Analysis of Human Emotion via Speech Recognition Using Viola Jones Compared with Histogram of Oriented Gradients (HOG) Algorithm with Improved Accuracy

Mahitha Sree E. and Nagaraju V.
Saveetha University, Chennai, India

Keywords: Human Emotion, Novel Viola Jones, Histogram of Oriented Gradients, Accuracy, Speech Signal, Research, Communication.

Abstract: The objective of this study is to enhance the precision in predicting human emotions through speech signals. This is achieved by introducing a novel approach, the Viola Jones (VJ) method, in contrast to the conventional Histogram of Oriented Gradients (HOG) algorithm. In this research we used Toronto Emotional Speech Set (TESS) as a dataset for this with a G-power of 0.8, alpha and beta values of 0.05 and 0.2, and a Confidence Interval of 95%, sample size is calculated as twenty (ten from Group 1 and ten from Group 2). Viola Jones (VJ) and Histogram of Oriented Gradients, both with the same amount of data samples (N=10), are used to perform the prediction of human emotion recognition from speech signals. The performance of the proposed viola jones is much greater than the accuracy rate of 88.65 percent achieved by the histogram of oriented gradients classifier. This is because the success rate of the proposed viola jones is 95.66 percent. The level of significance that was assessed to be attained by the research was $p = 0.001$ ($p<0.05$) which infers the two groups are statistically significant. For the performance evaluation of human emotion classification from speech data, the proposed Viola Jones (VJ) model achieves a greater level of precision than Histogram of Oriented Gradients (HOG).

1 INTRODUCTION

Automatic detection of speech emotions is a very recent Human Computer Interaction (HCI) field of study (Shukla et al. 2022) . As computers have become a vital part of our lives, the need for a more natural interface for human-computer communication has increased. Speech recognition systems attempt to facilitate communication between humans and machines (Meyer and Wiesmann 2006). In the construction of Human–Computer Interface (HCI) frameworks, emotion identification from speech signals is a subject of extensive investigation since it offers insights into human mental states. Identifying the emotional state of persons as cognitive feedback is frequently necessary in HCI. This study compares the outcomes of a novel Viola Jones (VJ) method for recognizing human emotional speech with the Histogram of oriented gradients (HOG) algorithm (Kirana, Wibawanto, and Herwanto 2018) (G. Ramkumar et al 2022). The classification outcomes of the Viola Jones (VJ) and Histogram Of Oriented Gradients (HOG) classifiers are analyzed for comparison purposes. The speech signal has been of interest to researchers for decades due to its multiple uses, such as emotion perception, HCI, fingerprints, etc (Junqua and Haton 2012) (Padma, S et al. 2022).

There has been a lot of work in the past few years on emotion recognition using speech data (Lin and Wei 2005; Zvarevashe and Olugbara 2020; Kerkeni et al. 2020; Shami and Verhelst 2007; Gao et al. 2017). There are 145 research papers available on IEEE Xplore, and 133 articles in Google Scholar. Otsuka and Ohya implemented local eye and mouth regions for emotion detection (Otsuka and Ohya 1998). However, the above approach has issues with noise and data loss. To identify emotions, Wang et al. (Wang et al. 2006) performed geometric displacements, specifically the manual placement of features extracted as lines and dots around the eyes, eyebrows, and lips. Unfortunately, this method does not lend itself to automatized prediction of feature points. With the use of colorful plastic dots, Kaliuby and Robinson were able to clearly distinguish facial muscle moments in the photograph (Kaliouby, El Kaliouby, and Robinson). It's more accurate than the

