Research Article

Face Recognition from Video using Threshold based Clustering

R. P. Dahake[1], M. U. Kharat[2]

[1]MET's Institute of Engineering, Research Center, SPPU Nasik, India
[2]MET's Institute of Engineering, Research Center, SPPU Nasik, India

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Video processing has gained significant attention due to the rapid growth in video feed collected from a variety of domains. Face recognition and summary generation is gaining attention in the branch of video data processing. The recognition includes face identification from video frames and face authentication. The face authentication is nothing but labelling the faces. Face recognition strategies used in image processing techniques cannot be directly applied to video processing due to bulk data. The video processing techniques face multiple problems such as pose variation, expression variation, illumination variation, camera angles, etc. A lot of research work is done for face authentication in terms of accuracy and efficiency improvement. The second important aspect is the video summarization. Very few works have been done on the video summarization due to its complexity, computational overhead, and lack of appropriate training data. In some of the existing work analysing celebrity video for finding association in name node or face node of video dataset using graphical representation need script or dynamic caption details As well as there can be multiple faces of same person per frame so using K- Means clustering further for recognition purpose needs cluster count initially considering total person in the video. The proposed system works on video face recognition and summary generation. The system automatically identifies the front and profile faces of users. The similar faces are grouped together using threshold based a fixed-width clustering which is one of the novel approach in face recognition process best of our knowledge and only top k faces are used for authentication. This improves system efficiency. After face authentication, the occurrence count of each user is extracted and a visual co-occurrence graph is generated as a video summarization. The system is tested on the video dataset of multi persons occurring in different videos. Total 20 videos are consider for training and testing containing multiple person in one frame. To evaluate the accuracy of recognition. 80% of faces are correctly identified and authenticated from the video.

Keywords: face recognition, fixed-width clustering, co-occurrence graph, video summarization

1. Introduction

There is rapid growth in video feed collected from a variety of domains like a collection of movies, documentaries, series, etc from Over The Top platform, the surveillance camera video collection, users' personal video collection, etc. In the domain of data mining, unknown information and interesting patterns are extracted. The aim of video data mining is to extract important information, relationship among data, and interesting patterns from video content. As compared to the other data mining domain, video mining is still in the stage of infancy. The video is treated as a collection of still images called as frames, arranged in a sequence with the help of temporal information. The information form video can be extracted by comparing:

1. Two videos: The two videos are input to the system. The comparison between these two videos can be performed based on the frames data and similarity can be extracted by matching objects or persons' faces.

2. Video to still images: The video and still image set is input to the system. In the video processing, the still frames are initially extracted and processing steps are applied on these frames. Video processing is an extended version of image processing. The system extracts frames from the video and compares the frames with the still image set.

In information extraction, generally, object recognition and person recognition plays an important role. The recognition terms refer to identification and verification. In face recognition, the system initially scans the frame data and identifies the facial shapes from videos. Then the authentication i.e. verification is performed to validate the identity of each user present in the video.

Along with the person recognition more statistical information can be collected from the video like, leading character those appears on the screen for a longer time span or appearance count for each person, Co-occurrence information of persons present on a screen, the co-occurrence persistence, etc. This generated statistical information helps in deep learning process. Face recognition, especially in video processing faces two major challenges: accuracy and efficiency. The accuracy refers to the correctness of face identification and verification. The following factors affect the accuracy of face recognition:

A. Illumination variation: The slight variance in video affects the face identification. The face appears different with respect to different relative positions with light and with different light intensities. In low-light conditions, it is difficult to find and capture facial imagery.

B. Motion blur: In advance camera setting or in photo editing, focusing variation can be done. The camera exposure time can be set to long or background blur may obscure the face present in the video frame.

C. Occlusion: In the video, some parts of the face may be blocked by other objects present in the video. This creates obstacles in the face identification process.

D. Pose variation: The CC camera captures the environment movements with uncontrolled conditions. This may capture the non-ideal face shots. The variation of pose may lead to the problem in face authentication.

E. Expression variation: The facial expression changes the appearance of the face in the video. This leads to problems is face authentication.

F. Scale Variation: The scale variation is represented by the percentage of the screen occupied by the face. The nearer face occupies the largest portion of the frame. As a person goes away from the camera, the occupation section becomes smaller. After some distance the face becomes unrecognizable.

The efficiency of the system depends upon the video size and number of faces present in the actual video and the training data. The number of frames generated from video depends upon the video length and frame rate per second. The processing time get increases as the number of frames increases in the video. For face authentication, the facial data is matched with previously known faces provided as training the system. The authentication process time depends upon the training dataset size.

