Performing Media Operations using Feature Based Model and Gesture Recognition by Implementing Machine Learning Techniques.

[1]Jyoti Avhale, [2]Prof. Ashwini Gaikwad
[1][2]Department of Computer Science and Engineering
[1][2]Deogiri Institue of Engineering and Management Studies

Abstract

When someone calls while you are viewing a video, you have to look elsewhere or leave the computer for a while, which means you miss some of the film. You must drag the video back from the last place you saw it. A media player that pauses itself when the user is not gazing at it is called a Gaze Based Media Player with gesture recognition for the hands. The moment the user glances at it, the player resumes running. The web camera or camera module on top of the PC is used for this. With this project, we're creating a sophisticated media player that automatically plays and pauses videos based on facial recognition, and moves videos forward and backward based on hand gestures. Using a Haar-cascade Classifier, the system maintains record of whether the user is looking at the screen or not. The system is going to terminate the video if the user is not paying attention or if it is unable to identify the user's face. Convolutional Neural Networks are used to control the media player's other features, including playing the next and previous videos. Additionally, it has the capability of regulating media player features like noise detection and evaluation between machine output and input from the environment; if the input is greater, the media player will stop.

Introduction

Human Computer Interaction can acquire several advantages with the introduction of different natural forms of device free communication. Gestures are a natural form of actions which we often use in our daily life for interaction, therefore to use it as a communication medium with computers generates a new paradigm of interaction with computers. With the emergence of many natural types of device-free communication, human-computer interaction can gain a number of benefits. The use of gestures as a communication tool with computers creates a new paradigm of connection with computers since gestures are a natural form of activity that we frequently employ in our daily lives. We'll go through computer vision and gesture detection methods that were created on a low-cost, vision-based input device for controlling media players using gestures. The VLC programme has a central computational module that analyses gesture images using Principal Component Analysis to identify the gesture's feature vectors and save them as an XML file.

The Recognition of the gesture is done by Convolutional Neural Networks. The theoretical analysis of the approach shows how to do recognition in static background. This hand gesture recognition technique will not only replace the use of mouse to control the VLC player but also provide different gesture vocabulary which will be useful in controlling the application.. The face recognition is a technique to identify or verify the face from the digital images or video frame. A human can quickly identify the faces without much effort. It is an effortless task for us, but it is a difficult task for a computer. There are

various complexities, such as low resolution, occlusion, illumination variations, etc. These factors highly affect the accuracy of the computer to recognize the face more effectively.

Convolutional Neural Networks provide the gesture recognition. How to recognise objects against a static background is demonstrated by the theoretical study of the method. In addition to eliminating the need for a mouse to manage the VLC player, this hand gesture detection approach also offers a variety of gestures that can be used to control the programme. Face recognition is a technique for locating or authenticating the face in digital photographs or video frames. A human can easily and quickly recognise the faces. For us, it is a simple task, but for a computer, it is challenging. There are many difficulties, including low resolution, occlusion, different lighting conditions, etc.

Face detection and character recognition come in many different varieties. The Haarcascade Classifier has been utilised in this instance to detect faces. Our project's objective is to develop a sophisticated media player that utilizes facial expressions and hand motions. To accomplish the goal, we have set the following goals:

A. The media player's user interface needs to be effective and user-friendly.

B. The media player's output should be precise.

D. If the user's face cannot be quickly identified by the media player, the video will pause.

C. No video content is missing.

D. Accurately recording hand motions and performing actions that go along with them are essential.

When someone calls when you are viewing a video, you typically have to look elsewhere or step away from the computer for a while, which means you miss some of the movie. You must drag the video back from the last place you saw it. Here is a fix for this issue, though. a media player with a look-based system that pauses when the user isn't looking at it. The user looks at it again, and the player immediately begins to move again. The computer's built-in camera or webcam is used for this. The media is played as long as the camera recognises the user's face gazing at it. As soon as the user's face cannot be seen in its whole, the player pauses.