approaches we've been using up until now. When it comes to true human-computer interaction, however, manual point labeling isn't adequate. Z. Han and J. Wang (Han and Wang 2017) proposed a method for emotion recognition in spoken language using SVM and Gaussian Kernel Nonlinear Proximal SVM. Hugo L. Rufiner, M. Albornoz, and Diego H. Milone (Albornoz, Milone, and Rufiner 2017) A significant challenge in creating humanlike voice interface systems is the analysis of emotional states. The time-frequency analysis of an audio signal provides the research characteristic. Alionte and Lazar (Alionte and Lazar 2015) used the computer vision toolbox in MATLAB to create a Viola-Jones-based cascade face detector similar to the Haar face detector. Face detection in Python was written by Adouani et al. (Adouani, Ben Henia, and Lachiri 2019) using Haar-like cascade, LBP, and HOG in OpenCV and Dlib with SVM. Four emotions like happiness, sorrow, anger, and fear were identified by Frank Dellaert and colleagues using their own dataset (Dellaert, Polzin, and Waibel 1996). They used a total of 17 features chosen from 5 classes and three different methods (MLB classifier, KR, and KNN) with the latter yielding the best results. C. H. Wu et al. (Wu, Lin, and Wei 2014) presented an overview of the theoretical and empirical attempts that give different and complete views of the most new findings in emotion detection from bimodal data, which combines facial and voice expressions. In (Ooi et al. 2014), an unique architecture for an intelligent audio emotion identification system was proposed. This architecture's design module fully incorporates prosodic and spectral characteristics.

The most significant drawbacks of using Hog features are, first, that it has a slow training pace, and second, that it is very sensitive to noisy input, which might result in a poor final classification performance. In this study, a novel Viola Jones (VJ) classifier is developed with the goal of resolving this issue. The results of this classifier are compared with those of the Histogram of Oriented Gradients (HOG) classifier. The recognition performance of the speech recognition model that is based on VJ is shown to be superior to that of the HOG model, as shown by the results of the experiments.

2 MATERIALS AND METHODS

The Research was done in the Computer Science and Engineering Department's Software Laboratory at Saveetha Institute of Medical and Technical Sciences. Toronto Emotional Speech Set (TESS) repository is where the dataset was obtained for this research. The database is divided in such a way that 75% of it is taken for training, and the rest 25% is for testing. The two algorithms were divided into 2 Groups each with a sample Size of 10. Python is the software that is used for the online buying prediction model, and it is this software that generates the output. The sample size was determined by using previous research from (Jason, Kumar, and Others 2020) at clinicalc.com.

Histogram of Oriented Gradients (HOG)

HOG is often used to extract texture-based information from photos. Its purpose is to extract the images' local features for further analysis. Human Object Grammar (HOG) was created by Dalal and Briggs and was primarily used for human recognition. In terms of both brightness and invariance, it possesses the strongest texture characteristics possible. Human Oriented Gradient (HOG) is a potent approach for detecting pedestrians and objects. Not only are HOG features capable of directly adjusting to variations in lighting, but they also have the additional property of being geometrically invariant. This descriptor can be implemented by segmenting the speech into smaller connected sections (cells) and then creating a histogram of gradient directions or edge orientations for the pixels within each cell. When added together, these histograms stand in for the HOG descriptor. The goal of this approach is to draw out HOG characteristics. You can utilize these characteristics in your classifications. As HOG is a rotation-invariant descriptor, it has found application in both optimization and computer vision settings. The Pseudocode for HOG algorithm is given in Annexure.

Viola Jones (Vj)

The Viola–Jones feature extraction technique is among the most widely used. This algorithm was invented by Viola and Jones in 2001, and its benefits include high performance and rapid processing speed. This algorithm consists of Haar features, an integral picture, Adaboost, and a cascade classifier. For lip image tracking, the Viola–Jones algorithm was utilized. The Viola-Jones algorithm is an algorithm for quick speech detection. It detects emotions using a cascade of weak classifiers rather than a single strong classifier. Using the Viola–Jones method, faces are retrieved from speech signals of subjects. Using the retrieved facial images and the Viola–Jones algorithm, the lip images are located. The primary premise of the ViolaJones method is to scan sub windows inside an image to locate things of interest

across an area. It offers a rapid and precise framework for use in real-time object identification applications. This study uses the Viola-Jones emotion detection method to detect the audio and lip region. Viola and Jones discuss the algorithm's steps discussed in Annexure.

