Color Histogram Matching Approach for Speech Emotion Recognition

Ryan Jaeger[1] and Dylan Shoemaker[1]

Abstract— **In human speech, information is transmitted through the actual words spoken, but also through the emotion behind the utterance. The problem of speech emotion recognition looks to add this emotional context to machines understanding of speech. Speech emotion recognition, or SER, can be framed as a classification problem where a given audio clip can be labeled as one of the seven basic emotions. In this study, we used label audio clips from the Berlin Emotional Dataset to create spectrograms, visual representations of these images. We introduce a novel approach to emotion classification through an evaluation of color histograms generated from these spectrograms. Our Color Histogram Matching (CHM) method performs comparably well to both a convolutional neural network based on spectrograms, and an ensemble classifier based on traditional audio classification features, namely energy, zero crossing rate, and MFCC. With an accuracy rate of approximately 37%, there is evidence the CHM methods holds some predictive power for classifying emotions from speech. Although there is significant room for improvement, our findings indicate that color histograms offer an effective and robust means of conducting speech emotion recognition.**

I. INTRODUCTION

Speech signals are incredibly important to human interaction. Vocal utterances are the fastest and most natural method of communication between humans [1]. As more and more humans interact with intelligent machines on a daily basis, speech can be an important method of communications. Ultimately, the goal is to enable a very natural interaction with the computer by speaking instead of using traditional input devices and not only having the machine understand the verbal content, but also more subtle cues such as affect that any human listener would easily react to [2]. Since the 1950s, progress has been made on speech recognition, which converts human speech into a sequence of words processed by a machine. However, this is not enough to have a natural interaction between humans and computers.

[1]The Pennsylvania State University, University Park, PA. DS 340 Term Project. Dr. James Z. Wang and Dr. Jia Li

A. What is Speech Emotion Recognition?

The next logical step to improving human interactions with machines is speech emotion recognition, or SER. Simply put, speech emotion recognition is the ability for a machine to understand the emotion conveyed by a clip of human speech. More formally, speech emotion recognition is the process of predicting the high-level affective content of an utterance from the low-level signal cues produced by a speaker [3]. The ultimate goal of speech emotion recognition is to add semantic meaning to our conversations with machines. It is easy to see how speech emotion recognition could be framed as a classification problem, where we assign an emotional label to a given audio input. However, this task is more computationally difficult than it might initially seem.

B. Challenges of Speech Emotion Recogntion

The challenges of speech emotion recognition are numerous but not insurmountable. To begin, if we decide to frame speech emotion recognition as a classification problem, we need to define a finite number of emotions to act as possible states. In this study, we use the seven primary emotions as defined by our dataset - the Berlin Emotion dataset. The seven emotions used in labeling this data are: anger, boredom, disgust, fear, happiness, sadness, and neutral. While this technical challenge was easily addressed, the other challenges are more difficult. From a given audio clip, it is not always obvious which characteristics are most predictive in distinguishing emotion [1]. Acoustic variability caused by difference in sentences, speakers, speaking styles, and rates of speech impact energy and pitch, which are two most commonly used features in speech. Therefore, it becomes difficult to separate emotion from simple variability. Next, there may be more than one perceived emotion in a given utterance [1]. Emotions do not exist

in a vacuum. It is possible and sometimes likely for a speaker to be expressing more than one emotion when they are saying something. Then, individual variations in the culture, environment, and personality of the speaker can also increase the complication of the task [1]. Finally, audio data can be high dimensional depending on the feature extraction method, so this can be a computational challenge [3].

C. Applications to Human-Computer Interaction

Because of its potential to imbue more intelligence into communications with technology, speech emotion recognition is one of the emerging fields within human-computer interaction [2]. The applications of speech emotion recognition within HCI are near limitless, encompassing topics as far-reaching as robots, plane cockpit systems, and information retrieval for medical analysis. One particularly interesting application is to customer service call centers [4]. The idea is to provide human operators with information regarding the emotions given off by their own voice to help them better improve their interactions skills. Alternatively, it could be used to alert the human operators to particularly angry customers to help triage the service. SER has also been suggested for computer tutorial systems to help tutors adjust their methodology based on the detected emotion of the student [4]. Finally, it has also been proposed to be integrated into in-car monitoring systems to respond to changes in emotion of the driver that might lead to erratic driving [4].

