HUMAN FACE RECOGNITION IN VIDEO USING CONVOLUTIONAL NEURAL NETWORK (CNN)

Keerthi G S[1], Usha C R[2]

[1]*Student, Department of Computer Science and Engineering, BNMIT, Bangalore, India*

[2]*Assistant Professor, Department of Computer Science and Engineering, BNMIT*

ABSTRACT*: Video-based Face Recognition (VFR) is ability of system to recognize one or more people in video based on their facial characteristics. VFR system faces challenges such as pose variation and illuminance condition. Convolutional Neural Network (CNN) technique is one of state-of-the-art machine learning approaches that improves performance and cost effective. The main aim is to detect human faces in video with vary in poses and illuminance condition. CNN techniques improve efficiency in detecting human faces in video and achieves high accuracy on detection of faces.*

Keywords: Video-based Face Recognition, Ytcelebrity dataset, Illuminance, Varying poses, BRISK features, Recognition, RGB2GRAY, Viola-jones Algorithm, Convolutional Neural Network (CNN).

1. INTRODUCTION

VFR technology is a computer vision technology where it collects important information's on the video sequences in a frame. The VFR system as established its are of interest in the field of pattern recognition and computer vision because of its application on various field such as RFID (Radio Frequency Identification) cards, smart cards, surveillance systems, pay systems and access control. The technology can be used in surveillance video, voter identification, pay system, security systems criminal identification and so on [2]. For developing a useful and enhancing application of face recognition some of the factors need to be taken care, that is the system from detecting faces to recognizing face should be acceptable. The face recognition should be easily able to update and detect a greater number of people with high accuracy rate, it helps in enhancing the capacity of the system. The viola-jones algorithm is used in detection of faces and non-faces in video. It gives high accuracy and fast detection rate when compared with other methods such as Eigenfaces, Linear Binary Patterns (LBP) and so on. The viola-jones algorithm reduces computation rate and famous in detecting face with very low false positive [7]. Face Recognition is always a challenging task and its applications are useful for personal verification and recognition. Recognizing a face been very difficult due to all different situation that a human face can be found. The main objectives of video-based face recognition are detection of human faces in the video and identify the persons face accurately. Secondly, focusing on different pose while detecting human faces from video. Final objective is to focus on varying illumination conditions or lighting conditions while detecting human faces.

2. RELATED WORK

Recognition of human face the main challenge faced is varying illumination conditions and pose variation. The dataset Yale Facial Image helps in detecting human faces in varying illumination conditions and pose variations. Yale dataset improves performance of system by considering horizontal reflections of facial images. Convolutional Neural Network (CNN) classifier can learn local features from input data for discriminating facial images [1]. Recognizing faces in surveillance videos is difficult due to poor quality that is in terms of resolution, noise, varying lighting conditions and blurriness. The ChokePoint database has better image quality for face recognition. The benefit of using ChokePoint database is instead of using single cameras, multiple cameras are been used to get image quality. Face recognition uses a Unified Face Image (UFI) that generates several consecutive frames from each camera. Even though probe sequence takes from multiple cameras only few UFI's are extracted. Face Detection from surveillance cameras though UFI generated by fusing images from different cameras [2]. The VFR system face recognition involves matching two or more image sets which contains some of the facial images that are obtained from each video in the dataset. The captured images are taken from datasets such as Honda, MoBo (Mobile Body) and YouTube. The limitation is it performs better on small scale video, hence need to work on performance of large-scale video sequence [3]. VFR faces challenges like variation in pose and occlusion and its main aim is to recognize the face from video-based face patches. The dataset is Yaleface database that are taken frame by frame from video. The first step is to crop face patches from video frame by frame and then to extract the face portion from video frames using an alignment algorithm. The alignment algorithm align it and it also normalize the face image, but here the patches are not identified. A full-face proper precision image stitching algorithm is developed, and it reduces overlapping errors between the patches. The patch-based method is flexible, its computational cost is also vulnerable, but the large changes in pose variation, illumination and expression cannot be detected and noise need to be removed to provide the high-quality image [4]. The challenge is to extract dynamic features for VFR for analyzing facial expressions. The database used is Cohn-kanade (CK+) database. The Local Binary Patterns from three orthogonal planes (LBP-TOP) is used to convolve with multiscale and multi orientation Gabor filters. The database used is ChokePoint database. Instead of using single cameras, multiple cameras can be used to get high image quality. It only focusses on matching the images but further it focuses on pose variation, occlusion and on quality of image [5]. The main challenge in VFR system is improving the performance on computation and accuracy of recognizing the human face with multiple image of a single person. The detection and localization of face on each image is carried out by using a cascade object classifier. The image localization is carried out by fixing its center then cropping only the facial region, later it is resized to 64X64 pixels. In face selection, after testing the quality of test images checked in prior which enables to select the high-quality face image and it is done through face recognition algorithm due to these approaches there is a huge increase in recognition accuracy, it reduces the computational efficiency. The limitation is that it needs to concentrate on performance of the human face recognition from video [6]. Video-based Face Recognition system suffers severe performance degradation under uncontrolled real-world conditions. There are two datasets to determine the performance on detection of faces in video namely Labelled Faces in the Wild (LFW) and Honda/UCSD dataset. The Support Vector Machine (SVM) classifier increase the quality of image, determines the face more accurately. The human faces with extreme pose variation as large pitch angles so facial image information is lost by using homograph matrix. Hence there is a need to concentrate on pose, illumination and variation of facial image from video sequence [8]. In automatic facial expression recognition, the main goal is to classify each facial image as one of seven facial expressions. Extraction of features from the database CK+ and JAFFE helps in recognizing facial expressions on face. The CNN network has convolution layers, pooling layers and fully connection layers. The most benefit by using CNN is obtained by detecting the human faces from the image and video files for achieving better accuracy rate [9].