The proposed system works on face recognition form video by comparing the video data with a set of still images of video sequences. The system extracts statistical information from the video. The system initially extracts the frames from the video. From the extracted frames the faces are extracted. The system finds profile and front faces from the video. Then from the faces clusters are created using fixed-width clustering. Similar faces are grouped together. This helps to reduce the time of comparison. From the whole cluster, only top k images are matched with the training dataset. This reduces the time required for the recognition process. At the time of the matching system can use co-occurrence information. At the time of recognition, the face is matched with every training set face. To improve the efficiency of the system the system start matching with high probability occurrence faces with video data and after matching the face, the co-occurrence information is useful for finding the next probable face. After the face recognition system generates/updates the co-occurrence information graph and finds the occurrence frequency of each face in the video. The following section includes related work done in section II followed by proposed system in section III. Implementation details are mentioned in section IV and section V provides result and discussion followed by the conclusion in section VI.

2. Literature Survey

The recognition of faces from video frames is an image processing part. There are multiple face recognition strategies (Li B, Liu J, 2011) like PCA, Locally Linear Embedding, Isomap, etc. In video-based face recognition, important aspects are frame selections, recognition strategies, information extraction from videos, and efficiency improvement strategies.

1. Frame Selection: The faces from each video frame are extracted. Matching all the faces with training data for authentication is a time-consuming task. For efficient execution of face matching, there should be some strategy to select some key frames and/or key faces for authentication purposes. manifold- manifold distance framework was proposed by (Wang R., S. Shan, X. Chen, Q. Dai, and W. Gao, 2012) This technique extracts faces from video frames and generates subspaces called manifold based on pose and illumination variation. The variation is captured using local linearity property. The distance between the two manifolds is calculated by measuring the similarity between the input and reference subspaces/manifolds. However, the performance of a subspace/manifold based approach depends on maintaining the image set correspondences Kayal S. 2013) proposed a clustering approach. It uses GMM-based hierarchical clustering for face recognition. Before the face authentication process, the system generates clusters of faces using spatial features of faces and the temporal information. The final number of clusters are generated using ground-truth value computation of the temporal hierarchical agglomerative clustering algorithm.

2. Recognition strategies: (Barr J. R., K. W. Bowyer, P. J. Flynn, and S. Biswas, 2012) categorized the video face recognition strategies into 3 different types:

A: Set-Based Approach: In this approach, the video is treated as a set of images. The training image dataset can be retrieved from a static image set or can be derived from a reference video. The system compares the two image datasets. (Cui Z., S. Shan, H. Zhang, S. Lao, and X. Chen, 2012) proposed an image set alignment method before face matching. The training image dataset is predefine and pre-structured. In accordance with the training dataset, the face images are aligned.

(Harandi M. T., C. Sanderson, S. Shirazi, and B. C. Lovell, 2011) proposed a system that represents images set as a linear structure called as subspaces. This technique treats the image subspaces as a point on Grassmannian manifolds. This technique uses a graph embedding framework. This structure derives the intra-class and between class structure and finds the correlations among subspaces. It finds the similarity between reference i.e training and test image set. For matching the training and testing data, enough training dataset is required with an abundant variation. (Y. Hu Y., A. S. Mian, and R. Owens, 2011) proposed an image set classification technique. The images set is represented as a triplet containing sample image, mean of sample images, and affine hull model. The system uses Sparse Approximate Nearest Points dissimilarity measure. It finds between set dissimilarity. It improves the scarcity of nearest points using the scalable accelerated proximal gradient method.

B: Sequence-Based Approach: In this approach, 2 videos are compared. The training dataset contains a set of single person videos. (Ding S, Ying Li, Junda Zhu, Yuan F. Zheng, Dong Xuan, 2013) proposed a sequential sampling and updating scheme called as SSC. The training and testing datasets are videos. The probability mass function is sequentially updated to find the identity of a person. It also works on pose variations and identity switching during the recognition just because it treats the video of a single person. The system requires multiple videos of single persons with abundant pose variations. (Franco Annalisa, Dario Maio, F. Turroni ,2014) proposed a Spatio-temporal analysis for the video to video comparison. It works on dynamic facial changes with temporal information in running video. It works on the temporal dimension of the analysis of facial expression in terms of key points and uses SURF descriptor.

