A Comparison of Facial Emotional Recognition Accuracy Among Machine Learning Algorithms and Humans

Elliott T. Ruebush

Christopher Kramer

October 31, 2017

Abstract

This experiment examined the question of how machine learning methods compare to humans in their capability to detect emotions based off facial expression. Three different machine learning facial recognition methods were used to perform facial emotional recognition on the extended Cohn-Kanade dataset, and humans were asked to assign emotion to a set of images from the same dataset. I hypothesised that humans would be more accurate than machine learning methods, which would vary somewhat in accuracy and be generally lower in accuracy than humans. In the end, for the specific emotions compared, the results of the experiment showed surprising similarity (approx. 10% difference) in the accuracy of the Fisherfaces method and the accuracy of human emotional evaluation.

This study aimed to answer the question of how computer based algorithmic machine learning methods compare to humans in the task of identifying emotion based on facial expression. Studying this idea holds importance because doing so can help in determining where research in the field of computer emotional recognition should focus. Additionally, on a broader scale, teaching computers to recognize emotion is an essential step in developing artificial intelligence. Furthermore, facial emotional recognition serves as a useful tool in any project that aims to utilize emotions for implementation of a component. Overall, discerning the differences between humans and computers in recognizing emotions presents itself as an enlightening exercise that provides insight into how advanced today's machine learning algorithms are. Throughout my background research, I was able to find studies comparing machine learning algorithms (such as the 2011 work by Jang et al examining optimal emotional recognition algorithms), however papers that directly compared humans with machine learning algorithms were less common, therefore I saw an opportunity for contributing some additional research and data to the field. On a more personal side note, this project appealed to me because it motivated me to develop skills with Python and some of its various libraries.

The design of this study was relatively simple. It simply included running tests on the machine learning data that output accuracies, and then comparing those accuracies to human accuracies. In order to ensure that there was more to compare than just two averages, I also tracked machine learning accuracies per emotion and calculated human accuracies per emotion. Due to this relatively simple design, as well as the general nature of the experiment, I was able to

easily classify the method used for facial recognition (humans, fisherfaces, LBPH, eigenfaces) as the independent variable, leaving the accuracy of facial emotional recognition as the dependent variable. To restate the purpose of the experiment: this study sought to compare facial emotional recognition accuracy among multiple machine learning algorithms as well as humans. I hypothesized that the machine learning methods would vary mildly in accuracy (0% - 25% difference), and generally be less accurate than humans.

Method

Participants

High school students comprised the participants surveyed in this experiment, with a majority of the students being male. Participants filled out a survey in which they were asked to choose which emotion out of the 7 given (Anger, Contempt, Disgust, Fear, Happiness, Sadness, Surprise) best described each facial expression in a set of 32 color images from the extended Cohn-Kanade dataset. Age of students surveyed was not recorded, but all participants were members of high school computer science classes. The only personal data collected by the survey was gender of the participant.

Assessments and Measures

For this study, the images used in both the survey and the computer testing were taken from the extended Cohn-Kanade dataset (Lucey et al, 2010) (). The Anaconda distribution of the programming language Python was used for all computerized portions of the experiment. Libraries used include OpenCV, NumPy, Matplotlib (Specifically PyPlot), Pandas, and Jupyter

(specifically a Jupyter Notebook). OpenCV was used for image processing and three of its built-in machine learning facial recognition methods were used for emotional recognition. NumPy was used for formatting of lists into arrays and for various mathematical calculations. PyPlot was used for data visualization. Pandas was used for reading .csv files with the survey data into a format that could be manipulated using Python. A Jupyter Notebook was used as the environment in which code was written and results were displayed. The survey conducted on human participants was created using Google Apps Script and Google Forms. 13 of the 60 surveys were conducted on June 1st, 2017 between approximately 12:05 pm and 12:20 pm. 47 of the 60 surveys were conducted on September 26th, 2017 between approximately 7:50 am and 11:55 am. All surveys were given to students during computer science classes at the Charter School of Wilmington.