D. Related Work

Typically, speech emotion recognition systems work by extracting features from speech, followed by a classification procedure to predict emotions [5]. Common audio features extracted and used for analysis include pitch, energy, zero-crossing rate, linear predictor coefficients (LPC), linear predictor cepstral coefficients LPCC, mel-frequency cepstral coefficients MFCC, and Teager-energy-operator. Because of the high dimensionality of the audio feature space, it is almost always necessary to perform feature selection, and only use a subset of the possible features [3]. Classification models such as Support Vector Machines [6], [7], Hidden Markov Models [8], [9], [10], and others are very commonly used. In recent years, deep learning methods have also been applied to the problem of speech emotion recognition [5], [11], [12], [13].

E. Research Contributions

In this paper, we present a novel application of color histograms, an image analysis tool, for the purpose of audio emotion classification. Our rapid, robust model for SER performs comparably well to deep learning and traditional audio classification methods. In the next section, we will introduce the dataset used in our analysis and the preprocessing involved. We then present our methodology for the Color Histogram Matching method and two contemporary SER classifiers. We next present our results and discuss the implications of our findings, and then conclude with a discussion of future work.

II. DATA AND PREPROCESSING

A. Berlin Emotional Dataset

The collection of labeled emotional speech samples can be a difficult and time-consuming task, so for our analysis, we utilized the Berlin Emotional Dataset [14]. The Berlin Emotional Dataset is a collection of spoken German sentences from 1997 to 1999. Utilizing five male actors and five female actors ranging in age from 21 to 35, these researchers prompted the actor to say a given line while working to convey a given emotion. The emotions they used for their collection were Anger, Boredom, Disgust, Fear, Joy, Neutral, and Sadness. The audiologists selected 10 everyday sentences to get the most realistic expectation of what emotions sounds like in the real world, and recorded the actors in an anechoic chamber for maximum sound quality. The final dataset used for analysis is 535 utterances with associated emotional labels, with the length of each utterance ranging from approximately 1-3 seconds.

B. Data Augmentation

Because of the small sample size of the dataset, we looked to augment our data in some way. By altering and manipulating the audio samples, we could simultaneously achieve a larger set of data, but also a more robust set, as it contains lower quality audio than the expertly produced Berlin utterances. Specifically, for the audio clips, we employed rolling, stretching, rate changing, pitch changing, and white noise addition to each utterance. Ultimately, we were able to create 6 new audio clips from each original clip, resulting in a combined dataset size of 3,745 utterances used for analysis.

C. Spectrograms

Next, one of the primary tools for our analysis of the speech emotions is the spectrogram. A spectrogram is the visual representation of a signal strength over time at different frequencies present in certain waveform [5]. The image encodes data about time along the horizontal axis, frequency along the vertical axis, and the amplitude at each point is associated with color. Lower amplitudes are shown in dark blue colors and stronger amplitudes are shown in bright colors up to red. A spectrogram can be computer via a Fast Fourier Transform , and spectrograms have been used in other speech analysis contexts [5]. We generated our spectrograms with the MatPlotLib package of Python for each of the 3,745 audio clips in our combined dataset. See Figure 1 for a sample spectrogram.