C. Dictionary-based: Approach: In this approach face images from a test video are matched with a large number of faces in a dictionary. (Chen Y.C., V. M. Patel, P. J. Phillips, and R. Chellappa, 2012) proposed a technique in which a video sequence is partitioned into a number of subsequence and then built the sequence-specific dictionaries. Based on the dictionary face identification and authentication is performed. This approach increases the overhead of the creation of sequence-specific dictionaries for pose variation and illumination variation. (Chen Y.C, V. M. Patel, S. Shekhar, R. Chellappa, and P. J. Phillips, 2013) also proposed an approach for dictionary specific learning. It uses the joint sparsity coefficient approach and concatenates n subdirectories and generates a combined decision. In the training phase, the system filters the different faces, and partitions are created for pose and illumination variation. Each subdirectory defines the face in a particular viewing condition. The same kinds of partitions are created for the test video query and comparison is made according to joint sparse representation. In (Dahake R., M. U. Kharat, Priti Lahane, 2016) authors have discussed challenges in face recognition process from real world scenario. Face recognition is affected by various parameters like:

1: Illumination and color variation: (Chai D. and K. N. Ngan, 1999) proposed a skin color filter. In this system, a universal skin color map is derived. The system detects the pixels with skin color appearance by comparing the chrominance component of each pixel value in an image. (Balasubramaniam, Vivekanandam & Dr. Babu, 2017) worked on illumination variation in the video. The system uses feature extraction techniques such as Local Binary Pattern, Histogram Orientation Gradients Elastic Bunch Graph matching and Weber Local Descriptor. These features are then classified using ELM with K- means classification technique.

2: Pose Issue: In a video capturing the camera angles and human pose variation faces many issues for face identification and recognition process. (Annis Fathima, V Vaidehi, S Vasuhi, Mukund Murali, S K Parulkar, 2015) proposed a pose invariant face detection. It works on frontal and profile faces. It uses a cascaded structure of HAAR-like features along with Adaboost classification. It includes a measurement called as: the degree of freedom. This measurement counts the pose as Roll, Pitch, and Yaw orientations. The HAAR cascade is a machine learning technique where the training data contain a lot of training faces with its orientation.

3: Low Resolution: The low-resolution problem is generally faced by live video capturing during surveillance cameras. (Li, P, Patrick J. Flynn, Loreto Prieto, Domingo Mery, 2019). proposed a survey on face recognition

strategies developed for low-quality images. The technique is known as Low-resolution face recognition. It contains a survey of techniques proposed in the last 6 years.

4: Information extraction: Important concepts or some patterns repetitively occur in multiple videos (Zhang X, F. Pala and B. Bhanu (2017). The video summarization is the concept of summary generation of video by finding shots, objects and/or persons who occur frequently. (Chu W., Yale Song and A. Jaimes, 2015) have proposed an algorithm for Maximal Biclique Finding algorithm for finding sparsely co-occurring patterns and discard the less occurring one. The system discards the dominant pattern in a video if it less occurs in other dataset videos. **(**Haq, I. U. K. Muhammad, A. Ullah and S. W. Baik, 2019) proposed a system to detect main characters from Hollywood movies. The star character occurrence is identified as a video summarization. The system initially extracts the faces clear shots then face clustering is applied on discriminative deep features and then generates the occurrence matrix. Using this occurrence matrix star characters of movies are detected. As there can be multiple faces of same person per frame so they can be cluster together using K- Means clustering further for recognition purpose. This approach needs cluster count initially considering total person in the video **(**Pande N, R. P. Dahake, 2014). In (Tadge S., Ranjana Dahake, 2017) authors have worked on celebrity dataset for face name association from the video and generating the face name graph. Many of the approaches need script, tags information or celebrity video details for finding association in name node or face node of video dataset using graphical representation.(Workie, Ashenafi & Rajendran, Rajesh Sharma & Chung, Yun, 2020) proposed a survey on digital video summarization. In summarization, the important information from video key frames is extracted. The author states that the video summarization is still a changing problem due to its complexity, computational overhead, and lack of appropriate training data. Different clustering techniques for content-based feature extraction from Image are discussed in (Kharat M. U. R. P.Dahake1, Kalpana V. Metre, 2018). In (Surva Remanan, 2020) authors have given deep learning-based video summarization and explored that many advances will be made in the near future to create and optimize the best summaries based on the audience, delivery medium, and intent of summarization. Threshold based clustering for facial image clustering from video is one of the novel approach in face recognition process.

