Enhanced Media Player using Face Recognition, Hand Gestures and Voice Detection

Ruchita Paithankar, Aditi Pusalkar, Jayshree Saindane, Sinu Mathew

Department of Computer Engineering, University of Mumbai, Atharva College of Engineering, Malad, Mumbai, India.

Publication Info

Article history:
Received : 19 February 2020
Accepted : 23 May 2020

Keywords:
Computer vision, Face detection, Haar Cascade, Hand gestures, Human-computer interaction (HCI), Media player, Voice detection, Web camera.

****Corresponding author:***
Ruchita Paithankar
e-mail: *ruchitapaithankar@gmail.com*

Abstract

When watching a video on a computer, many interruptions can distract the user away from the system, e.g., a laptop or a desktop. This causes an important part of the video to be missed. We are building up a media player which is one step ahead of the usual media players. It plays and pauses the video by identifying the client's face, utilizing a web camera. If the user is looking at the screen, then the video is not interrupted. In case if the user is not looking, and the system couldn't detect the user's face, then it immediately stops the video. We are adding additional functionality to control other features of our enhanced media player, such as increasing and decreasing the volume, forwarding, and backwinding the video, using hand gestures and voice detection.

1. INTRODUCTION

Normally when you are viewing a video, and somebody calls you, you need to look elsewhere or leave from pc for quite a while, so you miss some piece of the video. Later you have to haul back the video from where you saw it. With the improvement of universal registering, current client connection approaches with console, mouse, and pen are not adequate. Because of the restriction of these gadgets, the useable order set is likewise constrained.

The prime objective of the project is to create a system that can recognize human-initiated hand gestures, detection of face and voice, and make use of this data for device control. The user plays out a signal before a camera, which is connected to the PC. The image of the gesture is then handled to distinguish the motion showed by the user. When the motion is distinguished, relating control activity appointed to the gesture is incited.

The media player will also work on voice commands, e.g., If a voice command is given like play, then the media player will play the video; if it says pause, then it will pause the video. Hidden Markov algorithm is used where a raw speech is transmitted to it a text signals are generated.

2. LITERATURE REVIEW

The survey, face detection to filter selfie face image on Instagram [2], is to strain and refine face images of a selfie, on search results based on hashtags on Instagram. This is done by integrating two techniques using the Haar Cascade method, the web information extraction strategy, and the human face recognition system. The face detection frameworks meant to diminish the bogus positive rate and increment the exactness of distinguishing face, particularly in compounded and complicated foundation pictures. The proposed study presents a timely analysis and evaluation of face detection techniques comprising characteristic-based, aspect-based, theory-based, and arrangement comparing [4]. Our project is based on the task of face detection. Viola and Jones proposed a statistical point of view to manage the diversity in human face structures. For that, in their algorithm, the forming of an "integral image" is employed to reason an expensive array of Haar-like options. In their algorithm, the construct of an "integral image" is employed to reason an expensive deposit of Haar-like options. There is a database in which various hand gestures are stored. When a user shows a particular hand gesture to the application, the application searches the particular hand gesture in the database and then provides the output.

The paper on research on the hand gesture recognition based on deep learning comprehends the division of hand signals by setting up the skin color model and AdaBoost classifier established on Haar in consonance with the peculiarity of skin color for hand gestures. While handing problems with noisy data, the study, fuzzy neural network (FNN) with audio-visual data for voice activity detection in noisy environments [1], validates an FNN to exercise the unpredictability. By scrutinizing and examining the geometrical shapes, the lip outlines ad curves attribute and characteristic of the person who is speaking, can be withdrawn. Then, both audio and visual information is considered by the proposed fuzzy neural network. In our venture, we are also using speech recognition. The media player will work on voice commands. Hidden Markov algorithm is used where a raw speech is transmitted to it a text signals are generated.

3. METHODOLOGY

The initial step of the implementation process is that the user will show his/her face or a gesture to the webcam, which should be persistent for an interval of time. This is obligatory for dynamic computing. A set of gestures is already defined as the correct gesture for processing. The gestures shown by the user should match this predefined set. Haar cascade XML files are used for face detection. These files have a lot of feature sets corresponding to a very specific type of use case. The web camera captures the human-generated hand gesture or the user's face, and stores it in memory. A package is used for storing the image in memory and again calling the same program after a particular interval. This package is called Emgu-CV.