The Haar-like features are arranged in the Viola-Jones object identification framework in a structure known as a classifier cascade to create a powerful learner or classifier. A Haar-like feature's primary benefit over most other features is its calculation speed. Digital image properties that resemble Haars are employed for object recognition. They were employed in the first real-time face detector and got their name from how intuitively they resembled Haar wavelets. The target size window is dragged over the input image during the detection phase of the Viola-Jones object detection framework, and the Haar like feature is generated for each subsection of the image. Then, this difference is contrasted with a learnt threshold that distinguishes between objects and non-objects. A large number of Haar-like features are required to accurately describe an item because one such Haar-like feature is merely a weak learner or classifier (its detection quality is only marginally better than random guessing). The Haar-like features are arranged in the Viola-Jones object identification framework in a structure known as a classifier cascade to create a powerful learner or classifier. A Haar-like feature's primary benefit over most other features is its calculation speed. A Haar-like feature of any size can be determined in constant time by using integral pictures (approximately 60 microprocessor instructions for a 2-rectangle feature).

Literature Review

A significant research issue spanning many professions and disciplines is face recognition. This is due to the fact that face recognition is a fundamental human behaviour on the part that is necessary for successful human interactions and communication, in addition to having numerous practical applications like bankcard identification, access control, searching for mug shots, video surveillance, and surveillance systems. The first formal system for classifying faces was put forward in [1]. The authors suggest assembling face profiles as curves, calculating their average, and then classifying other profiles based on how far they differed from the average. This classification produces a vector of independent measures that can be compared to other vectors in a database since it is multi-modal. Face recognition technology has improved to the point where it is already being used in actual settings [2].

This observation provides us with the person's "Biometric Signature." (2) The biometric signature is "normalised" by a computer programme to match the format (size, resolution, view, etc.) of the signatures stored in the system's database. We obtain a "Normalized Signature" of the person by the normalising of their biometric signature. (3) A matcher evaluates the normalised signature in comparison to the set (or subset) of normalised signatures stored in the system's database and calculates a "similarity score" for each signature in the database set (or sub-set). The application of the biometric system determines what is performed with the similarity measure after that. Face recognition begins with the identification of face patterns in occasionally cluttered scenes, then moves on to normalising the facial image to account for simple geometric and lighting changes, possibly using knowledge of the position and visual appeal of facial landmarks. Finally, the results are post processed using model-based strategies and logistic feedback [3].

Face detection and normalisation and face identification are the two main components of every face recognition system. Fully automatic algorithms are those made up of both parts, whereas partially automatic algorithms are those made up of just the second part. The centre of the eyes' coordinates and a facial image are provided to partially automatic algorithms[4].
Just facial picture data is provided to fully automatic algorithms. On the other hand, as face recognition has advanced over the years, it is now possible to group recognition techniques into three categories, namely frontal, profile, and viewtolerant recognition, based on the type of photos and the algorithms. The traditional method of recognition is frontal, however view-tolerant algorithms typically execute recognition in a more sophisticated way by taking into account parts of the underlying physics, geometry, and statistics. Profile schemes have only a very minor role in identification as stand-alone systems.
They are, nevertheless, particularly useful for quick, rough pre-searches of big face databases to lessen the computing strain for a subsequent, complex algorithm, or as a component of a mix recognition scheme. These hybrid approaches, which combine various recognition techniques either in a serial or parallel way to tackle the shortcomings of the separate components, have a unique position among face recognition systems. Examining whether face recognition algorithms are based on models or exemplars is another approach to classify them. In [5] the Quotient Image is computed, and in [6] the Active Appearance Model is derived. These models give tight limitations when dealing with variance in appearance and capture classifier (the class face).

Exemplars can also be used for acknowledgment, which is the other extreme. The ARENA approach in [7] only stores every training image and compares it to the task image. As far as we can determine, exemplars are not used in contemporary approaches that use models, and vice versa. This is due to the fact that these two strategies are not necessarily incompatible. A recent method of mixing models and exemplars for face recognition was put forth by [8]. In which models are employed to create new training photos, which may then be utilised as examples during a face recognition system's learning phase.

