
6).style("font-size", "10px")
                .attr("dy", ".71em").style("text-anchor", "end").text("Prob. (%)");
            svg.selectAll("bar").data(data).enter().append("rect").style("fill", "steelblue").attr("x", function(d) {
return x(d.emotion); })
                .attr("width", x.rangeBand()).attr("y", function(d) { return y(d.value); }).attr("height", function(d) {
return height - y(d.value); });
            svg.append("text").attr("x", (width / 2)).attr("y", 0 - (margin.top / 2)).attr("text-anchor",
"middle").style("font-size", "16px")
                .style("text-decoration", "underline").text("Emotion Category Proportion");
        }

        // VAD Plotting Function
        function plot_1D(dimension, value) {
            // Setup Margins and Size Parameters
            var margin = {top: 30, right: 20, bottom: 10, left: 80},
                width = 600 - margin.right,
                height = 60 - margin.top - margin.bottom;

            // Setup Graph
            var svg = d3.select("#"+dimension);
            var line = svg.append("line").attr("x1", margin.left).attr("y1", margin.top).attr("x2", width).attr("y2",
margin.top).attr("stroke-width", 1).attr("stroke", "black");
            for (var i=0; i<=10; i++) {
                svg.append("line").attr("x1", margin.left + ((width-margin.left)/9)*i).attr("y1",
margin.top-10).attr("x2", margin.left + ((width-margin.left)/9)*i).attr("y2", margin.top+10).attr("stroke-width",
1).attr("stroke", "black");
                svg.append("text").attr("x", margin.left + ((width-margin.left)/9)*i).attr("y",
margin.top+25).attr("text-anchor", "middle").style("font-size", "10px").style("fill", "gray").text(i+1);
            }

            // Plot Label
            svg.append("text").attr("x", margin.left-50).attr("y", 35).attr("text-anchor", "middle").style("font-size",
"12px").text(dimension.charAt(0).toUpperCase() + dimension.slice(1));

            // Plot Value
            svg.append("circle").attr("cx", ((width-margin.left)/10)*(value+1)).attr("cy", margin.top).attr("r",
5).style("fill", "red");
        }

        // Plot VAD Values
        plot_1D('valence', 6.67);plot_1D('arousal', 7.15);plot_1D('dominance', 7.01);
```