Outline of Experiment

To begin, the Cohn-Kanade dataset was downloaded from its website at http://www.consortium.ri.cmu.edu/ckagree/. Next, the dataset was uploaded to Google Drive and Google Apps Script was used to iterate along constant intervals through a portion of the color images and add them to a Google Forms survey. The survey was conducted as detailed earlier. After completing that portion of experimental preparation, the computer tests began. First of all, the emotions of the dataset (read in from .txt files included in the dataset that corresponded with specific images) were matched with the file paths to images using a Python dict. Afterwards, OpenCV was used to classify the images as faces and then convert the images to all be the same dimensions. A new directory was created and the new images were moved to that directory.

Following the reformatting of the images, three different methods of machine learning based facial recognition included in the OpenCV library were used to classify the images based on emotion. The three methods used were Fisherfaces, Eigenfaces, and Local Binary Patterns Histograms (LBPH). The execution of each method and subsequent data collection followed the same formula. First, the data was organized into a training set containing 80% of the images and a testing set containing 20% of the data. Then the method currently being used was run on the training set. After being run on the training set, the method was then run on the testing set and count of total correct guesses as well as amount of correct guesses per emotion was kept. Following the running of machine learning tests, the data was organized and plotted on different charts using PyPlot.

Data collected

The data collected varied between the computerized tests and the humans surveyed because requiring humans to analyze the same amount of images as an algorithm was realistically unfeasible due to time constraints and willingness of students to participate. Average percent correct for all guesses over 15 trials was collected for each method, as well as the average percent correct for each emotion over those same 15 trials. From the human test subjects, average percent correct for all participants was collected, as well as the average percent correct among all participants for each individual image included on the survey.