Face recognition methods can be separated into two groups based on how they handle posture invariance: I global techniques and (ii) component-based approaches. With a global method, a classifier is fed a single feature vector that represents the entire face image. In the literature, a number of classifiers have been suggested, including Fisher's discriminant analysis [11], neural networks [12], and minimal distance classification in the eigenspace [9, 10].

A significant research issue spanning many professions and disciplines is face recognition. This is due to the fact that face recognition is a fundamental human behaviour on the part that is necessary for successful human interactions and communication, in addition to having numerous practical applications like bankcard identification, access control, searching for mug shots, video surveillance, and surveillance systems.

The first formal system for classifying faces was put forward in [1]. The authors suggest assembling face profiles as curves, calculating their average, and then classifying other profiles based on how far they differed from the average. This classification produces a vector of independent measures that can be compared to other vectors in a database since it is multi-modal. Face recognition technology has improved to the point where it is already being used in actual settings [2].

Frontal images of faces can be categorised effectively using global approaches. Yet, because global characteristics are so sensitive to the face's translation and rotation, they are not resistant to changes in position. Before classifying the face, an alignment stage can be introduced to eliminate this issue. Calculating correspondence between the two face images is necessary for aligning an input face image with a reference face image. The connection is typically established for a select few significant facial features, such as the corners of the mouth, the nostrils, and the eye centre. The input face image can be warped to a reference face image based on these correspondences. [13] computes an affine transformation to carry out the warping. [14] employs active shape models to match input face with model faces. In [15], the combination of a semi-automatic alignment stage and a classification phase using support vector machines was suggested.

The classification of local facial components is a complement to the global strategy. Component-based recognition's major goal is to account for pose variations in the categorization stage by providing a variable geometrical link between the components. By independently comparing templates of two facial regions, face recognition was carried out in [16]. (eyes, nose and mouth). As the system didn't have a geometric representation of the face, the component arrangement during categorization was unrestricted. In [17], a similar strategy with an additional alignment stage was suggested. A 2D elastic graph was used in [18] to create a face's geometrical model. The wavelet coefficients obtained on the elastic graph's nodes served as the foundation for the recognition.

In [19], a window was placed across the facial picture, and a 2D Hidden Markov Model was fed the DCT coefficients obtained inside the window. In some specialised areas, such stance and lighting variations, face detection and recognition study still has challenges. Even though several approaches have been put up to address these issues and have shown a great deal of promise, obstacles still exist. Because to these factors, automatic face recognition currently performs matching rather poorly when compared to fingerprints and iris match, despite the fact that it can be the only measuring instrument available for an application. Error rates between 2 and 25% are usual. When used in conjunction with other biometric metrics, it is beneficial.

This section provides a summary of the main human face recognition methods that mostly apply to frontal faces, along with the benefits and drawbacks of each methodology. The techniques taken into account include geometrical feature matching, neural networks, dynamic data design, hidden Markov model, and template matching. The facial representations used in the approaches are examined.

One of the methods for facial recognition that has been the most fully researched is Eigenface. It is also referred to as a primary component, eigenpicture, eigenvector, and the Karhunen-Loève expansion. Principal component analysis was employed in references [16] to effectively depict images of faces. They claimed that with a minimal set of weights for every face and a common face image, any detected faces could be roughly rebuilt (eigenpicture). Projecting the facial image onto to the eigenpicture yields the weights that describe each face. Reference [16] employed eigenfaces, a face identification and detection method inspired by Kirby and Sirovich's method.

Either the eigenvalues of the covariance of the collection of face images or the principal component analysis of the distributions of faces, in mathematics, are referred to as eigenfaces. The eigenvectors are arranged in descending order to indicate varying degrees of variance among the faces. A combination of the eigenfaces can accurately depict each face. Another method is to only use the "best" eigenvectors with the highest eigenvalues to make an estimate. The best M eigenfaces create the "face space," which is an M-dimensional space.