The captured image is firstly pre-processed. The techniques which are used are like color space detection, skin color detection using OpenCV [Emgu–CV wrapper], color space conversion [YCrCb, HSV, RGB], and differentiation, and finally, for finger detection, line segment detection is used.

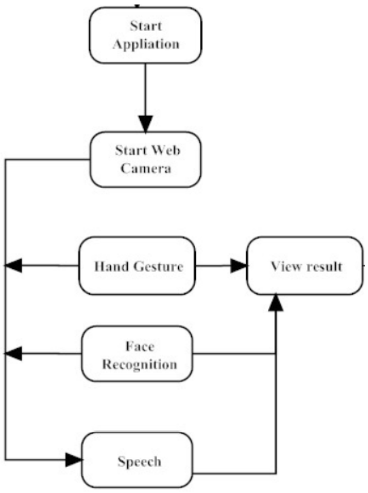

Figure 1: State chart

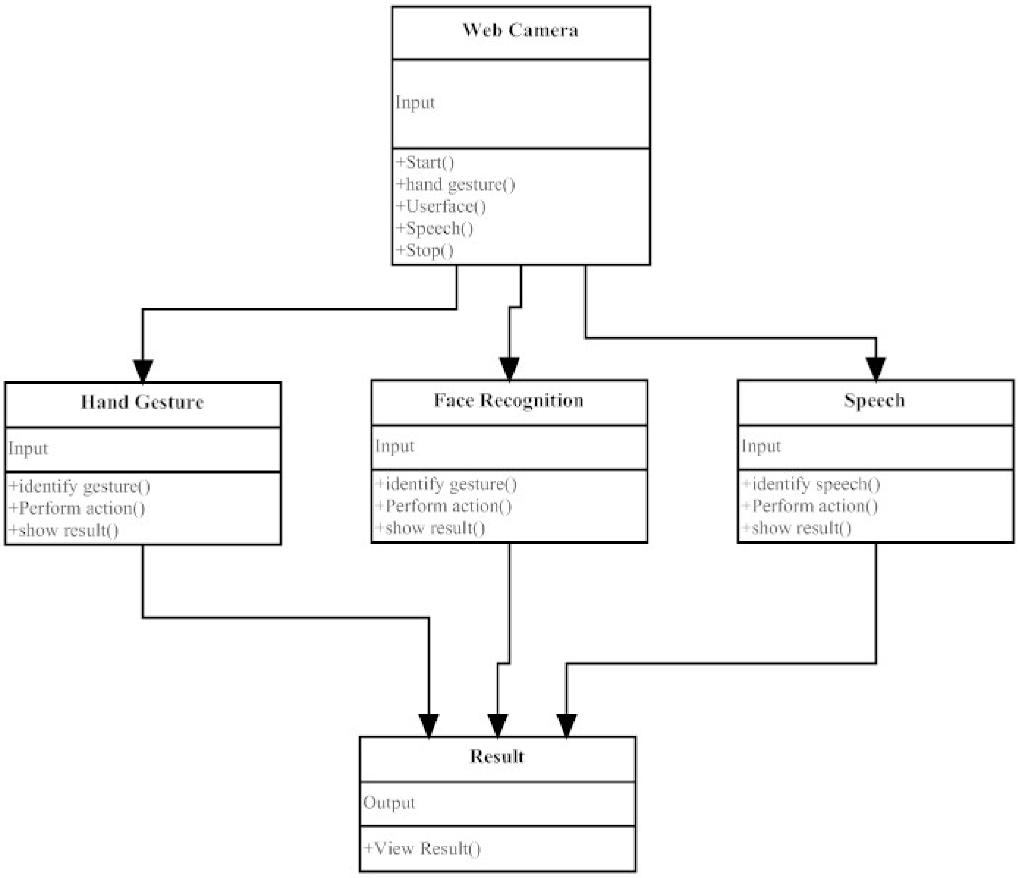

Figure 2: Class diagram

The algorithm will first either detect the face or count the number of fingers shown by the user. This will work as the input for further processing. Once the gesture is recognized, or the user's face is detected, the appropriate command for it will be executed. Our system then goes back to capturing images.