Statistical Analysis

The generated output is produced using Python software (Milano 2013). The training of these datasets necessitates a display resolution of 1024x768 pixels, on a system featuring a 10th generation Intel Core i5 processor, 12GB of RAM, and a 500 GB HDD. To conduct a thorough statistical analysis of the VJ and HOG algorithms, we utilize SPSS (Pallant 2010). SPSS is employed to perform computations of means, standard deviations, and standard errors of means. An independent sample t-test is executed through SPSS, facilitating a comparison of the two sets of data. The accuracy serves as the dependent variable, while inter scale matrix, intra scale matrix, and covariance stand as the two independent variables.

3 RESULTS

Figure 1 compares the accuracy of the VJ classifier and the HOG classifier. The accuracy rate of the VJ prediction model is higher than that of the HOG classification model, which is 88.65. The VJ classifier is notably distinct from the HOG classifier (test of independent samples, p 0.05). Along the X-axis, the VJ and HOG precision rates are displayed. Y-axis: Mean keyword identification precision, ±1 SD, with a confidence interval of 95 percent.

Table 1 presents the performance measurements for the comparison of VJ and HOG classifiers. The VJ classifier has a 95.66 percent accuracy rate, whereas the HOG algorithm has a rating of 88.65 percent. The VJ classifier is more accurate than the HOG when predicting human emotion from a voice input.

The computations for the VJ and HOG classifiers, including mean, standard deviation, and mean standard error, are displayed in Table 2. In the t-test, the accuracy level parameter is utilized. The Proposed technique has an average accuracy of 95.66%, whereas the HOG classification algorithm has an average accuracy of 88.65%. Standard Deviation for VJ is 0.1553, whereas the HOG method yields a value of 3.5356. VJ's Standard Error is 0.1905 on average, but the HOG method is 0.6355.

The statistical computations for VJ's independent variables in comparison to the HOG classifier are presented in Table 3. The level of significance for the accuracy rate is 0.001. Using a significance threshold of 0.98452 and a confidence interval of 95%, the VJ and HOG algorithms are compared using the independent samples T-test.

Table 1. The performance measurements of the comparison between the VJ and HOG classifiers are presented. The VJ classifier achieves a precision of 95.66%, while the HOG classification algorithm demonstrates an 88.65% accuracy level. With a greater rate of accuracy, the VJ classifier surpasses the HOG in predicting human emotion from speech signals.

SI.No.	TEST SIZE	ACCURACY RATE (in %)	
		VJ	HOG
1	Test1	94.23	86.70
2	Test2	94.44	86.83
3	Test3	94.56	87.19
4	Test4	94.84	87.32
5	Test5	95.12	87.52
6	Test6	95.16	87.71
7	Test7	95.24	87.85
8	Test8	95.26	88.08
9	Test9	95.35	88.18
10	Test10	95.54	88.34
Average Test Results		95.66	88.65

Table 2. The VJ and HOG classifiers undergo statistical analysis, encompassing metrics such as mean, standard deviation, and mean standard error. The accuracy metric serves as a crucial factor in the t-test. In terms of accuracy, the Proposed method yields an average of 95.66 percent, in contrast to the HOG classification algorithm, which achieves an average accuracy of 88.65 percent. VJ has a Standard Deviation of 0.1553, and the HOG algorithm has a value of 3.5356. The mean of VJ's Standard Error is 0.1905, while the HOG method is 0.6355.

GROUP		N	MEAN	STANDARD DEVIATION	STANDARD ERROR MEAN
ACCURACY RATE	HOG	10	88.65	3.5356	0.6355
	VJ	10	95.66	0.1553	0.1905

Table 3. The statistical calculation for independent variables of VJ in comparison with the HOG classifier has been evaluated. The significance level for the rate of accuracy is 0.001. Using a 95% confidence interval and a significance threshold of 0.98452, the VJ and HOG algorithms are compared using the independent samples-t-test.