3. Proposed Work

Problem Formulation and analysis: There is exponential growth in video data collected from a variety of media. Analysis and summarization is an important concept. The analysis represents the mining of video data and collecting important information. The extracted information can be summarized in some statistical format to represent a single and/or group of video. Video face recognition and summary generation is important parts of video processing. The image processing techniques cannot be directly applied to video processing. The video processing techniques face multiple problems such as pose variation, expression variation, illumination variation, camera angles, etc. There is a need to develop a system that helps to analyze and generate statistical analysis from the video.

The following figure 1 shows the architecture of the system. The video and training dataset is input to the system. The video can be previously recorded, from the given video system identifies and then authenticate the faces. After authentication system generates statistical information such as the occurrence frequency of each person in the video and the co-occurrence weighted graph of persons in the video. The occurrence frequency and co-occurrence information obtained from previously uploaded videos is input to the face authentication phase. This frequency information and co-occurrence information helps to select the next probable face in training data for matching. This reduces the searching iterations and hence improves the system efficiency. Important steps in system design are given in here.

A. Frame Generation: The stored video or live camera capture is input to the system. Using this framework system automatically finds the video properties like frame rate, frame size, etc. and accordingly creates and saves the video frames at the defined location. The frames are saved with the sequence number with respect to the appearance of the frame in the video.

B. Face Extraction: For face extraction from the frame, the HAAR cascade classifier is used. It is based on the HAAR Wavelet technique. This technique analyses the pixels of images on multiple squares by functions. It detects a number of small important features and then uses a cascading technique to detect the face in an image.

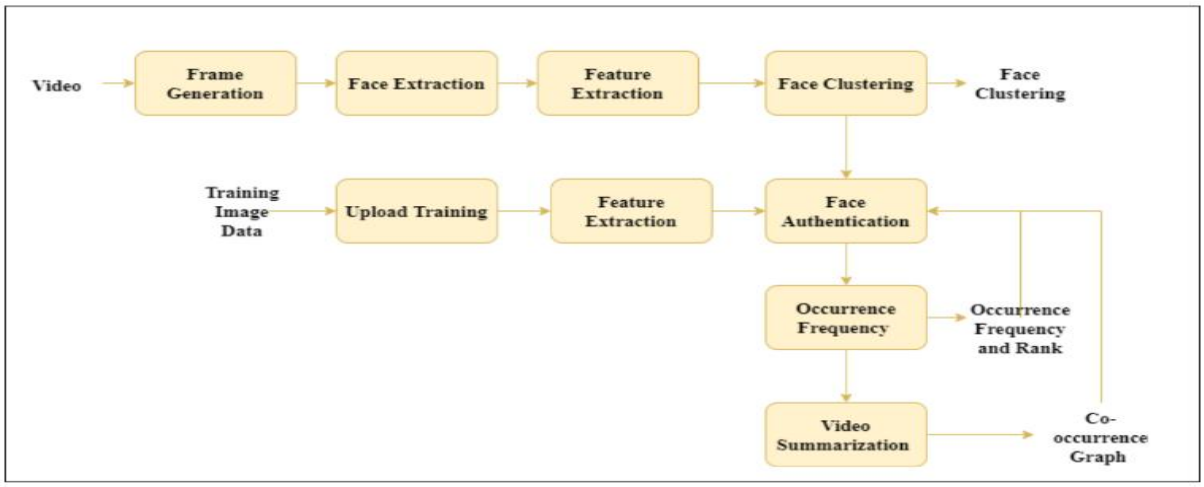

Figure 1. System Architecture for video summarization

This classifier is present in the OpenCV Framework (docs.opencv.org). The system extracts front and profile faces using this technique. After detecting the faces, the faces are cropped from the frames and resized to a fixed-size square of 96X96. The cropped faces are saved at the defined location with the frame number appended with the generated random name.

C. Feature Extraction: For feature extraction Principal Components Analysis PCA is used. PCA is a powerful method for finding similarities and differences among multiple images. The main advantage of PCA is the reduced dimension set generation. PCA reduces redundant information and generates a low dimension subspace without much loss. The features of cropped faces are extracted using PCA and saved in a CSV file. The feature extraction phase is also applicable to the training dataset. The features of each face in the training image dataset are extracted and given to the authentication phase.

D. Face Clustering: The similar faces are grouped together in this phase. Rather than authenticating each face, top k representative faces from the group are used for the authentication process. This reduces the time of authentication. The grouping is done using k means and fixed-width clustering. In the k-mean algorithm number of persons should be previously known and this infeasible condition for real-life videos. Then overestimation of k provides the solution. But overestimation count varies with respect to video. The count k is too high then the cluster may contain very few faces while in other cases where the cluster count is too low then it may generate a group of faces with different persons looking similar to each other. The K-Mean clustering is also unable to handle noisy data and outliers.