The average correct classifications across lighting, direction, and size variables were reported to be 96%, 85%, and 64%, respectively, by the authors. 16 people were represented by 2,500 photos in their database. The results above are influenced by backdrop because the photos contain a significant amount of background space. The authors used a strong correlation between pictures and variations in illumination to explain the resilient system's ability to function under various lighting situations.

The connection between whole-face pictures, as demonstrated by [17], is ineffective for achieving adequate recognition performance. With the eigen faces technique, illumination normalisation [17] is typically essential. If the object is Lambertian, reference [30] suggested a new approach to construct the covariance matrix using three photos, each obtained in a different lighting environment, to compensate for arbitrary illumination effects. The initial stuff on eigen face in Reference [18] was expanded to include eigen features that correlate to facial features like the eyes, mouth, and lips. They employed a modular eigen space made up of the aforementioned characteristics are important (i.e., eigen eyes, eigen nose, and eigen mouth). The usual eigenface technique would be more susceptible to changes in appearance than this one. For the 7,562 images of nearly 300 people in the FERET database, the system had a 95% recognition rate. In summary, eigenface appears to be a quick, easy, and useful technique. Nevertheless, broadly speaking, it does not provide invariance over changes in scale and lighting.

Recent ear and face recognition experiments [19] using the conventional principal component evaluation approach showed that the efficacy of recognition using ear images or facial images is essentially identical, and that using both for multi - modal recognition results in a statistically meaningful performance improvement. For instance, the multimodal biometric has a rank-one identification rate difference of 90.9% compared to 71.6% for the ear and 70.5% for the face for the day variation experiment using the 197-image training sets. There is extensive related work in multimodal biometrics. For instance, [17] and [18] both used voice and face for multimodal biometric identification. Yet, it appears that using the eye and ear together is more pertinent for surveillance purposes.

The use of neural networks is appealing perhaps because the network is not linear. In light of this, the feature extraction stage might be more effective than the linear Karhunen-Loève approaches. A single layer adaptive network called WISARD that has a different network for each stored individual was one of the earliest artificial neural network (ANN) solutions for facial recognition [19].

For successful recognition, the manner a neural network is built is essential. It heavily depends on the programme that is being used. For face recognition, machine learning model [19] and convolutional neural networks [19] have been utilised. [38] is an inter pyramid for face verification.

Self-organizing map (SOM) neural network, convolutional neural network, and local image sampling are all elements of the hybrid neural network suggested in reference [20]. In the interest of discretization and invariance to small changes in the image sample, the SOM quantizes the image samples into a topological space where input that are close to one another in the original space are likewise close to one another in the output space. With complete invariance to rotation, rotation, scaling, and deformation, the convolution operation retrieves progressively larger features from a collection of hierarchical layers. Using 400 photos of 40 people in the ORL database, the studies report 96.2% correct recognition. Despite the training taking up to 4 hours, the classification process takes less than 0.5 seconds.

A decision-based neural network (DBNN modular)'s structure was handed down to the probabilistic decision-based neural net (PDBNN) used in the reference [20]. The face detector, which locates a human face in a crowded image, the eye localizer, which determines the locations of both eyeballs in order to provide useful feature vectors, and the face recognizer are all scenarios where the PDBNN can be applied effectively. A fully connected network topology does not available for PDNN.

Instead, it creates K subnets to split the network. Each subset is used to identify a single database user. The output of each "facial subnet" in the PDNN is the weighted total of the neuron outputs, and its neurons are activated according to the Guassian activation function. In other words, the popular mixture-of-Guassian model is used by the face subnet to estimate the likelihood density. Mixture of Guassian offers a far more complicated and flexible framework for estimating the time likelihood density in the face space than the AWGN scheme does. Each network is trained with its own set of face photos during the initial phase of the PDNN's two-phase learning process.

The subnet variables may be educated by some specific samples from different face classes during the second phase of the learning process, referred to as decision-based learning. Not all of the training samples are used in the decision-based learning strategy's training. Only incorrect patterns are employed. The correct subnet will adjust its settings in order to relocate its decision-region closer to the incorrectly classified sample if the sample was misclassified to that subnet. The distributed computing

theory of the PDBNN-based biometric authentication system is very simple to build on parallel computers, and it has the advantages of both neural nets and statistical techniques.