3. EXISTING SYSTEM

The VFR system involves detecting the human face in video for various purposes. The detection of faces is mainly carried on images or frames but lacks in detecting the person face in video. The VFR defined previously lacks in their performance that it indulges in performance degradation of a system. The person in images are video frame may vary in pose, but even though the face of the right person needs to be detected. The image or video frame may be affected by varying lighting or illuminance conditions, still the faces should be detected. These are the some of disadvantages of existing system in detecting human face in video.

4. METHODOLOGY

Detection of human faces in video-based face recognition involves following image processing stages as shown in figure 1.

4.1 Pre-processing

The video is uploaded, then sequence of frames is generated from the video. These frames are off different illuminance conditions, hence to detect the face with this background is difficult. So, the RGB frames generated are converted into Gray scale images for better detection purposes.

4.2 Face Detection

Here the pre-processed frames are taken for detection of human face. The face is tracked in each frame of the video using Viola-jones algorithm, it enables in differentiating between faces and non-faces of humans in video. The viola-jones method was selected because of its robustness in detection of faces and these faces detected by algorithm is embedded within a rectangular boundary box around face, by doing so there can be a clear few of only human faces selected in video.

Viola-jones Algorithm Steps

Step 1: Haar Feature Extraction

All human faces share some similar properties. These regularities may be matched using Haar Features. A few properties common to human faces are the eye region is darker than the upper-cheeks. Secondly, the nose bridge region is brighter than the eyes.

Step 2: Creating an Integral Image

An image representation called the integral image evaluates rectangular features in constant time, which gives them a considerable speed advantage over more sophisticated alternative features. Because each feature's rectangular area is always adjacent to at least one other rectangle, it follows that any two-rectangle feature can be computed in six array references, any three-rectangle feature in eight, and any four-rectangle feature in nine.

Step 3: AdaBoost Algorithm

Features evaluated does not compensate for their numbers, for example in standard 24X24 pixel sub-window there are total of 1,60,000 features possibly extracted and it is expensive to evaluate all by testing an image. For any classifier with accuracy higher than 50%, the weight is positive. The more accurate the classifier, the larger the weight. While for the classifier with less than 50% accuracy, the weight is negative. It means that we combine its prediction by flipping the sign. For example, we can turn a classifier with 40% accuracy into 60% accuracy by flipping the sign of the prediction. Thus, even the classifier performs worse than random guessing, it still contributes to the final prediction. We only don’t want any classifier with exact 50% accuracy, which doesn’t add any information and thus contributes nothing to the final prediction.