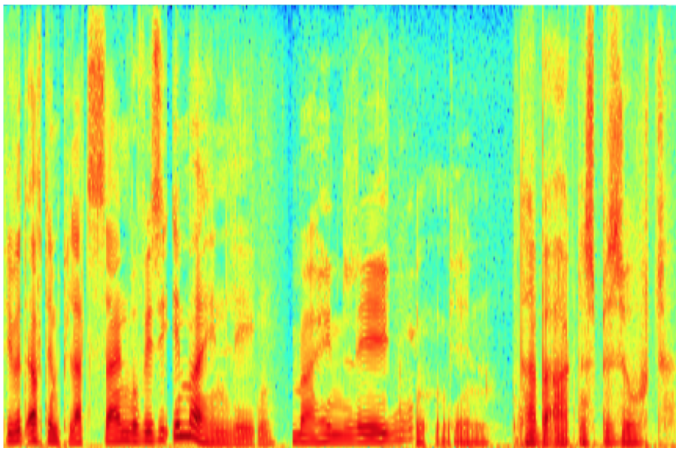

Fig. 1. Sample spectrogram generated from the Berlin Emotional Dataset

III. COLOR HISTOGRAM MATCHING

The proposed methodology consists of a novel approach to speech emotion recognition involving color histograms and nearest centroid classification based on the Wasserstein distance metric. This method is more robust than many machine learning classification models and less computationally-intensive than other deep learning models. The main components of the approach and comparisons to these concurrent methods are outlined in the following sections.

A. What is a Color Histogram?

A color histogram is a visual representation of the color composition of an image. Color histograms have been used in classical image classification tasks such as histogram intersection and histogram backprojection

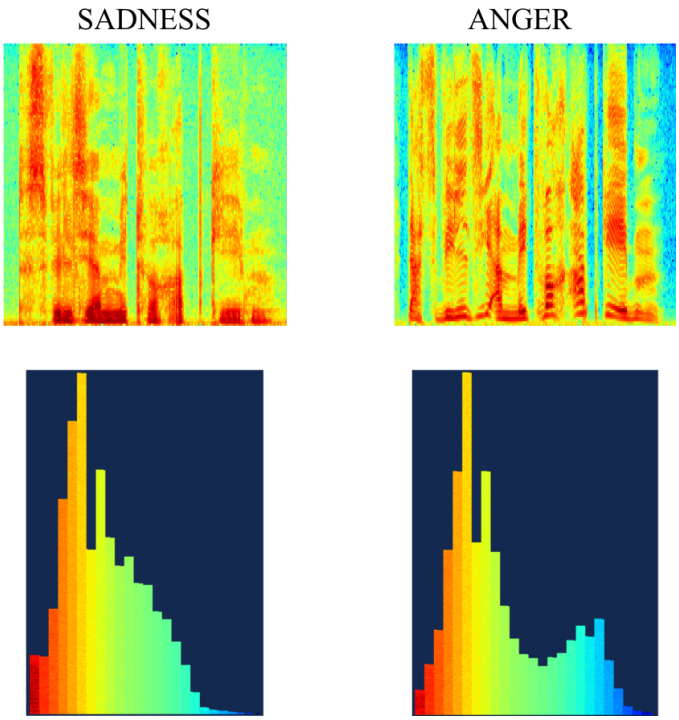

Fig. 2. Comparison of the Color Histograms for sample audio clips of Sadness and Anger

[15]. One of the most significant drawbacks of the use of color histograms in the typical image classification setting is that important information regarding the shape of the object(s) depicted in the image. This becomes less of an issue in the case of spectrograms since the images depict signals rather than shapes or objects. To create color histograms, the pixels in the chosen image are converted from their RGB representations to their corresponding HSL (Hue, Saturation, Lightness) representations. The choice of adding this transformation stems from the school of thought that HSL color representations more closely model how humans perceive color than RGB components. Each pixel can then be placed into one of the corresponding bins (number of bins is a model hyperparameter) for the color histogram based on the H component. This results in a pictorial distribution of the colors present in the chosen image.