To overcome this problem, a fixed-width clustering is used. This provides a better solution and removes the dependency of k. The cluster width is input to the fixed-width clustering algorithm. Based on the fixed-width system automatically generates the number of clusters depending on the variation in data. The width is again a dependent parameter. It is set by analyzing the results of multiple datasets. The system generates multiple clusters based on the width and similarity in faces. Following the algorithm shows the detailed steps of execution for fixed-width clustering algorithm.

Fixed width Clustering

Input: Face image feature vector set(S),
pre-defined radius/width (w)

Output: Set of clusters.
Processing:
1. Initialized: set of clusters C_n, their centroid and initial cluster as (n=0)
2. for face feature vector fe_i in S set do
3. if the number of created cluster is zero then
4. n=n+1
5. put fi in C_n; put fe_i in Centroid
6. else
7. for each centroid C_i in C_n

8. Ed: Compute Euclidean distance between the centroid C_i and fe_i.
8. if Ed< w and n > 0 then
9. fe_i is added into that cluster C_i
10. M_i = Compute mean of (C_i, fe_i)
11. Update Cluster centroid: $C_i = M_i$
10. else
create new cluster C_i+1 and set fe_i as its centroid
11. END

E. Training: In the training phase, the number of images is input to the system with its label. The faces are cropped from images and faces and its label is extracted as a preprocessing step. The features of faces are extracted using PCA. The extracted feature of each face and its label is input to the authentication phase.

F. Face Authentication: In the authentication phase, test video faces data is compared with the training dataset and the label of each face is identified. For face comparison, the features extracted by PCA of test faces are compared with training data faces using Euclidian distance. The KNN-1 technique is applied and the first neighbor of test image from training data is identified and accordingly, a label is assigned to the test image. For efficient system execution, the following two techniques are applied:

1: Cumulative Gain (CG) is the sum of the graded relevance values of all faces in each clusters. The CG at a particular rank position is defined as:

$$CG = \sum_{i=1}^{p} rel_i \quad (1)$$

Where reli is the graded relevance of the cluster face at position i from the centroid. The cumulative gain of each face is extracted from a given cluster. The top m faces are selected using cumulative gain value. The faces near to the centroid have a higher relevance. Only top m faces from each cluster are tested with the training dataset and matching labels are extracted. The cumulative voting result of matched labels is calculated and the maximum occurred label is assigned to all the faces in the cluster. This reduces the time required for matching of each face from the test video.

2. Co-occurrence Information: Generally, a scene in video occurs for a longer or shorter time span at a single location with the same group of people. The adjacent frames in a video generally represent the same group of persons. The co-occurrence information represents the occurrence of two or more persons in the same frame. This information is generated by analyzing the co-occurrence of persons in each frame in a video.

The co-occurrence information helps to find candidate faces from training datasets that may occur in the video. The co-occurrence information helps to reduce iterations of face matching at the time of the face authentication process. Initially, the one face is matched with the whole training dataset and get the matched label. For other faces, the probable matching candidate faces are extracted by checking the co-occurrence frequency of previously matched faces with the new one. The face having the highest co-occurrence frequency has the highest probability of matching. This reduces the time for matching each face with the complete training dataset by generating a candidate list.

G. Video Summarization: The summary is generated in terms of Users' Occurrence Frequency and co-occurrence graph.

Occurrence Frequency: An occurrence frequency is calculated using on the count of faces of each individual person in a video.

According to Zipf's law, the collection frequency cf_i is proportional to the inverse of the rank i.

$$cf_i = \frac{1}{i} \quad (2)$$

Based on Zip's law the rank of each individual is calculated. The higher the occurrence frequency, the lower is the rank.