Step 4: Cascade Classifier

On average only 0.01% of all sub-windows are positive (faces) Equal computation time is spent on all sub-windows. Must spend most time only on potentially positive sub-windows. A simple 2-feature classifier can achieve almost 100% detection rate with 50% FP rate. That classifier can act as a 1st layer of a series to filter out most negative windows2nd layer with 10 features can tackle “harder” negative-windows which survived the 1st layer, and so on. A cascade of gradually more complex classifiers achieves even better detection rates. The evaluation of the strong classifiers generated by the learning process can be done quickly, but it isn’t fast enough to run in real-time. For this reason, the strong classifiers are arranged in a cascade in order of complexity, where each successive classifier is trained only on those selected samples which pass through the preceding classifiers. If at any stage in the cascade a classifier rejects the sub-window under inspection, no further processing is performed and continue searching the next sub-window. The cascade therefore has the form of a degenerate tree. In the case of faces, the first classifier in the cascade called the attentional operator uses only two features to achieve a false negative rate of approximately 0% and a false positive rate of 40%. The effect of this single classifier is to reduce by roughly half the number of times the entire cascade is evaluated.

4.3 Feature Extraction

The selection of video for detecting human faces has few features which are extracted from each frame of video using BRISK (Binary Robust Invariant Scalable Keypoints) feature extraction method as shown in Figure 3. This method involves three main steps namely keypoint detection, keypoint description and matching. The BRISK feature extraction method concentrates on extracting the unique keypoints from images or video frames while detecting human faces in VFR. The extraction of features is provided by pyramid layers which consists of n octaves and intra-octaves layers where the intra-octaves are located between the octaves. The salient keypoints obtained keypoint detection involves DAISY method for positioning of the unique sampling patterns. Finally, the most common features between two or more keypoints are calculated using the hamming distance method which measures the distances between the pair of binary string points.

BRISK Feature Extraction Algorithm

Step 1: Keypoints Detection

It is performed at the first moment. The goal at this step is to identify salient points in the image that,

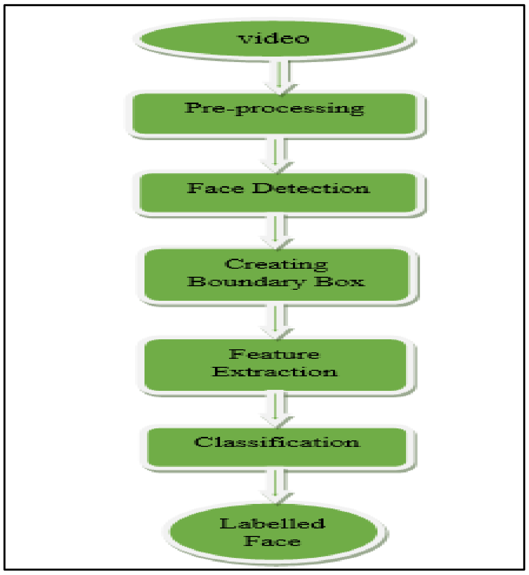


Figure 1: System Architecture of Video-based Face Recognition

ideally, could be uniquely differentiated from any other point. To do so, these points must be searched across the image and scale dimensions using a saliency criterion. The search through scale space is fundamental to achieve scale invariance and it is realized in BRISK using a pyramid of images. This pyramid is formed by many layers which correspond to resamples of the original image. In BRISK method, these layers are divided in n octaves ci and n intra-octaves di, with i = {0, 1,....,n} and usually n = 4. The intra-octaves di are located between octaves ci and ci+1. These intra-octaves and octaves are formed by the original image subsampled by factors 1/ 2 (half-sampling) and 2/3rd (two-third-sampling). The original image is progressively half-sampled to form the octaves, while the intra-octaves are formed by progressives half-sampling of the first intra-octave d0, which in turn is composed by a two-third-sampling of the original image. After the octaves and intra-octaves calculation, the corner detector is applied to each one of them. All the points of each layer will be evaluated as a keypoint candidate, and to be elected a point must be salient among their intra-layer and inter-layers neighbors. It means that each point is compared with its neighbors in the same layer and with those in the above and below layers. After the analysis of all octaves and intra-octaves, the detection stages finally end, producing a set of keypoints with space and scale coordinates, what means that each one of them can be exactly located.