B. Motivation

Upon inspecting the spectrograms generated in our dataset, we noticed some general patterns in color composition among the emotional classes (see fig. 2). We conducted further screening by computing the color histograms for samples from each emotional class. One general trend that we observed was that the spectro-

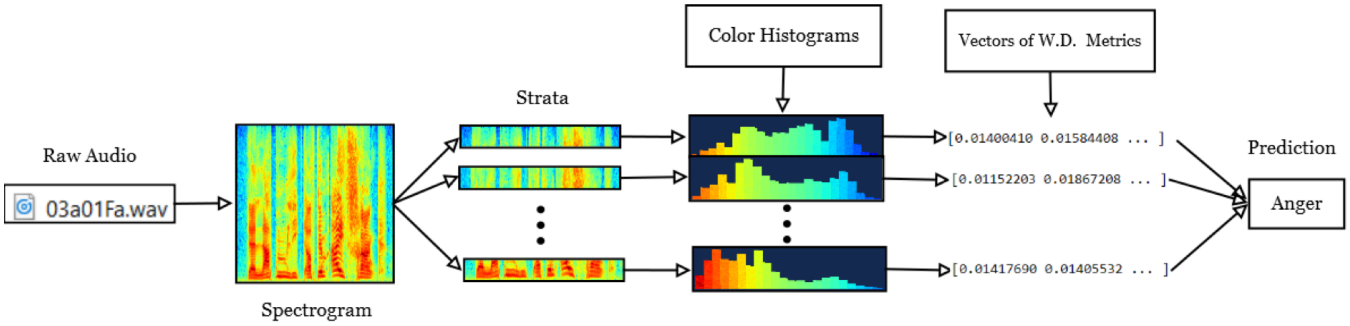

Fig. 3. Diagram of the Color Histogram Matching Data Flow

grams for anger tended to be dominated by brighter colors like yellow, red, and orange, with secondary peaks around deeper blues, while spectrograms for sadness were composed of higher proportions of intermediate colors like lighter blues and greens. Additionally, we saw that spectrograms belonging to the boredom class were dominated largely by yellow and light orange hues, with very little representation of reds and blues, while disgust was characterized by a primary peak of yellow and light orange with a secondary peak of light green and blue. Furthermore, we observed general trends across all emotional classes in terms of how the color composition of the spectrogram changed with respect to changes in frequency (i.e. changes in the y-axis of the spectrograms). Altogether, this led us to conjecture that each emotional class might possess a distinct color distribution at each of the frequency levels depicted in the spectrograms.

C. Proposed Framework

The conjecture outlined in the previous section served as the basis for our algorithmic design in our novel SER classification approach. Functionality-wise, the algorithm could be described as stratified nearest centroid classification using the Wasserstein distance metric. For the sake of simplicity, we referred to it as color histogram matching (CHM). Further elaboration of the Wasserstein distance metric, the nearest centroid classification algorithm, and our algorithmic data flow for CHM is detailed in the following sub-sections.

1) Wasserstein's Distance: The Wasserstein distance metric can be thought of as the amount one needs to alter a probability distribution to obtain a second probability distribution. Intuitively, if the original distribution is thought of as a pile of dirt, then the Wasserstein distance between the original pile and the final pile can be viewed as the minimum cost of altering the original pile to become the final pile; where it is assumed the cost of transforming a pile is the product of the amount of dirt to move and the distance to move it. For this reason, Wassersteins distance is often colloquially referred to as earth movers distance. Formally, Wassersteins distance is computed by the following equation:

$$W_1 = \inf_{\pi \in \Gamma(u,v)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| d\pi(x,y)$$

[16]

where $\Gamma(u,v)$ is the set of distributions on whose marginal distributions are u and v on the first and second factors respectively [16].

2) Nearest Centroid: The nearest centroid classification algorithm essentially classifies observations according to the label for the class whose mean or centroid the observation is nearest to. The algorithm consists of two steps: training and predicting. During training, the centroids are computed by taking the mean of all features belonging to each class present in the dataset. Prediction is then performed by assigning class observations based on the class label which minimizes the distance from the observation to the centroid of that class.

3) Dataflow Diagram: With these definitions in mind, we can now introduce the CHM algorithm in Figure 3 above.