Generate a co-occurrence Graph: The co-occurrence information of previously analyzed videos is input to the system. After the face authentication system updates the co-occurrence information graph with respect to the

current input video. The co-occurrence graph is generated by analyzing occurrences of persons in each video frame. Following algorithm 2 elaborates on the steps of co-occurrence graph generation. Co-occurrence graph

Generation Finding Co-occurrence between persons

Input: Pl: person list containing faces with label
Output: co [][]: co-occurrence matrix
Processing:

1. N: Find unique person count
2. Initialize co-occurrence matrix co[n][n]={0}
3. Initialize hashmap po = null for the frame-wise occurrence of a person
4. For each person
5. Pl: Generate frame-wise occurrence list of each person
6. po.add (person_label, Pl)
7. For each person i in po
8. Get Pl_i = get list at i^{th} index
9. For each person j in po
10. Get Pl_j = get list at j^{th} index
11. If i==j
12. continue;
13. Find lst = $p_i \cap p_j$
14. Update co[i][j] = size(lst)
15. Update co[j][i] = size(lst)

The complete system is summarized in the following video face recognition VFR algorithm.

VFR Algorithm:
Input: V=Video
W: constant for fixed-width clustering
Tr = Training folder containing faces
folders with face_label as the folder name
Output: res= Similar faces cluster
Updated_ci= Face co-occurrence information

Processing:
/** Frame Generation **/
1. Fset =Divide video V in frames and label the frames with sequence no
/** Face Extraction **/
2. For each frame f_i in Fset
3. Fif = Extract front faces
4. Fip = Extract profile faces
5. Resize to the size 96*96 and Save Fif and Fip with frame no f_i and serial no. in folder faces
6. End For
/** Feature Extraction **/
7. For each face fc in folder faces
8. Fpca = Extract features using principal component analysis (PCA)
9. F_csv = Save feature in CSV with Fpca
10. End For
/** Face Clustering **/
11. Apply fixed width clustering on F_csv data with width W using algorithm 1
/** Upload Training **/
12. Generate Face clustering result Fc containing similar faces in the same folder

```
    /** Feature Extraction **/
13. T = Load Training Dataset and extract feature
      with PCA
   /** Face Authentication **/
14. CI = Load Co-occurrence Information
15. Initialize PI = null
16. For Each folder in Fc
17.         If PI is null
18.             ci = Select top m faces using formula 1 to match
                with the training dataset
19.             Match the ci images with Training set T with
                PCA features
20.             Get match label cumulatively
21.             Pl = Read the co-occurrence information CI
                with the matched label and generate next
                Probable candidate matching info list
22.         Else
23.             Match the ci images with Faces in Training set
                T with the matched label in PI
24.             If match not found
25.                 Match the ci images with Faces in Training
                    set T with PCA features
26.                 Get match label cumulatively
27.                 Read the co-occurrence information CI with
                    matched label
28.                 Generate the next Probable candidate
                    matching information list
29.       In Res folder
30.             Find the folder with the match label
31.             If folder not found
32.                 Create folder
33.             Add images in a folder
34.             Set PI = match label
35. End for
/** find occurrence and co-occurrence information**/
36. Find count of each face and rank using eq2
37. Find co-occurrence information using algorithm 3
38. Return
```

4. Implantation Details

The system is implemented using Java 8 platform and tested on windows 10 operating system.

4.1. Dataset

Short video sequences are captured for experimentation. This Real World Video dataset consists of 10 subjects of age group from 10 to 70 years. Total 20 videos containing multiple person in one frame are consider for training and Testing. The dataset has a frame rate of 60 fps and the image resolution is 1280X720 pixels. Faces in this datasets have variations in terms of expressions, illumination conditions, pose, and sharpness, as well as misalignment, facial occlusion like goggle. Figure 2 shows the sample video frames of the dataset

Figure 2. Illustrative frames from video sequences

4.2. Metrics of Evaluation

Evaluation of the proposed approach with respect to the parameters such as

A. Face Identification accuracy: The faces from frames are cropped. The face identification accuracy is calculated by comparing the actual faces that occurred in the frames and the cropped faces.

B. Cluster analysis:

i. Visual outlier

The visual outliers are detected from data. The outliers found using k-means clustering and fixed width clustering is compared.

ii. Cluster Purity:

The generated faces clusters are verified using the face cluster purity factor. The Cluster purity is calculated as:

$$purity = \frac{1}{N}\sum_{i=1}^{N}\frac{Max_j|C_i \cap Name_j|}{|C_i|} \qquad (3)$$

Where,

C_i is the set of face tracks in the i^{th} cluster and $Name_j$ is the set of face tracks with the j^{th} label (person name)

C. Face Authentication Accuracy: The face authentication process is validated by checking the correctness of the label assigned to the face.

5. Results and Discussions

A. Face Detection: The system focuses on finding the front and profile faces of users. For face detection, the HAAR cascade classifier is used. The following graph shows the accuracy achieved in the face detection process. The figure 3 includes the graph containing result of face tracking and detection for 10 different videos in dataset.