Step 2: Keypoints Description

In this stage, around each keypoint, which has its coordinates found in the previous step, a sampling pattern is positioned. The BRISK algorithm uses a sampling pattern like the one in DAISY descriptor. However, it is important to note that the use of this pattern is quite different. The pattern relies on 60 equally spaced points located in four concentric rings. In order to produce a scale and rotation invariant description, the sampling pattern is exactly scaled and rotated according to each keypoint. The BRISK descriptor is composed as a binary string of length 512. This string concatenates the results of simple brightness comparisons tests between each keypoint and its 60 neighbors in the pattern. This approach was inspired by the BRIEF method.

Step 3: Matching

Finally, to perform the comparison between two or more keypoint descriptors, the Hamming distance is used. This distance measures the number of different bits between two binary strings and it represents the degree of inequality of the descriptors being compared.

4.4 Classification

The final step in detection of human faces in video-based face recognition system is classification using Convolutional Neural Network (CNN) technique. Here its main aim is classifying the person according to alphabets which further helps in labelling a person face detected. Then is to specify the training option and set a training CNN network. In final step the prediction of faces by labeling the new data and enhancing the accuracy of classification through VFR.

CNN Classifier Model

Step 1: Load and explore image data.
Step 2: Define the network architecture.
Step 3: Specify training options.
Step 4: Train the network.
Step 5: Predict the labels of new data and calculate the classification accuracy.

5. RESULTS

Datasets: The public dataset used is ytcelebrity which consists of about 200+ videos for detection of face in video. The dataset consists of more subjects (at least 10 videos for each subject) where the video sequence is composed of varying poses and illuminance conditions. The ytcelebrity dataset consists of face movement of a person in video. It has single person and multi-person in video where there is a need to identify trained person in a video. For more complexity for detecting faces in video-based recognition system one can download a video with a greater number of people surrounded around the trained faces and then identify the right person in video, by doing so more complexity is obtained.

Comparison with existing system: In existing methods processing lead to a smaller number of video frame, it lacks in performance on large datasets when compared with small size dataset [5]. The Unified Face Image (UFI) technique [14] only focuses on matching images but it needs to focus on pose, occlusion and on quality of image. SVM classifier [11] trains for long time when training large datasets and it is not robust in understanding the finalized models and variable weights.

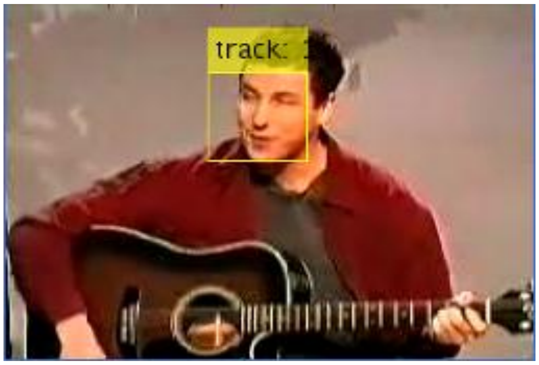


Figure 2: Tracking of Faces

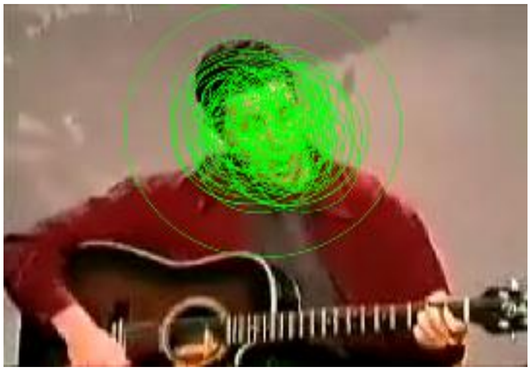


Figure 3: BRISK Feature Extraction

Analysis: A video is loaded to detect the faces in video, hence it is preprocessed from converting RGB to Gray scale for eliminating the lighting effects in video. A cascade face detector is applied to detect faces in each video sequence as shown in figure 2. For comparison purpose the videos are been taken for training and testing, so that it can recognize the person correctly. Figure 4 shows the labelled face detection in video-based face recognition system. Feature keypoints are extracted using BRISK feature extraction method. It identifies salient points in the image that is uniquely differentiated from any other point. The BRISK features are obtained by more number layers that are divided into n octaves and intra-octaves and finally obtains some keypoints by using corner detector.