Results

Figure 1

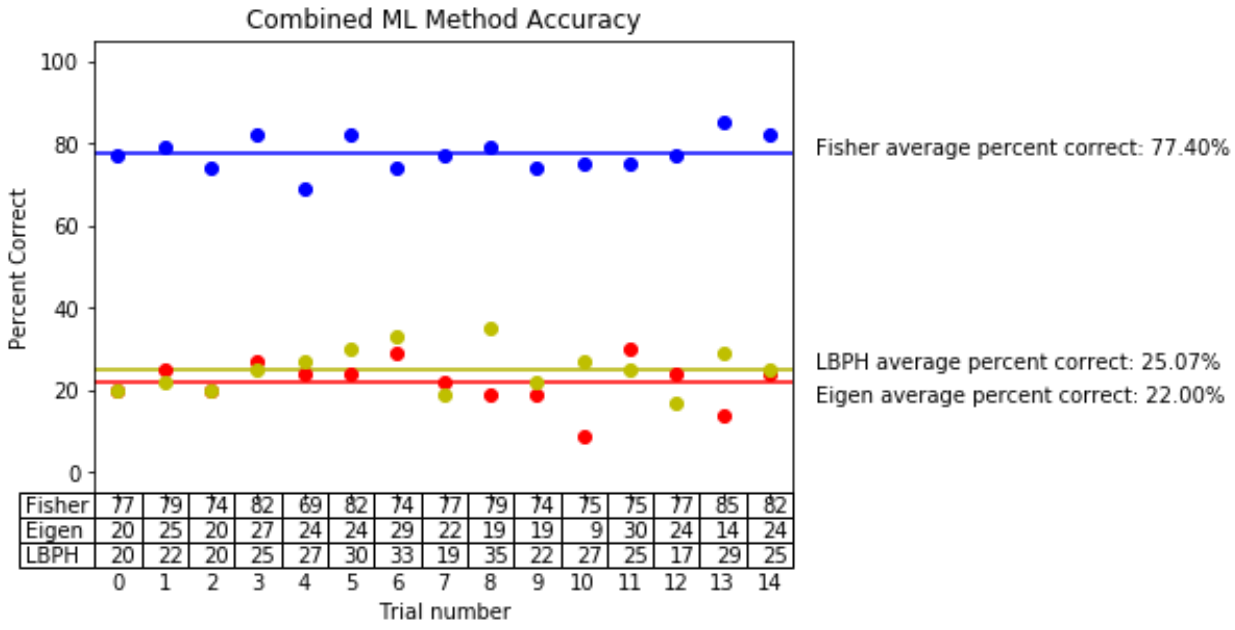


Figure 2

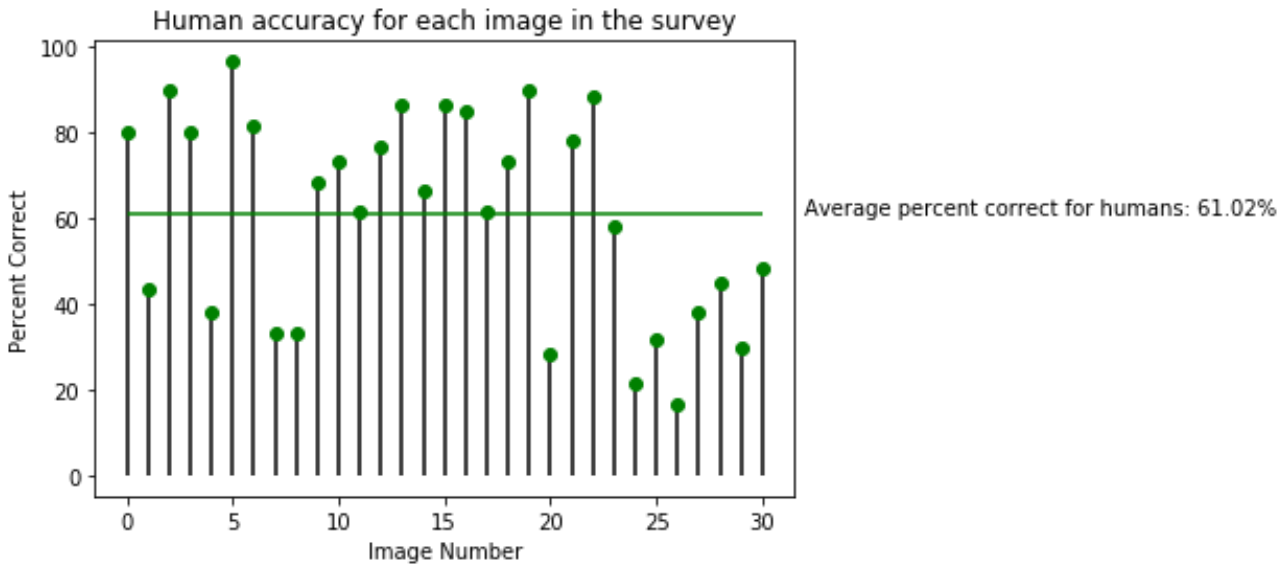


Figure 3