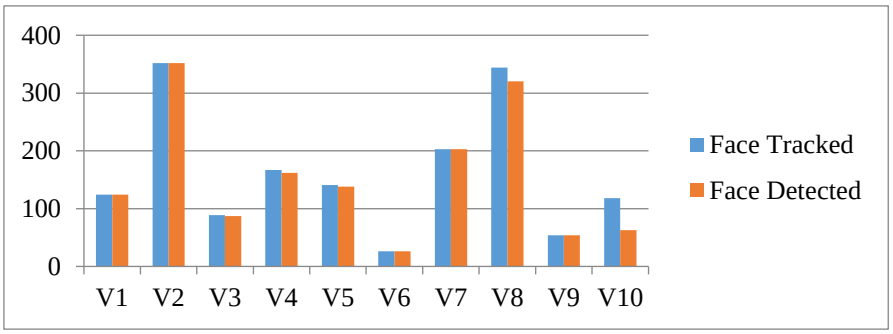

Figure 3. Face tracking and Detection for videos from Dataset

B. Cluster Generation: Following figure 4 shows the generated illustrative face clusters for a video in dataset. Each row represents a single cluster. Similar faces are grouped together in a single cluster. The cluster count is not the same as the number of persons in the video. The cluster count is dynamically defined as per the given fixed width. The fixed-width threshold is set as 1000, after testing with the multiple test videos. The person having the same pose, with the same illumination, and the same expressions are grouped in the same cluster.

Figure 4. Sample clusters of fix width clustering of two person from selected video sequence

C. Cluster Outliers: The following table shows the visual outlier count per cluster using K-Means and fixed-width clustering algorithm. The table contains multiple test cases from the videos in dataset, the cluster count, and outlier found in each cluster. There is too much variation in outlier count when we change the K value for K-Means clustering. For the sake of convenience, initially, fixed-width clustering results are collected. The value of k for K-Means clustering is defined the same as the cluster generated suing fixed-width clustering. The fixed-width clustering generates better results without having dependency with the cluster count. The outliers found using K-Means are higher as compared to the fixed-width clustering.

Table 1. Comparison of face clustering for visual outliers in K-Means and Fix width Algorithm for similar cluster count

Sr. No.	Video Number	Number of Clusters	Outliers in Fixed-Width clusters	Outliers in K-Means clusters
1	V1	19	00	00
2	V2	10	00	Cluster: 01
3	V3	06	00	Cluster: 01
4	V4	16	00	Cluster8: 01 Cluster 11: 01 Cluster 16: 01
5	V5	15	Cluster6: 01	
6	V6	14	Cluster1: 02	Cluster1: 04 Cluster 4: 01
7	V7	14	Cluster2: 10	Cluster 13 : 01 Cluster 14: 08
8	V8	29	Cluster1: 11	Cluster 1: 01 Cluster 4: 01 Cluster 6: 01
			Cluster2: 02	Cluster 7 : 02
			Cluster13: 01	Cluster 13: 01
			Cluster15: 01	Cluster 17: 01
			Cluster17: 04	Cluster 18 : 02
			Cluster20: 01	Cluster 20: 07 Cluster 23: 01 Cluster 27: 03 Cluster29: 06
9	V9	11	00	Cluster 6: 01 Cluster 9: 03
10	V10	14	00	Cluster 5: 02 Cluster 8: 01

D. Cluster Purity: It is calculated for cluster generated using K-Means clustering and fixed-width clustering. For K means clustering, the cluster count is set the same as the number of clusters generated using fixed-width clustering still the cluster generated have mixed data, for the previously known value of k cluster purity lesser as

K- means clustering is unable to handle mixed data. Fixed width clustering has a higher cluster purity value than the K-Means clustering results as shown in figure 5.