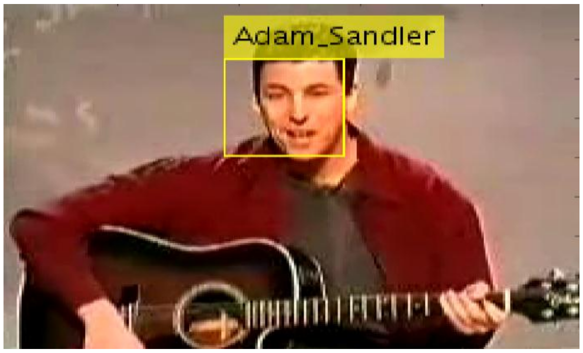
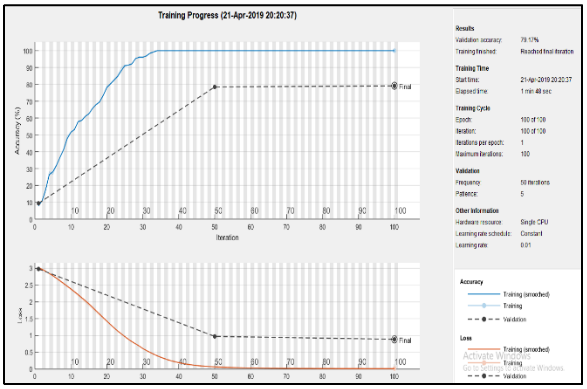


Figure 4: Labelled Face detection



Results
Validation accuracy: 79.17%
Training finished: Reached final iteration

Training Time
Start time: 21-Apr-2019 20:20:37
Elapsed time: 1 min 48 sec

Training Cycle
Epoch: 100 of 100
Iteration: 100 of 100
Iterations per epoch: 1
Maximum iterations: 100

Validation
Frequency: 50 iterations
Patience: 5

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.01

Figure 5: Performance Analysis

In figure 5 it represents each step of performance analysis along with data divided by training step as well as validation step data sets, there are 750 images or above in every category, the validation dataset will have left images for labelling. The validation data and training digit data are split into newer datastores. The training can either be stopped by clicking on the stop button present on the top-right corner, while training a network it can return to its current state completely without affecting the training data. For example, during training there may be a need to stop the training when the accuracy as reaches its high value and is further the accuracy is not increasing, then can click on a stop button to let the training complete and once the training gets completed the training network returns for training. When the training of network is completed then can view the results that accomplishes validation accuracy and the reason for completion of the training. The final validation metrics are been labelled in the final plots, if networks works on batch normalized layers then these networks are completely different from the validation metrics that has been evaluated while training a network. Figure 6 shows multiple face recognition in video, it recognizes the face of various person, but the accuracy level is degraded when compared with detection of single person in video.

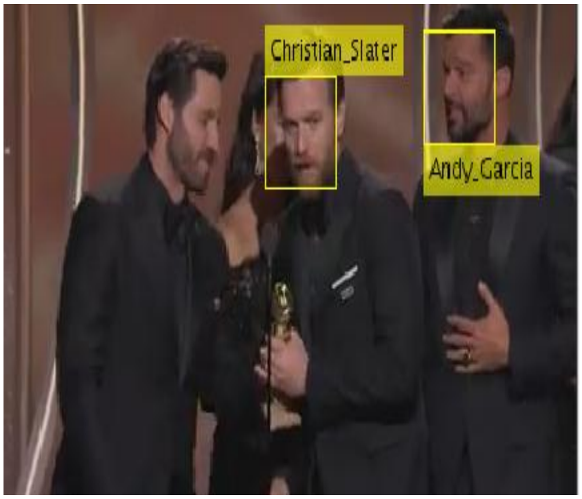


Figure 6: Multiple Face Detection

6. TEST CASES

The test cases are defined using every module and sub-modules involved in developing VFR system. The table 1 describes all modules description and their possible outcomes in detecting human faces in video. The test cases help in gaining information on each step of detecting face in video, it checks whether they are successful in each step or not.