3.1. Haar Cascade Classifier

In the Viola-Jones object detection structure, the Haar-like highlights are sorted out in a classifier course to shape a solid student or classifier. For object recognition, Haar-like highlights, which are advanced picture highlights, are utilized. They get their name from Haar wavelets, to which they are instinctive similar.

An aperture of the calculated size is transferred to the top of the information picture. For every minute segment of the information image, the Haar-like component is determined. This is called the identification span of the Viola-Jones object discovery system. This dissimilarity is then contrasted with an educated edge that differentiates non-objects from objects. Since such a Haar-like component is just a powerless student or classifier (its recognition quality is slightly better to irregular speculating), countless Haar-like highlights are significantly necessary to represent an item with sufficient accuracy. The Haar-like highlights are, in this way, collected in something many introduce to as a classifier course to shape a solid student or classifier, in the Viola-Jones object discovery system.

The key bit of leeway of a Haar-like element is its estimation rate. Because of the usage of vital pictures, a Haar-like component of any dimension can be resolved in constant time (roughly 60 chip directions for a two square shape include). Open CV's calculation is presently utilizing the accompanying Haar-like highlights, which are the contribution to the fundamental classifiers:

- Feature = w1 x RecSum(r1) + w2 x RecSum(r2).
- Weights can be definite or negative.
- Weights are accurately parallel to the region.
- Computed at each point and scale.

3.2. Speech Recognition

In our project, we are also including speech recognition. Voice commands will serve as an input for the media player, For e.g., if a voice command is given like play then the media player will play the video; if it says pause then it will pause the video; if it says play the previous one then it will play the desired clip and vice versa. Hidden Markov algorithm is used where a raw speech is transmitted to it a text signals are generated. When associate in nursing HMM is registered to speech recognition, the circumstances are decoded as acoustic models, showing what sounds are likely to be caught and perceived during their analogous portions of speech, while the transitions offer temporal constraints, indicating how the states may follow each other in sequence.

3.3. Hand Gesture Recognition

Initially, Viola and Jones proposed a demographic perspective to handle the great number of human faces for the task of face tracking and detection. In their algorithm, the construct of "integral image" is employed to reason an expensive collection of Haar-like options. There is a database in which various hand gestures are stored. When a user shows a particular hand gesture to the application, the application searches the particular hand gesture in the database and then provides the output. The gesture will be mapped according to the gray level values. Every gesture has a different task to be performed. The gesture could be a representation of physical conduct or emotional expression. It includes both body and hand gestures. It falls into two categories: static gesture and dynamic gesture.