GROUP		Levene's Test for Equality of Variances		t-TEST FOR EQUALITY OF MEANS						
		F	Sig.	t	Df	Sig.(2-tailed)	Mean Diff	Std. Error Difference	95% CI (Lower)	95% CI (Upper)
Accuracy	Equal variances assumed	2.56	0.02	12.2	34	.001	9.627	0.9845	8.548	13.781
	Equal variances not assumed			10.23	33.40	.001	6.885	0.9845	6.57	12.87

PSEUDOCODE FOR HOG ALGORITHM

```
Import necessary libraries
#import necessary libraries
import cv2
import numpy as np
Initialize image path
#image path
img_path = 'emotion.jpg'
Read the input image
#read the image
img = cv2.imread(img_path)
Initialize the Histogram of Oriented Gradient (HOG) feature vector
#Initialize the HOG feature vector
hog = cv2.HOGDescriptor()
Compute the histogram gradient components
#compute the Histogram of Oriented Gradient components
hist_ grad = hog.compute(img)
Flatten the feature vector
#Flatten the feature vector
flattened_grad = hist_grad.reshape((1,-1))
Feed the feature vector to the model
#Feed the feature vector to the model
emotion = model.predict(flattened_grad)
Print the predicted emotion
#print the predicted emotion
print(emotion)
```

PSEUDOCODE FOR VIOLA-JONES ALGORITHM

Input: original test image
Output: image with face indicators as rectangles
for i & 1 to num of scales in pyramid of images do
Downsample image to create image;
Compute integral image, images
for j <- 1 to num of shift steps of sub-window do
for k < 1 to num of stages in cascade classifier do
for 1 < 1 to num of filters of stage k do
Filter detection sub-window
Accumulate filter outputs end for
if accumulation fails per-stage threshold then
Reject sub-window as face
Break this & for loop end if
end for
if sub-window passed all per-stage checks then
Accept this sub-window as a face
end if
end for
end for

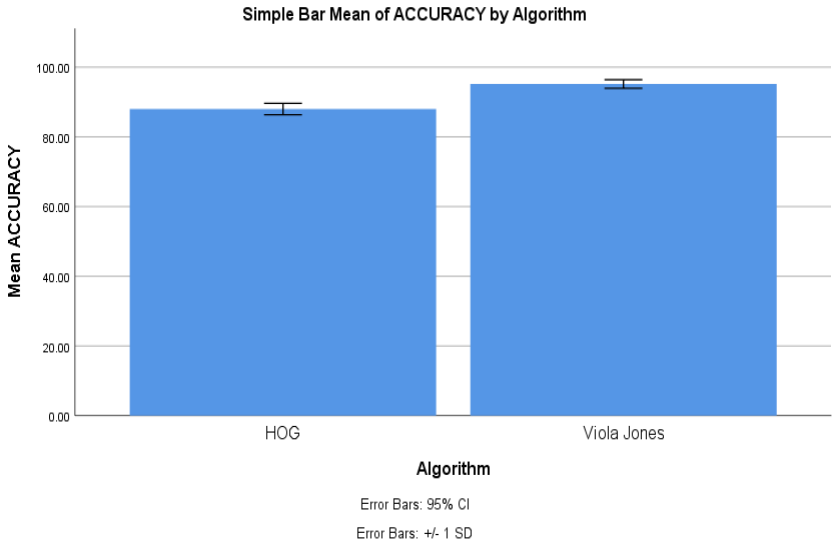

Fig. 1. Comparing the accuracy of the VJ classifier to that of the HOG algorithm has been evaluated. The Proposed method has a mean accuracy of 95.66 percent, whereas the HOG classification algorithm has a mean accuracy of 88.65 percent. The VJ prediction model has a greater accuracy rate than the HOG classification mode. The VJ classifier differs considerably from the HOG classifier (test of independent samples, $p<0.05$). The VJ and HOG precision rates are shown along the X-axis. Y-axis: Mean keyword identification accuracy, ± 1 SD, with a 95 percent confidence interval.