Figure 3

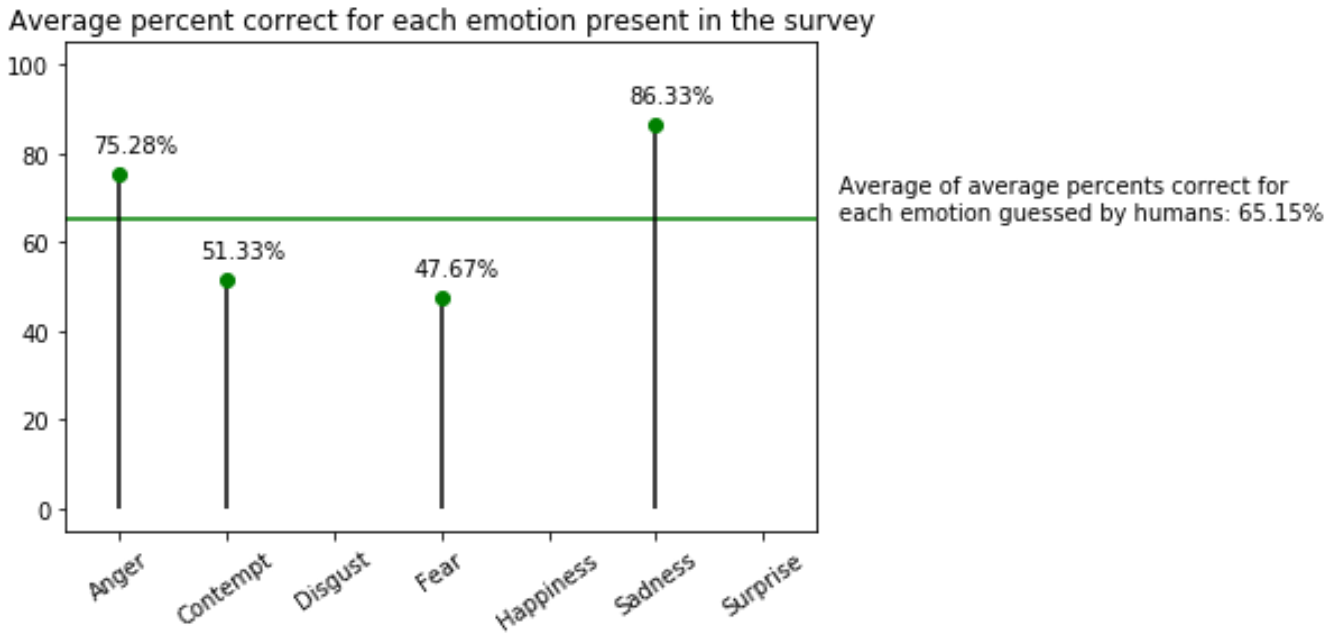
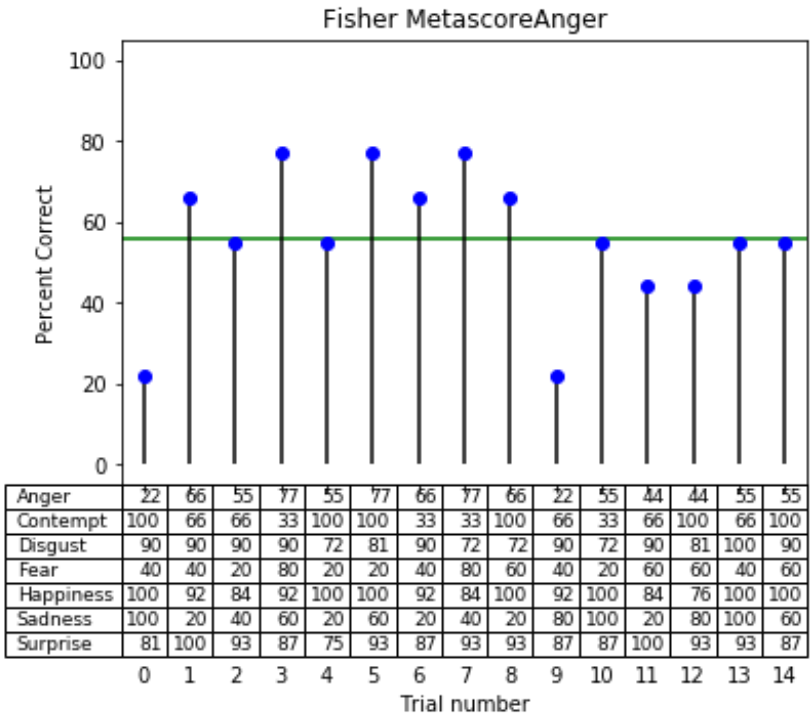
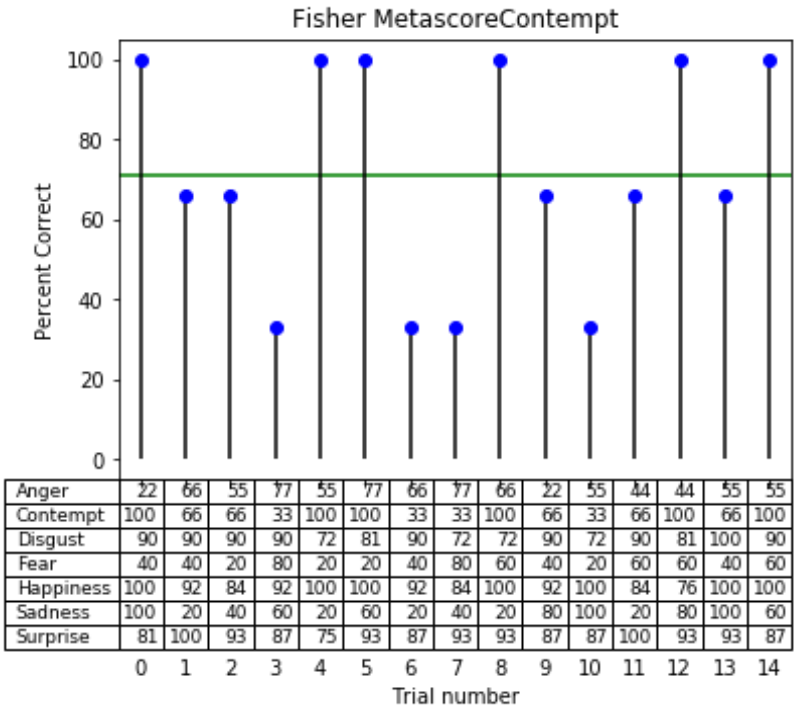