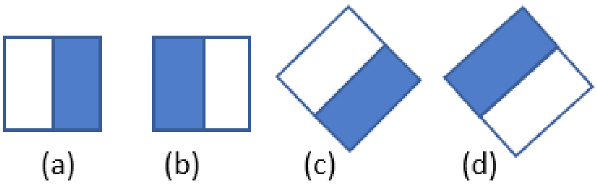

Figure 3: Line features

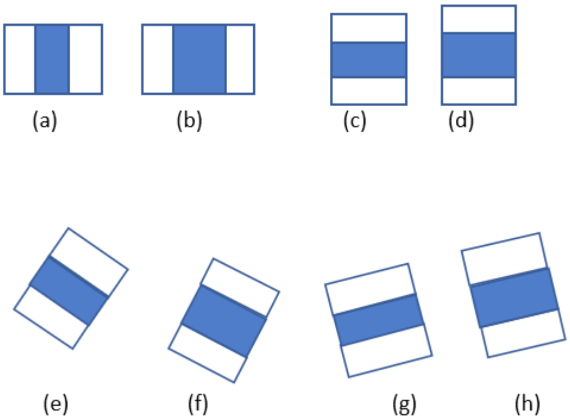

Figure 4: Edge features

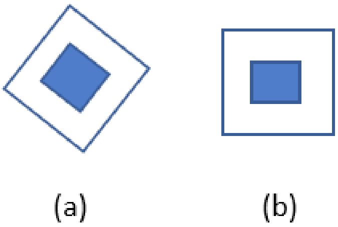

Figure 5: Centre surround features

Compared with different approaches that should care for numerous image scales, the integral image can do accurate scale invariableness by getting rid of the requirement to work out a multi-scale image pyramid and remarkably lowers the processing time of an image. The Viola and Jones rule is around fifteen times quicker than any previous approaches, whereas accomplishing certainty that's cherish the simplest printed results. Viola-Jones algorithm uses basic Haar-like features. The "integral image" at the situation of the pixel (x, y) contains the total of the constituent values higher than and left of this constituent that is comprehensive.

The "integral image" at the location of the pixel (x, y) holds the summation of the pixel numbers above and left of this pixel, which is comprehensive.

3.4. Face Detection

Face detection calculates a lot of features using all the probable sizes of locations of each kernel. For each attribute computation, we are required to obtain the summation of pixels under both white and black rectangles. To resolve this problem, they initiated the integral images. It streamlines computation of add of pixels, however giant is also the number of pixels, to a functioning requiring just four pixels. It is nice because it makes things fast-paced.

With all of these options, we have an inclination to calculated; most of them are unimportant. The first feature designated looks to target the attribute that the area of the eyes is usually darker than the part of the nose and cheeks. The second attribute designated depends on the trait that the eyes are darker than the bridge of the nose. But constant windows pertaining to cheeks or the other place is moot. It is achieved by Adaboost.

4. PROPOSED SYSTEM

This project consists of three methods by which our media player can be controlled, namely, face detection, speech recognition, and hand gesture recognition. To pause and play the video, face recognition is used. The additional functionalities of our media player used hand gestures as input commands. The user's voice is captured for speech recognition related features of the media player.

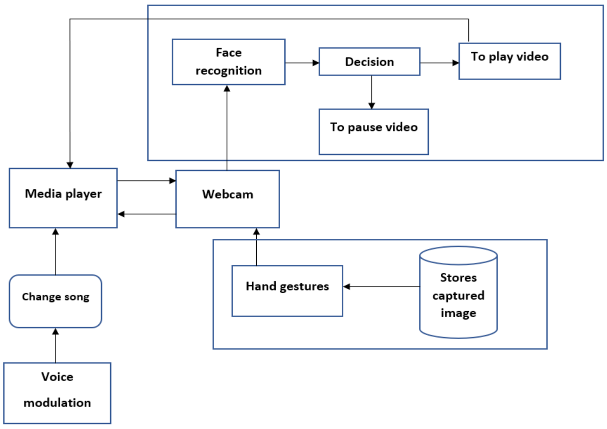

Figure 6: Proposed system

5. FUTURE SCOPE

The proposed system is limited for individual use only. It can be extended to two or more people. Also, it can be extended so that the algorithm can detect the emotions of the user and then suggest videos accordingly, which can then be played on our media player.

6. CONCLUSION

The primary concern of this proposed system is to facilitate the user to acquire the utmost experience of media player. To achieve this goal, we have tried to automate the media player considerably. We are doing this by using face recognition and hand gestures for running the wide-range of functions of the media player. These functions consist of pausing and playing the video using face recognition; and forwarding, backwarding, increasing, and decreasing; the volume of the video using hand gestures.

7. ACKNOWLEDGMENT

We would like to express our regards and respect towards our project guide Prof. Sinu Mathew, for her helpful assistance and supervision and for wholeheartedly supporting the project. Her technical guidance and constant motivation made this project a success. It was a great experience learning under her. We are also grateful to our principal Dr. S. P. Kallurkar, and HOD of Computer Department, Prof. Suvarna Pansambal, project coordinator Prof. Mamta Meena and Prof. Shweta Sharma, and all the faculty members of the Computer Department, who guided us throughout to develop this project idea.