Test Id	Test Case Description	Input	Output	Test Case Status
1	Frame Extraction	Video	Number of frames extracted	Pass
2	Pre-processing	RGB Frames	Gray Scale frames	Pass
3	Face Detection	Processed Image	Faces and non-faces detected	Pass
4	Feature Extraction	Detected Faces	Features extracted	Pass
5	Classification	Classified Person Face	Labelled face	Pass
6	Classification	Classified Person Face	Labelled face	Fail

Table 1: Test Cases

7. CONCLUSIONS

A detection of human faces in VFR task is achieved. The tracking of faces in each frame of a video is accomplished using Viola-jones algorithm, hence faces and faces in each frame are distinguished properly. The Brisk feature extraction method extracts interested keypoints in faces which helps in matching persons face in video. CNN classifier achieves better recognition performance, memory management and it's easy to use on detecting faces in video. Video-based face recognition system help in tracking and recognizing person face with different poses and illuminance condition, that leads to improve performance of recognition.

In future the training of a person with more in number and detecting more people in video can be done along with more accuracy rate.

REFERENCES

[1] N. Pattabhi Ramaiah, Earnest Paul Ijjina and C. Krishna Mohan, "Illumination Invariant Face Recognition Using Convolutional Neural Networks", IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), 2015.

[2] Le An, Bir Bhanu and Songfan Yang, "Face Recognition in Multi-Camera Surveillance Videos", International Conference on Pattern Recognition, 2012.

[3] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen, "Image sets alignment for video-based face recognition," in Proceeding IEEE Conference Computer Visual Pattern Recognition, 2012.

[4] Changbo Hu, Josh Harguess and J. K. Aggarwal, "PATCH-BASED FACE RECOGNITION FROM VIDEO", 2009.

[5] XIE Liping, WEI Haikun, YANG Wankou and ZHANG Kanjian, "Video-based Facial Expression Recognition Using Histogram Sequence of Local Gabor Binary Patterns from Three Orthogonal Planes", IEEE Proceedings, 2014.

[6] Vignesh S, Manasa Priya K. V. S. N. L., Sumohana S. Channappayya, "Face Image Quality Assessment for Face Selection in Surveillance Video using Convolutional Neural Networks", IEEE Global Conference on Signal and Information Processing, 2015.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceeding IEEE Conference Computer Visual Pattern Recognition, 2015.

[8] Walter J. Scheirer, Patrick J. Flynn, Changxing Ding, et al., "Report on the BTAS 2016 Video Person Recognition Evaluation", IEEE International Conference on Biometrics: Theory, Applications and Systems, 2016.

[9] Cong Wang, "A Learning-based Human Facial Image Quality Evaluation Method in Video-based Face Recognition Systems", IEEE International Conference on Computer and Communications, 2017.

[10] Xiaoguang Chen, Xuan Yang, Maosen Wang and Jiancheng Zou, "Convolution Neural Network for Automatic Facial Expression Recognition", IEEE International Conference on Applied System Innovation, 2017.

[11] Marko Arsenovic, Srdjan Sladojevic, Andras Anderla and Darko Stefanovic, "FaceTime – Deep Learning Based Face Recognition Attendance System", IEEE International Symposium on Intelligent Systems and Informatics, 2017.

[12] Changxing Ding and Dacheng Tao, "Trunk-Branch Ensemble Convolutional Neural Networks for Video-based Face Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[13] Maksym Kovalchuk, Anatoliy Sachenko, Vasyl Koval and Diana Zahorodnia, "Development of Real-time Face Recognition System Using Local Binary Patterns", IEEE Second International Conference on Data Mining &Processsing, 2018.

[14] Wen Yang, Xiaoqi Li and Bin Zhang, "Heart Rate Estimation from Facial Videos Based on Convolutional Neural Network", IEEE Proceedings, 2018.

[15] Savitha G and Keerthi G S, "Video based Face Recognition Using Image Processing Techniques", International Research Journal of Engineering and Technology, 2019.