4 DISCUSSION

In order to determine emotions in human speech, a comparative analysis between the HOG algorithm and the novel viola Jones algorithm has been presented. An accuracy study has been carried out in order to determine the relative significance of each of the input characteristics. When compared to the HOG method, the output accuracy provided by VJ is significantly higher. VJ is an effective method for determining the emotions present in human speech. The accuracy of the output obtained by VJ is superior to that produced by the HOG approach. The accuracy of classifications as well as the amount of time saved by using VJ can be

considerably improved. This shows that the VJ algorithm is capable of achieving the highest level of accuracy in a short amount of time. The results of the experiment show that the proposed VJ technique performed better than the HOG model in terms of accuracy, as it attained a high level of 95.66 percent accuracy and exceeded the HOG method, which achieved 88.65 percent accuracy.

Some similar studies are Joseph Juliana and Sharmila (Julina, Kulandai Josephine Julina, and Sree Sharmila 2019) used HOG and LBP traits from face characteristics such the nose, eyes, & lips to study and identify the three emotions joyful, sad, & angry. They used texture characteristics to train a traditional neural network classification method, and the resulting accuracy was 86% for HOG features and 65% for LBP data. In 2019, A. Bhavan et al. (Bhavan et al. 2019) proposed a method for recognising emotional states in people's voices by the extraction of a small number of spectral features that have been preprocessed (MFCCs and spectral centroids). This method proposes using a bagged ensemble of SVMs with a Gaussian kernel as the classification model. Accuracy of 83.21 percent was found. Separately, the discriminant temporal pyramid mapping method was utilised to collect features in (Zhang et al. 2018) a study using Mel spectrogram and the AlexNet deep learning network. The gathered data showed that the pre-trained deep learning model performed effectively when processing emotional speech. (Prasomphan 2015) used synthetic neural networks and the EMO-five DB's emotions to suggest a new approach to emotion detection using a spectrum analyzer. Five out of the ten emotions had an 82% success rate.

The Viola-Jones algorithm has the drawback that it is difficult to detect emotions when the background signal is complicated or when there are several noises present, and it also has a low detection rate. These are both limitations. Future work has to pay more attention to a wider range of emotional types. The system's ability to interpret the relevance of the speech signal would be an added bonus.

5 CONCLUSION

The model that is being suggested exhibits both the VJ and the HOG, with the VJ having obtained higher accuracy values than the HOG as a result of its use. The HOG has just a 88.65% accurate accuracy rating, however the VJ has an accuracy rating that is 95.66% more accurate than that of the HOG in an analysis of human emotion via voice signal with an enhanced accuracy rate.

REFERENCES

Adouani, Amal, Wiem Mimoun Ben Henia, and Zied Lachiri. (2019). “Comparison of Haar-Like, HOG and LBP Approaches for Face Detection in Video Sequences.” In *2019 16th International Multi-Conference on Systems, Signals Devices (SSD)*, 266–71.

Albornoz, Enrique M., Diego H. Milone, and Hugo L. Rufiner. (2017). “Feature Extraction Based on Bio-Inspired Model for Robust Emotion Recognition.” *Soft Computing* 21 (17): 5145–58.

Alionte, Elena, and Corneliu Lazar. (2015). “A Practical Implementation of Face Detection by Using Matlab Cascade Object Detector.” In *2015 19th International Conference on System Theory, Control and Computing (ICSTCC)*, 785–90.

Bhavan, Anjali, Pankaj Chauhan, Hitkul, and Rajiv Ratn Shah. (2019). “Bagged Support Vector Machines for Emotion Recognition from Speech.” *Knowledge-Based Systems* 184 (November): 104886.

Dellaert, F., T. Polzin, and A. Waibel. (1996). “Recognizing Emotion in Speech.” In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP ’96*, 3:1970–73 vol.3.

Gao, Yuanbo, Baobin Li, Ning Wang, and Tingshao Zhu. (2017.) “Speech Emotion Recognition Using Local and Global Features.” In *Brain Informatics*, 3–13. Springer International Publishing.