A.S. Tolba, " A parameter–based combined classifier for invariant face recognition,", the PDBNN face recognizer could identify up to 200 people in a single second and reach a recognition accuracy rate of up to 96%. To the contrary, as the population grows, the computing cost will get more difficult. In general, as the variety of classes (i.e., people) grows, neural network systems run into issues. Additionally, they are unsuitable for a single piece image recognition test since retraining the systems to "optimal" parameter necessitates the usage of several model photos for each individual.

Recently, the use of SVMs in computer vision problems has been proposed. In reference [22], the face identification problem was addressed using SVMs and a binary tree identification technique. SVMs train the discriminating function between each pair of features after the features have been retrieved. After that, the system receives the disjointed test set for recognition. For the purpose of identifying the testing samples, they suggest using a binary tree structure. There were given two sets of experiments. The first experiment uses 400 photos of 40 people from the Cambridge Olivetti Research Lab (ORL) face database. The second uses a bigger dataset with 1079 photos of 137 different people.

The face recognition issue is presented as a difference space problem in [22], which models the differences between two facial images. They formulate face recognition as a two class problem in a different environment. Cases include I facial differences between individuals, and (ii) facial differences between individuals. A similarity metric between faces is created by altering the interpretation of the decision surface and is learned through examples of how different faces differ from one another. Using a challenging set of photos from the FERET database, the SVM-based technique and a principle component analysis (PCA)-based approach are contrasted.

A component-based technique, two global techniques, and their performance in terms of resilience against pose changes were reported in Reference [23]. The component-based method identified, extracted, and sorted a team of 10 facial characteristics into a single feature vector, which was then classified by linear SVMs. The entire face is identified, taken from the image, and sent into the classifiers in both global systems. For each database user, a single SVM made up the first global system. In the second approach, a group of view-specific, SVM are trained after clustering each person's database.

By incorporating a 3D morphable model into the training process, Reference [23] demonstrated a novel advancement in component-based face recognition. They calculated the 3D face model of each individual in the database using two face pictures of a person and a 3D morphable model. A huge number of synthetic face photos are created by rendering the 3D models in various positions and lighting setups in order to train the component based recognition algorithm. For faces rotated up to 360 degrees in depth, a component-based identification rate of 98% is attained. The system's requirement for a sizable number of training photographs shot from various angles and in various lighting situations was a significant flaw.

An array of K optimal pairwise coupling classifiers (O-PWC) is created in [24] to address the multi-class classification problem for a K-class classification test. Each O-PWC is the most accurate and optimal for the corresponding class in terms of cross entropy of square error. The ultimate choice will be made by merging the K O-PWC findings. This technique is used on the 400 photos of 40 people in the ORL face collection, which contains quite a bit of variation in expression, position, and facial characteristics. 200

samples were included in the practise set (5 for each individual). The test set is comprised of the remaining 200 samples.

Nonetheless, SVMs were investigated as part of facial authentication by [25]. (verification). Their research supports the idea that the SVM approach's effectiveness over benchmark approaches is mostly due to its ability to extract the necessary discriminatory information from of the training data. SVMs lose their superiority when the representation space already captures and accentuates the discriminatory information content, as in the case of Fisherfaces. SVMs can handle changes in illumination as long as the training data appropriately accounts for them. SVMs, however, can get over-trained on data that has been cleaned up using extraction of features (Fisherfaces) and/or normalisation, which impairs their capacity to generalise.

Conclusion

Computer vision substantial research focuses recognition of hand gestures and facial detection in applications like sign language interpretation and human-computer interaction. The development of a system that can recognise hand movements and face detection to send information for controlling media units is the main objective of the proposed methodology. By using hand and facial gestures, sign language is a frequent, efficient, and alternate method of interaction for the deaf and hard of hearing. Here, there is no need for an intermediate medium because the hand and face are being used directly as the input to the device for successful communication and gesture identification.