Prior to any algorithmic implementation, we first partition the dataset into training and test sets the training set consisting of about 3,000 audio files (80% of the dataset, roughly 430 images per class) and the test set of about 750 audio files (20% of the dataset). To begin the CHM algorithm, we compute and store the spectrogram representations of each audio file in the training set. Then, we stratify each spectrogram by cutting each one into 8 horizontal strips, evenly spaced along the y-axis in order to preserve the frequency information of each spectrogram. Each stratum is then converted into a color histogram which captures the relative frequency

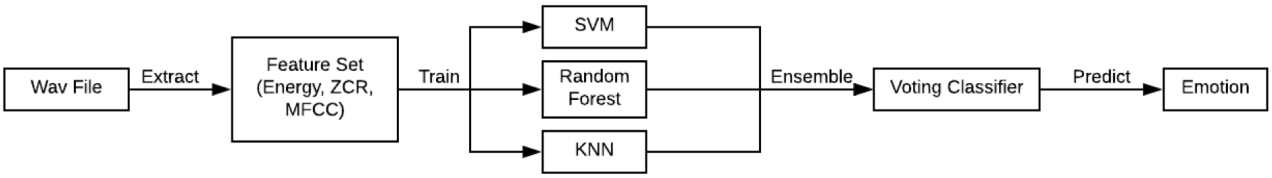

Fig. 4. Diagram of the Ensemble Classifier Data Pipeline

of each color present among all the strata. Centroids for each emotional class can be computed easily by taking the average relative frequency for each of the 36 bins of the color histograms at each frequency level of the respective class. For prediction, we can mimic the first several steps of computing the spectrogram representation of each audio file, stratifying it, and then computing the color histograms for each stratum. We then use Wassersteins distance to quantify how similar each stratums color histogram is to the centroid distributions of the respective levels of each emotional class and store this information in a vector. The vectors of centroid distances at each stratum level are then summed and the prediction is made based on the class with the minimum total Wasserstein distance to the original observation.

IV. CONVOLUTIONAL NEURAL NETWORK

A Convolutional Neural Network is type of deep, feed-forward Neural Network that works by sliding small filters across the input image. The main distinguishing aspect of CNNs from typical feed-forward neural networks is their convolutional layers, in which the input images are represented by feature maps. Pooling and fully connected layers may be used to identify activation features, and finally the classification is made with a softmax layer.

To establish a baseline CNN for emotion classification, we used a model with two convolutional layers, two max-pooling layers, and one fully-connected layer followed by a softmax layer. The input images for the CNN were the spectrograms generated from the raw audio files in the dataset (each resized to be 256 x 256). The first convolutional layer consisted of 16 5 x 5 kernels and the second convolutional layer consisted of 36 5 x 5 kernels. Each convolutional layer was followed by a max-pooling layer consisting of 2 x 2 kernels with a stride setting of 2. The following FC layer consisted of 128 neurons and the final softmax layer consisted of 7 neurons corresponding to the 7 emotional classes represented in the dataset.

V. ENSEMBLE CLASSIFIER

While the purpose of the convolutional neural network was to compare the Color Histogram Matching method to another model that uses spectrograms for audio analysis, we also wanted to compare the model to other classifiers with traditional audio features. To do this, we created an ensemble classifier of three models (Support Vector Machines, Random Forest, and K Nearest Neighbors). See Figure 4 above for a diagram of the ensemble classifier data pipeline.

A. Audio Feature Extraction

To conduct our feature extraction directly from the .wav files of the utterances from the Berlin Emotional Dataset, we utilized YAAFE (Yet Another Audio Feature Extractor) by means of a Python module. Based on [Seehapoch], we identified three particular features we wanted to focus on: Energy, Zero Crossing Rate, and MFCC. Energy is calculated as the root mean square of an audio feature sample. Zero crossing rate is defined as the weighted average of the number of times the speech signal changes sign within the time window [Seehapoch]. MFCC is a set of 13 coefficients for characterizing an audio based on characteristics of human hearing. Because each of these features was calculated for each frame of the audio clip, there were over 180 instances of the features generated for each 1-3 second audio clip. To handle this dimensionality, we took the mean, median, variance, minimum, and max of Energy and Zero Crossing Rate and the mean for each of the 13 MFCC coefficients, for a total of 23 features for each audio clip. Then, we normalized the feature set.