Average percent correct for each emotion present in the survey
100
80
60
40
20
0
75.28%
51.33%
47.67%
86.33%
Anger
Contempt
Disgust
Fear
Happiness
Sadness
Surprise
Average of average percents correct for each emotion guessed by humans: 65.15%

Figure 4



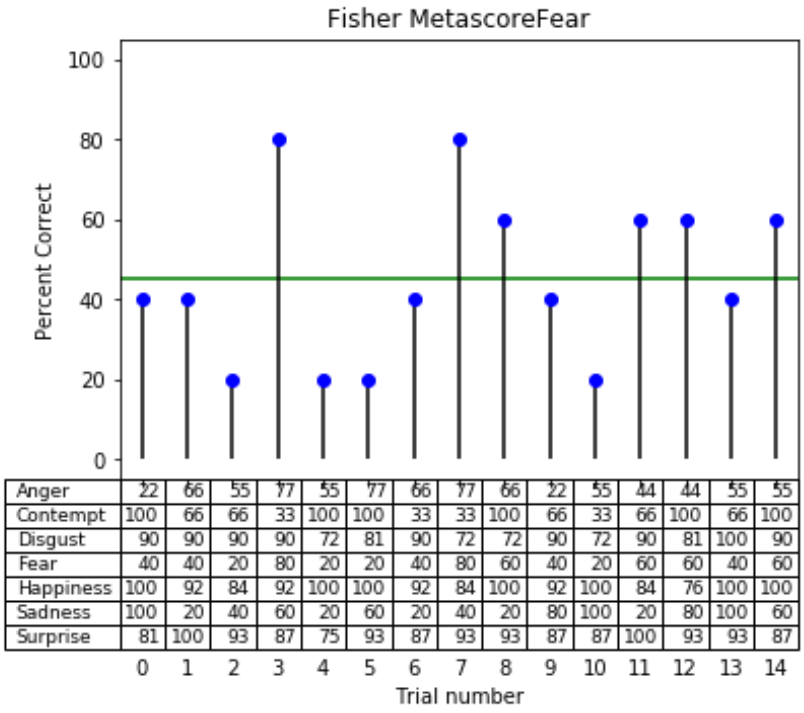
Anger	22	66	55	77	55	77	66	77	66	22	55	44	44	55	55
Contempt	100	66	66	33	100	100	33	33	100	66	33	66	100	66	100
Disgust	90	90	90	90	72	81	90	72	72	90	72	90	81	100	90
Fear	40	40	20	80	20	20	40	80	60	40	20	60	60	40	60
Happiness	100	92	84	92	100	100	92	84	100	92	100	84	76	100	100
Sadness	100	20	40	60	20	60	20	40	20	80	100	20	80	100	60
Surprise	81	100	93	87	75	93	87	93	93	87	87	100	93	93	87