References

[1] B. Takács, “Comparing face images using the modified hausdorff distance,” Pattern Recognition, vol. 31, pp. 1873-1881, 1998.

[2] Y. Gao and K.H. Leung, “Face recognition using line edge map,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 6, June 2002.

[3] V. Blanz and T. Vetter, “Face recognition based on fitting a 3D morphable model,” IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 25, no. 9, September 2003

[4] A. Lanitis, C. Taylor, and T. Cootes, “Automatic interpretation and coding of face images using flexible models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 743 – 756, 2012.

[5] Purdue Univ. Face Database, http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html, 2013.

[6] V.N. Vapnik, “The nature of statistical learning theory,” New York: Springverlag, 2013.

[7] C.J. Lin, “On the convergence of the decomposition method for support vector machines,” IEEE Transactions on Neural Networks,2014.

[8] G. Guo, S.Z. Li, and K. Chan, “Face recognition by support vector machines,” In proc. IEEE International Conference on Automatic Face and Gesture Recognition, pp. 196-201, 2014.

[9] P.J. Phillips, “Support vector machines applied to face recognition,” Processing system 11, 1999.

[10] B. Heisele, P. Ho, and T. Poggio, “Face recognition with support vector machines: Global versus component-based approach,” in International Conference on Computer Vision (ICCV'01), 2015.

[11] J. Huang, V. Blanz, and B. Heisele, “Face recognition using Component-Based support vector machine Classification and Morphable models,” LNCS 2388, pp. 334-341, 2016.

[12]K.Jonsson, J. Mates, J. Kittler and Y.P. Li, “Learning support vectors for face verification and recognition,” Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, pp. 208-213, Los Alamitos, USA, March 2017.

[13]G.D. Guo, H.J. Zhang, S.Z. Li. "Pairwise face recognition". In Proceedings of 8th IEEE International Conference on Computer Vision. Vancouver, Canada. July 9-12, 2017.

[14]O. Deniz, M. Castrillon, M. Hernandez, “Face recognition using independent component analysis and support vector machines,” Pattern Recognition Letters, vol. 24, pp. 2153-2157, 2018.

[15]Y. Li, S. Gong and H. Liddell. Support vector regression and classification based multi-view face detection and recognition. In Proc. IEEE International Conference on Face and Gesture Recognition, Grenoble, France, March 2018.

[16]K. I. Kim, K. Jung, and J. Kim, “Face recognition using support vector machines with local correlation kernels,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 16 no. 1, pp. 97- 111, 2018.

[17]K. Jonsson, J. Kittler, Y. P. Li, and J. Matas, “Support vector machines for face authentication,” in T. Pridmore and D. Elliman, editors, British Machine Vision Conference, pp. 543–553, 2019.

[18]Huang J., X. Shao, and H. Wechsler, "Face pose discrimination using support vector machines,” 14th International Conference on Pattern Recognition, (ICPR), Brisbane, Queensland, Aus, 2019.

[19]S. Pang, D. Kim, S.Y. Bang, “Membership authentication in the dynamic group by face classification using SVM ensemble,” Pattern

[20]Recognition Letters vol. 24, pp. 215-225, 2020.

[21]A.S. Tolba, " A parameter–based combined classifier for invariant face recognition," Cybernetics and Systems, vol. 31, pp. 289-302, 2020.

[22]A.S. Tolba, and A.N. Abu-Rezq, "Combined classifiers for invariant face recognition," Pattern Anal. Appl. Vol. 3, no. 4, pp. 289-302, 2020.

[23] Ho-Man Tang, Michael Lyu, and Irwin King, "Face recognition committee machine," In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), pp. 837- 840, April 6-10, 2003.

[24] Rui Huang, Vladimir Pavlovic, and Dimitris N. Metaxas, "A hybrid face recognition method using Markov random fields," ICPR (3) , pp. 157-160, 2020.

[25]A.S. Tolba, A.H. El-Baz, and A.A. El-Harby, "A robust boosted parameter- based combined classifier for pattern recognition,"submitted for publication.2021