B. Classifiers and Ensemble

Based on a review of the literature surrounding speech emotion recognition as a classification problem

	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	24.41	1.574	1.574	37.01	25.20	5.512	4.724
Boredom	0	29.63	1.235	28.40	0	17.28	23.46
Disgust	0	2.174	23.91	32.61	19.57	4.348	17.39
Fear	2.899	8.696	2.899	57.97	15.94	2.899	8.696
Happiness	7.042	2.817	11.27	25.35	30.99	4.225	18.31
Neutral	0	17.72	0	20.25	1.266	30.38	30.38
Sadness	0	16.13	0	19.35	0	1.613	62.90

Fig. 5. Confusion Matrix for Color Histogram Matching algorithm

[7], [17], [18], we identified 3 potential classifiers that showed promise and ease of implementation. The following are the specifications of our models:

- Support Vector Machines, with c=10, default gamma, linear kernel;
- Random Forest, with n_estimators = 10, random_state = 42;
- K Nearest Neighbors, with n_neighbors = 5, uniform weights

Each of these was implemented using SkLearn package for Python. Each of the three classifiers was trained on the combined set of original and augmented data with 10-fold cross validation. The predictive power of the three classifiers was pooled using an ensemble method; in particular, SVM was weighted more heavily than Random Forest and KNN based on an inspection of prediction accuracy.

VI. RESULTS

A. Color Histogram Matching

Upon training, the color histogram matching approach achieved relatively low per-spectrogram classification accuracy. The overall test accuracy was about 37.17% and the individual class accuracies ranged from 23.91% to 62.9%. From the confusion matrix for CHM (shown in Figure 5), we can observe that audio samples belonging to the anger class were often confused with fear and happiness while disgust samples were often also confused with fear. Generally speaking, fear was a very common prediction by the model (over-predicted). One explanation for this phenomenon is that the stratified centroid representation of the fear class closely resembles the centroid representation of all spectrograms in the dataset. Upon further analysis, we see this is a viable reason since the cumulative Wasserstein distance from the fear centroid to the centroids of each of the other emotions is smaller than for any other emotion class.

B. Convolutional Neural Network

After training our baseline CNN model, we observed an overall per-spectrogram accuracy of about 50%. It is plausible that with a more sophisticated CNN architecture such as in [5], an overall per-spectrogram accuracy in the range of 60% could be obtained. From the resulting confusion matrix in Figure 6, we see that boredom is often confused with neutrality and both happiness and fear are often confused with anger. Just as fear was over-predicted by the CHM model, anger is over-predicted for the CNN model. The complexity of the CNN classification process makes it difficult, however, to pinpoint why this occurs. Overall, the CNN approach obtained a higher accuracy than the CHM method; however, it should be noted that the CHM method performed comparably or better on 4/7 of the emotion classes.

C. Ensemble Classifier

The ensemble classifier of Support Vector Machines, Random Forest, and K Nearest Neighbors, trained on the combined original and augmented dataset with 10-fold cross validation, produced a test accuracy of 27%, with the accuracy for each class ranging from 63% to 3%. It is clear that the CHM model performed better than this ensemble classifier. Based on the confusion matrix shown in figure 7, it is interesting to observe that the ensemble classifiers predicts Anger significantly more than any other emotion. Further experimentation beyond the scope of this paper would need to be done to determine why this is the case. Also, while the ensemble classifier has a 27% accuracy on the combined dataset, it had a 70% accuracy when trained only on the original audio clips. What this indicates is that the ensemble classifier is particularly sensitive to

	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	78.26	0	2.044	6.180	13.419	0.097	0
Boredom	5.121	35.70	5.441	2.952	0.6392	42.58	7.569
Disgust	20.51	1.498	46.79	4.211	5.310	10.355	11.326
Fear	39.47	4.489	1.312	42.55	0.8643	10.474	0.8407
Happiness	70.61	2.458	1.970	2.033	11.05	3.961	7.918
Neutral	7.753	9.447	1.343	4.350	5.839	61.783	9.485
Sadness	1.159	12.46	2.735	2.575	0.589	13.09	67.392