Figure 5



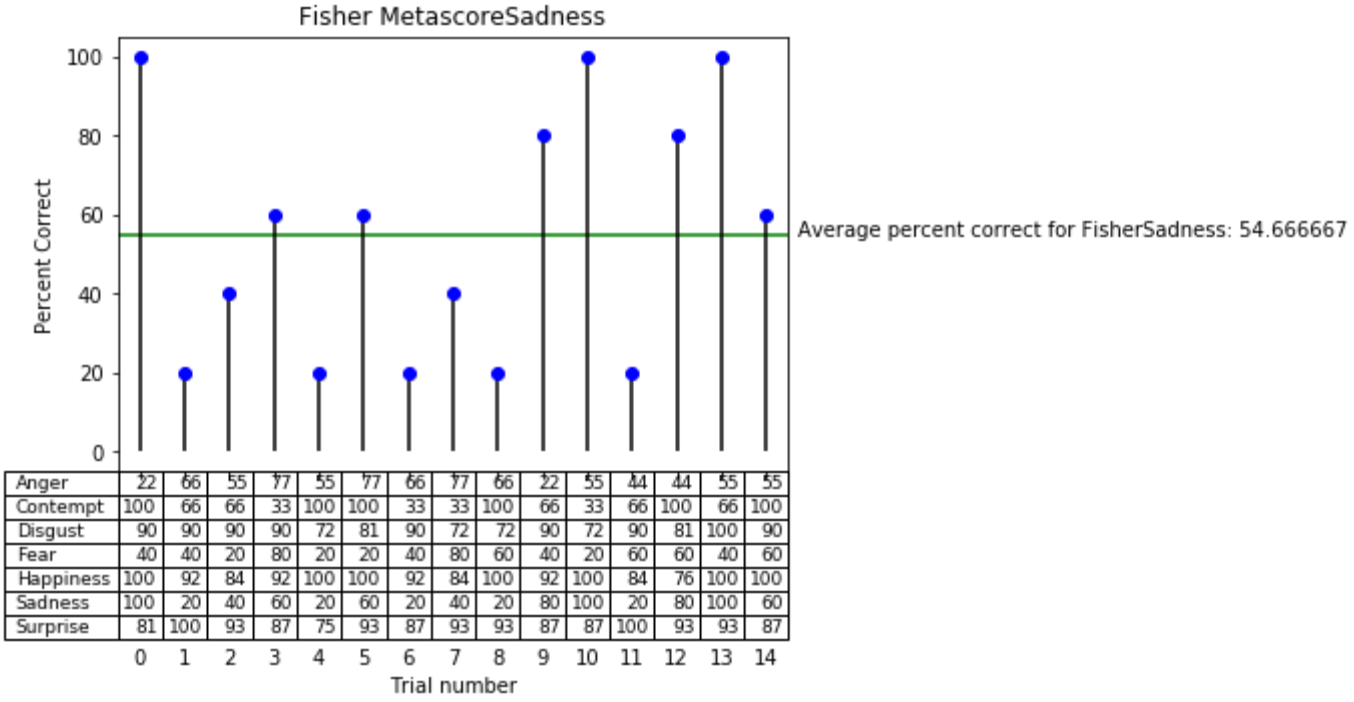
Anger	22	66	55	77	55	77	66	77	66	22	55	44	44	55	55
Contempt	100	66	66	33	100	100	33	33	100	66	33	66	100	66	100
Disgust	90	90	90	90	72	81	90	72	72	90	72	90	81	100	90
Fear	40	40	20	80	20	20	40	80	60	40	20	60	60	40	60
Happiness	100	92	84	92	100	100	92	84	100	92	100	84	76	100	100
Sadness	100	20	40	60	20	60	20	40	20	80	100	20	80	100	60
Surprise	81	100	93	87	75	93	87	93	93	87	87	100	93	93	87