8. REFERENCES

[1] G. Wu and Z. Zhu, "Fuzzy Neural Network with Audio-Visual Data for Voice Activity Detection in Noisy Environments," *2018 International Conference on Intelligent Autonomous Systems (ICoIAS)*, Singapore, 2018, pp. 141-145. doi: 10.1109/ICoIAS.2018.8494090

[2] J. Dey, M. S. Bin Hossain and M. A. Haque, "An Ensemble SVM-based Approach for Voice Activity Detection," *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*, Dhaka, Bangladesh, 2018, pp. 297-300. doi: 10.1109/ICECE.2018.8636745

[3] A. Priadana and M. Habibi, "Face Detection using Haar Cascades to Filter Selfie Face Image on Instagram," *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, Yogyakarta, Indonesia, 2019, pp. 6-9. doi: 10.1109/ICAIIT.2019.8834526

[4] J. Sun, T. Ji, S. Zhang, J. Yang and G. Ji, Research on the Hand Gesture Recognition Based on Deep Learning, "*2018 12th International Symposium on Antennas, Propagation and EM Theory (ISAPE)*, Hangzhou, China, 2018, pp. 1-4. doi: 10.1109/ISAPE.2018.8634348

[5] A. Sharifara, M. S. Mohd Rahim, and Y. Anisi, "A general review of human face detection including a study of neural networks and Haar feature-based cascade classifier in face detection, "*2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, Kuala Lumpur, 2014, pp. 73-78. doi: 10.1109/ISBAST.2014.7013097

[6] C. Czepa, S. Buchinger, H. Hlavacs, E. Hotop, and Y. Pitrey, "Towards an energy-efficient attention-aware mobile video player with sensor and face detection support," *2012 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, San Francisco, CA, 2012, pp. 1-6. doi:10.1109/WoWMoM.2012.6263801

[7] A. S. Ghotkar, R. Khatal, S. Khupase, S. Asati, and M. Hadap, "Hand gesture recognition for Indian Sign Language," 2012 International Conference on Computer Communication and Informatics, Coimbatore, 2012, pp. 1-4. doi: 10.1109/ICCCI.2012.6158807

[8] S. Sharma, S. Jain and Khushboo, "A Static Hand Gesture and Face Recognition System for Blind People," 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2019, pp. 534-539. doi: 10.1109/SPIN.2019.8711706

[9] S. Peng, K. Wattanachote, H. Lin, and K. Li, "A Real-Time Hand Gesture Recognition System for Daily Information Retrieval from Internet," 2011 Fourth International Conference on Ubi-Media Computing, Sao Paulo, 2011, pp. 146-151. doi: 10.1109/U-MEDIA.2011.45

[10] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Leveraging Large Face Recognition Data for Emotion Classification," 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, 2018, pp. 692-696. doi: 10.1109/FG.2018.00109

[11] L. Xu, L. Huang, and C. Liu, "Raw vs. Processed: How to Use the Raw and Processed Images for Robust Face Recognition under Varying Illumination," 2010 20th International Conference on Pattern Recognition, Istanbul, 2010, pp. 2692-2695. doi: 10.1109/ICPR.2010.660

[12] H. Kim, S. H. Lee, and Y. M. Ro, "Face image assessment learned with objective and relative face image qualities for improved face recognition," 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, 2015, pp. 4027-4031. doi: 10.1109/ICIP.2015.7351562

[13] N. L. Fitriyani, C. Yang, and M. Syafrudin, "Real-time eye state detection system using haar cascade classifier and circular hough transform," 2016 IEEE 5th Global Conference on Consumer Electronics, Kyoto, 2016, pp. 1-3. doi: 10.1109/GCCE.2016.7800424

[14] Y. Li, X. Xu, N. Mu, and L. Chen, "Eye-gaze tracking system by haar cascade classifier," 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), Hefei, 2016, pp. 564-567. doi: 10.1109/ICIEA.2016.7603648.

AUTHORS

Ruchita Paithankar
Student
Atharva College of Engineering

Aditi Pusalkar
Student
Atharva College of Engineering

Jayshree Saindane
Student
Atharva College of Engineering

Sinu Mathew
Professor
Atharva College of Engineering