Fig. 6. Confusion Matrix for Convolutional Neural Network

	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	63.44	15.05	1.08	4.30	4.30	10.75	1.08
Boredom	40.98	24.59	1.64	9.84	6.56	14.75	1.64
Disgust	42.42	9.09	15.15	6.06	9.09	12.12	6.06
Fear	74.35	7.68	0	2.56	10.26	2.56	2.56
Happiness	48.89	2.22	2.22	11.11	24.44	6.66	4.44
Neutral	33.33	21.59	1.96	7.84	5.88	25.49	3.92
Sadness	51.92	7.69	0	9.62	9.62	11.54	9.62

Fig. 7. Confusion Matrix for Ensemble Classifier

the quality and cleanliness of the audio sample. The further implication is that the CHM model is more robust and better suited for real-world audio samples that may not of high quality.

VII. DISCUSSION

Although the CNN model had superior overall classification accuracy, the color histogram matching algorithm performed at a comparable level or better than the CNN for 4 out of the 7 emotions. In this regard, the CHM algorithm is much more computationally efficient than the CNN model, as it took roughly 30 minutes to train the CHM model on the 3500 spectrogram training set and nearly 4.5 hours to train the CNN model on the same data. Furthermore, the CHM algorithm is more robust than other modern machine learning approaches like those outlined in the ensemble approach. With the addition of data augmentation, the performance of CHM is practically unchanged, but for the ensemble approach, the performance declines significantly with slight changes in the raw audio data.

While there are multiple benefits of the CHM algorithm, it of course has several shortcomings as well. In its current state, the CHM model can only obtain a classification accuracy of nearly 40% - a mark that is significantly lower than concurrent deep learning approaches to SER. Additionally, even with stratification, some information is still lost regarding the frequency and a lot of information is lost regarding the time depicted in the spectrograms. Since the audio test clips are so condensed, the CHM algorithm essentially discards information pertaining to the time aspect of each spectrogram. For clips longer than about 1 second this can become problematic if there are clear patterns in the shapes of frequency peaks within the spectrograms.

To solve the problem of classifying audio clips longer than 1-2 seconds, we could partition a longer audio clip into a chain of 1-2 second increments, then make predictions for each segment. This could help to model the semantic context of each very short clip since the overall phrase would be broken down into multiple utterances. We can then reason that if the per-spectrogram classification accuracy is at least 30% as proposed in [5], this method would likely obtain sufficient classification accuracy on longer phrases, assuming the same emotion is being conveyed throughout the phrases. Intuitively, we can treat this process as a binomial distribution, with the classification accuracy for each emotion being the probability of successfully classifying each 1-second segment, and the number of

Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
54-69%	66-80%	53-67%	98-99%	69-82%	68%-81%	99-100%

Fig. 8. Potential Accuracy Scores for Color Histogram Matching

experiments being the number of 1-second segments that compose the phrase. If typical phrases range from 7-9 seconds in length, we would only need to observe 2-3 successes to dominate the majority of votes for the phrase classification. From this, we see that the likelihood of correctly classifying the overall phrase increases dramatically (see Figure 8).

VIII. CONCLUSION

Speech emotion recognition is an open problem with numerous applications to the field of human computer interaction. In this paper, we presented an novel application of color histograms, a tool of image analysis, to conduct speech emotion recognition as a classification task. We compare the success of our Color Histogram Matching model with a convolutional neural network and an ensemble of other classifiers. While CHM did not outperform the neural network, it works significantly faster and requires very little computation time. In fact, the CHM did outperform the ensemble classifier and was more resistant against imperfections in the audio data.