Figure 6



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Anger	22	66	55	77	55	77	66	77	66	22	55	44	44	55	55
Contempt	100	66	66	33	100	100	33	33	100	66	33	66	100	66	100
Disgust	90	90	90	90	72	81	90	72	72	90	72	90	81	100	90
Fear	40	40	20	80	20	20	40	80	60	40	20	60	60	40	60
Happiness	100	92	84	92	100	100	92	84	100	92	100	84	76	100	100
Sadness	100	20	40	60	20	60	20	40	20	80	100	20	80	100	60
Surprise	81	100	93	87	75	93	87	93	93	87	87	100	93	93	87

Figure 7



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Anger	22	66	55	77	55	77	66	77	66	22	55	44	44	55	55
Contempt	100	66	66	33	100	100	33	33	100	66	33	66	100	66	100
Disgust	90	90	90	90	72	81	90	72	72	90	72	90	81	100	90
Fear	40	40	20	80	20	20	40	80	60	40	20	60	60	40	60
Happiness	100	92	84	92	100	100	92	84	100	92	100	84	76	100	100
Sadness	100	20	40	60	20	60	20	40	20	80	100	20	80	100	60
Surprise	81	100	93	87	75	93	87	93	93	87	87	100	93	93	87

Table 1

Comparing Averages for the Emotions Present on Human Survey

Emotion	Anger	Contempt	Fear	Sadness	Totals:
Humans Avg % Correct	75.28%	51.33%	47.67%	86.33	Avg: 65.15%
Fisherface Avg % Correct	55.73%	70.80%	45.33%	54.66%	Avg: 56.63%

Discussion

During the course of experimentation, I discovered that Fisherfaces was by far the most effective of the machine learning methods examined, outperforming the other methods by over

50% (*See Figure 1: "Combined ML Method Accuracy"*). Taken at face value, it appears that Fisherfaces was more effective than humans at recognizing emotion based on facial expression. However, although the overall performance of Fisherfaces (77.40%) (*Figure 1*) outstripped that of the average per image accuracy of the human test group (61.20%) (*Figure 2*), when averages of average accuracies for only certain emotions are compared, the human group (65.15%) actually outperforms Fisherfaces (56.63%) (*See Table 1*). The human test group was unfortunately not tested on Disgust, Happiness, or Surprise due to survey complications (The issue was not recognized until after the surveys had already been given, and there was not time to collect new, more comprehensive data). Therefore, although Fisherfaces' total average accuracy is higher than the average accuracy across all images from the human survey, no definitive claim can be made relating human performance across all seven emotions (Anger, Contempt, Disgust, Fear, Happiness, Sadness, Surprise) to machine learning performance across those same 7 emotions. Claims can only be made regarding performance among Anger, Contempt, Fear, and Sadness. This issue is perhaps the most glaring and major error present in the experiment. Due to repeated testing (15 training-testing cycles across the entire dataset for each method) with accurate results, the probability of irregularities is extremely low on the end of the machine learning methods. The human portion of the experiment holds greater room for error, especially in the accidental omission of multiple emotions that occurred during the running of the script used to create the survey (Script can be accessed here). However, the survey included sixty participants from the same population of high school computer science students, so conclusions

on emotional recognition accuracy of high school computer science students for the emotions present on the survey can still be made.