G. Ramkumar, G. Anitha, P. Nirmala, S. Ramesh and M. Tamilselvi, "An Effective Copyright Management Principle using Intelligent Wavelet Transformation based Water marking Scheme," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2022, pp. 1-7, doi: 10.1109/ACCAI53970.2022.9752516.

Han, Zhiyan, and Jian Wang. (2017). “Speech Emotion Recognition Based on Gaussian Kernel Nonlinear Proximal Support Vector Machine.” In *2017 Chinese Automation Congress (CAC)*, 2513–16.

Jason, C. Andy, Sandeep Kumar, and Others. (2020). “An Appraisal on Speech and Emotion Recognition Technologies Based on Machine Learning.” *Language* 67: 68.

Julina, J. Kulandai Josephine, J. Kulandai Josephine Julina, and T. Sree Sharmila. (2019). “Facial Emotion Recognition in Videos Using HOG and LBP.” *2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*. https://doi.org/10.1109/rteict46194.2019.9016766.

Kaliouby, R. El, R. El Kaliouby, and P. Robinson. n.d. “Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures.” *2004*

Conference on Computer Vision and Pattern Recognition Workshop. https://doi.org/10.1109/cvpr.2004.427.

Kerkeni, Leila, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder. (2020). "Automatic Speech Emotion Recognition Using Machine Learning." In *Social Media and Machine Learning*, edited by Alberto Cano. London, England: IntechOpen.

Kirana, Kartika Candra, Slamet Wibawanto, and Heru Wahyu Herwanto. (2018). "Emotion Recognition Using Fisher Face-Based Viola-Jones Algorithm." In *2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 173–77. ieeexplore.ieee.org.

Lin, Yi-Lin, and Gang Wei. 2005. "Speech Emotion Recognition Based on HMM and SVM." In *2005 International Conference on Machine Learning and Cybernetics*, 8:4898–4901 Vol. 8. ieeexplore.ieee.org.

Milano, Federico. 2013. "A Python-Based Software Tool for Power System Analysis." In *2013 IEEE Power Energy Society General Meeting*, 1–5.

Ooi, Chien Shing, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew. 2014. "A New Approach of Audio Emotion Recognition." *Expert Systems with Applications* 41 (13): 5858–69.

Otsuka, T., and J. Ohya. (1998). "Spotting Segments Displaying Facial Expression from Image Sequences Using HMM." In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 442–47.

Padma, S., Vidhya Lakshmi, S., Prakash, R., Srividhya, S., Sivakumar, A. A., Divyah, N., ... &Saavedra Flores, E. I. (2022). Simulation of land use/land cover dynamics using Google Earth data and QGIS: a case study on outer ring road, Southern India. Sustainability, 14(24), 16373

Pallant, Julie. (2010). "SPSS Survaival Manual: A Step by Step Guide to Data Analysis Using SPSS." McGraw-Hill Education. http://dspace.uniten.edu.my/handle/123456789/17829.

Prasomphan, Sathit. (2015). "Improvement of Speech Emotion Recognition with Neural Network Classifier by Using Speech Spectrogram." In *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 73–76.

Shami, Mohammad, and Werner Verhelst. (2007). "An Evaluation of the Robustness of Existing Supervised Machine Learning Approaches to the Classification of Emotions in Speech." *Speech Communication* 49 (3): 201–12.

Wang, Jun, Lijun Yin, Xiaozhou Wei, and Yi Sun. (2006). "3D Facial Expression Recognition Based on Primitive Surface Feature Distribution." In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:1399–1406.

Wu, Chung-Hsien, Jen-Chun Lin, and Wen-Li Wei. (2014). "Survey on Audiovisual Emotion Recognition: Databases, Features, and Data Fusion Strategies." *APSIPA Transactions on Signal and Information Processing* 3: e12.

Zhang, Shiqing, Shiliang Zhang, Tiejun Huang, and Wen Gao. (2018). "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching." *IEEE Transactions on Multimedia* 20 (6): 1576–90.

Zvarevashe, Kudakwashe, and Oludayo Olugbara. (2020). "Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition." *Algorithms* 13 (3): 70.