Because we outlined an entirely original approach to speech emotion recognition, there are many opportunities for extending this research. First, future work ought to focus on improving the accuracy of the CHM model. One could also investigate the impacts of using longer audio clips or expanding to more than 7 emotions. It would valuable to replicate the results of this paper on a dataset that is either in English or larger in size, such that it does not require audio augmentation. In summary, the application of image analysis techniques to the problem of speech emotion recognition is an interesting and potentially valuable avenue of research.

APPENDIX

As a team, we worked extraordinarily well together. This project went very smoothly and successfully because of the collaboration of Team 4. In terms of a big picture work breakdown, Dylan formulated and implemented the Color Histogram Method (18 hours), as well as the convolutional neural network (20 hours). He also led the efforts of data preprocessing, including audio augmentation and spectrogram generation (8 hours). Ryan spearheaded the extraction of audio features from the data (8 hours) and the development of the ensemble classifier (12 hours). He also took the lead on the communications of the project (16 hours), specifically the progress reports, presentation outline, and this report. Both team members contributed to the communication efforts, and Ryan handled the integration and coordination. While we are specifying tasks here, there are a lot of intangible tasks that we both worked on including problem formulation, review of literature, and analysis of findings (many more hours).

ACKNOWLEDGMENTS

We would like to express our thanks to our professors, Dr. James Wang and Dr. Jia Li, for their support and encouragement for this term project. We would also like to thank Yukun Chen, our project coach, who introduced us to this topic and provided us with helpful guidance throughout the entirety of the semester.

REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition, vol. 44, no. 3, pp. 572587, Mar. 2011.

[2] S. Ramakrishnan and I. M. M. El Emary, Speech emotion recognition approaches in human computer interaction, Telecommunication Systems, vol. 52, no. 3, pp. 14671478, Mar. 2013.

[3] Y. Kim, H. Lee, and E. M. Provost, Deep learning for robust feature generation in audiovisual emotion recognition, 2013, pp. 36873691.

[4] T. Vogt, E. Andr, and J. Wagner, Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation, in Affect and Emotion in Human-Computer Interaction, vol. 4868, C. Peter and R. Beale, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 7591.

[5] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network, 2017, pp. 15.

[6] Y. Pan, P. Shen, and L. Shen, Speech emotion recognition using support vector machine, International Journal of Smart Home, vol. 6, no. 2, pp. 101108, Apr. 2012.

[7] T. Seehapoch and S. Wongthanavasu, Speech emotion recognition using Support Vector Machines, 2013, pp. 8691.

[8] B. Schuller, G. Rigoll, and M. Lang, Hidden Markov model-based speech emotion recognition, 2003, vol. 2, pp. II-14.

[9] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Marino, Speech emotion recognitino using hidden Markov models, in Conference on Speech Communication and Technology, Denmark, 2001, pp. 26792682.
[10] Yi-Lin Lin and Gang Wei, Speech emotion recognition based on HMM and SVM, 2005, pp. 4898-4901 Vol. 8.
[11] V. Chernykh, G. Sterling, and P. Prihodko, Emotion Recognition from Speech with Recurrent Neural Networks, CoRR, vol. abs/1701.08071, 2017.
[12] Y. Niu, D. Zou, Y. Niu, Z. He, and H. Tan, A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks, CoRR, vol. abs/1707.09917, 2017.
[13] W. Lim, D. Jang, and T. Lee, Speech emotion recognition using convolutional and Recurrent Neural Networks, 2016, pp. 14.
[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, A database of German emotional speech, in 9th European Conference on Speech Communication and Technology, vol. 5, 2005, pp. 15171520.
[15] M. J. Swain and D. H. Ballard, Indexing via color histograms, 1990, pp. 390393.
[16] E. Jones, T. Oliphant, P. Peterson, and others, SciPy: Open source scientific tools for Python. 2001.
[17] J. Rong, G. Li, and Y.-P. P. Chen, Acoustic feature selection for automatic emotion recognition from speech, Information Processing & Management, vol. 45, no. 3, pp. 315328, May 2009.
[18] C. Busso et al., Analysis of emotion recognition using facial expressions, speech and multimodal information, 2004, p. 205.