As far as relation to previous research, the machine learning methods I used failed to seriously compete with the most accurate methods that can be employed (SVMs, such as the one used by the creators of the Cohn-Kanade dataset have reached accuracies of 95% or higher)(Lucey et al, 2010). A great variety of machine learning algorithms with different applications and methods of implementation exist (Brownlee, 2013). Therefore, this discrepancy in my accuracies and accuracies of other experiments dealing with machine learning and facial emotional recognition can be attributed to the features used in my experiment versus in other experiments. In my experiment, I used OpenCV algorithms designed for facial recognition and instead adapted them to attempt facial emotional recognition as demonstrated by Paul van Gent on his personal website (2016). This method of facial emotional recognition differed from some other methods as various facial landmarks are often used as features when doing facial emotional recognition. Cohn et al demonstrated the use of facial action coding sequences (FACS) as features for machine learning-based emotional recognition within their presentation of the original Cohn-Kanade dataset (2000). However, I had neither the expertise nor time to learn to implement these other methods and other features. Development of more specific and customized facial emotional expression related features (a 2017 paper published by Chu et al detailed such a process) likely would have yielded high accuracies.

If I were to redo this experiment, I would ensure the inclusion of all emotions in the survey, perhaps giving multiple surveys of randomized images from the Cohn-Kanade dataset.

Additionally, I would attempt to have a larger amount of survey participants from a larger variety of backgrounds. Furthermore, I would add additional machine learning methods and consider using facial landmarks as features to provide additional (and likely more accurate) data on the current state of advanced machine learning based facial emotional recognition as it relates to human facial emotional recognition. To ensure an even greater variety of data and to more closely emulate real world-esque situations I would also examine less structured data sets and random facial images pulled off the internet, comparing human accuracy versus machine learning accuracy for those images. In a broad sense, I would attempt to increase the overall scale of my experiment, making testing more rigorous and collecting a larger amount of more varied data.

Conclusion

The purpose of this experiment was to compare how accurately different machine learning algorithms can determine emotion based on facial expression along with how accurately humans are able to determine emotion based on facial expression. I found that humans and the more effective methods of machine learning perform similarly. The data I collected did not support my hypothesis that machine learning methods will vary mildly (0%-25% difference) and be less accurate on average than humans, but neither did it fully disprove it as the data collected for humans did not allow for a complete comparison for all emotions. To conclude, my results serve as a source of additional data regarding facial emotional recognition by machine learning methods versus by humans, however they also necessitate further, more exhaustive testing in order to glean a better idea of how machine learning algorithms truly compare to humans.

References

Brownlee, J. (2013, November 25). A tour of machine learning algorithms. Retrieved February 26, 2017, from http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Chu, W. S., De la Torre, F., & Cohn, J. F. (2017). Selective transfer machine for personalized facial expression analysis. IEEE transactions on pattern analysis and machine intelligence, 39(3), 529-545.

Jang, E. H., Park, B. J., Kim, S. H., Eum, Y., & Sohn, J. H. (2011, November). Identification of the optimal emotion recognition algorithm using physiological signals. In *Engineering and Industries (ICEI), 2011 International Conference on* (pp. 1-6). IEEE.

Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), Grenoble, France, 46-53.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010), San Francisco, USA, 94-101

van Gent, P. (2016). Emotion Recognition With Python, OpenCV and a Face Dataset. *A tech blog about fun things with Python and embedded electronics.* Retrieved from: http://www.paulvangent.com/2016/04/01/emotion-recognition-with-python-opencv-and-a-face-dataset/

Acknowledgements

I would like to acknowledge my father, Mitch Ruebush, for help in coming up with the idea for my research, as well as pointing me in the right direction for my preliminary background research. Additionally, I'd like to give a major thanks to my mentor, Mr. Kramer for general advice throughout the project and for generously helping with my surveys of humans. I would be severely lacking in survey data without Mr. Kramer's allowing me to give my survey during one class, as well as his agreeing to distribute my survey among multiple other classes. Furthermore, I would like to thank all survey participants for their help in facilitating my research. Finally, I would like to sincerely thank Cohn et al for their creation of the initial Cohn-Kanade dataset (2000) and Lucey et al. for their updating of the original Cohn-Kanade dataset to form the extended Cohn-Kanade dataset (2010), without which I would not have been able to perform my experiment.

Appendix

Anything that would potentially be included in this appendix can be found in the Jupyter Notebook for this project, which is hosted on my github account (email address for the account: ruebush.elliott@charterschool.org).