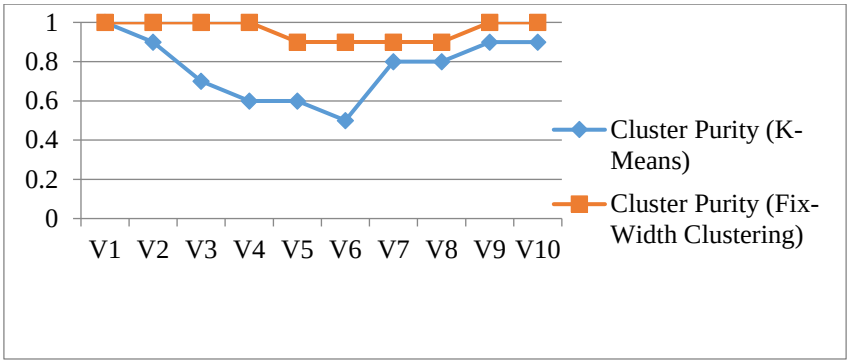

Figure 5. Cluster Purity for Illustrative videos from Dataset

E. Face Authentication: The user authentication process of the system is validated for the videos sequences in dataset. The face authentication process using K- Means clustering generate many false recognition result due to less clusters purity, 68% of faces are correctly identified and authenticated. Figure 6 shows the total number of faces occurred in a video and the total number of authenticated faces is compared in the authentication process. 80% of faces are correctly identified and authenticated from the video. The identified faces have variations in pose and illumination.

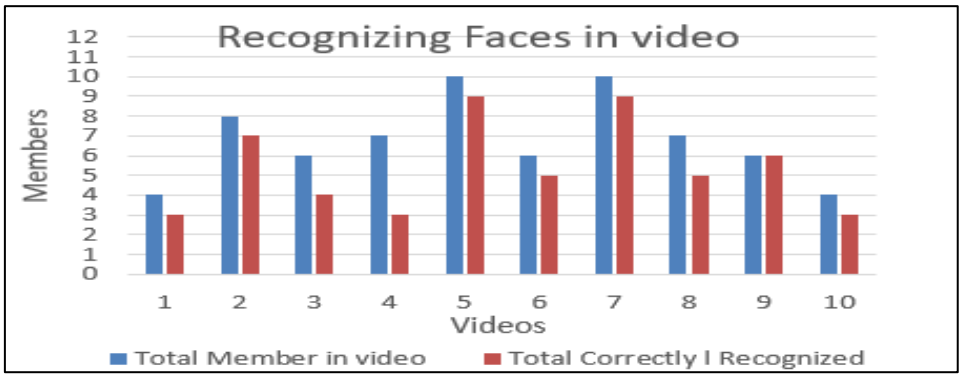

Figure 6. Face Authentication using fixed width clustering for videos from Dataset

Limitation of k-means is overcome with implementation of fixed width clustering process for faces from video here number of cluster before clustering process need not to specify. This clustering approach create new cluster if the new face cannot be the member of previously formed clusters. It takes the decision based width (w) parameter taken from user. So the number of faces are more compared to number of person in the video. As it forms a new cluster for persons, new pose, any change in expression etc. So there thus multiple cluster for same person. Clustering is done for faces so for face recognition input faces need not to compare all faces in the video whereas it has to compare to the top k faces of cluster only.

F. Authentication Time: The co-occurrence information of previously analysed videos is input to the system. After the face authentication system updates the co-occurrence information graph with respect to the current input video. Figure 7 shows face authentication time for different videos using fixed width clustering.

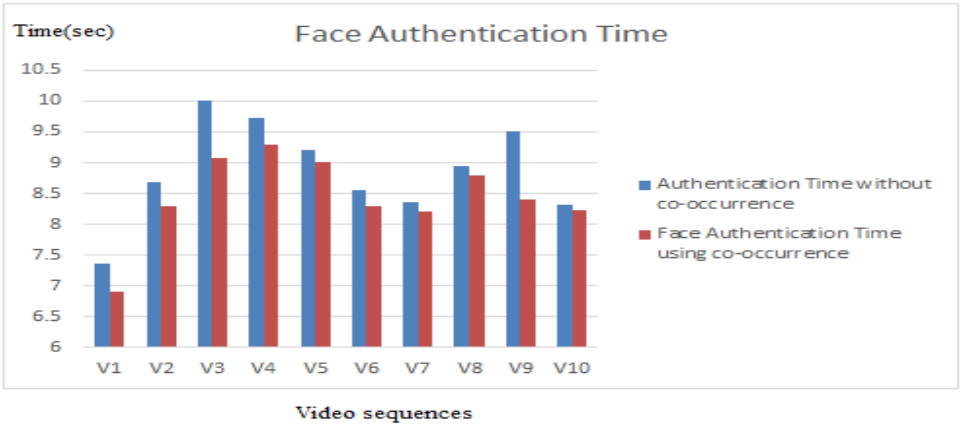

Figure 7. Face Authentication time using fixed width clustering

G. Co-occurrence Graph generation: Following fig 8 contains the summary generated from the video. The person occurrence count and Co-occurrence information is generated from the face recognition process. The Co-occurrence information is represented in a matrix of size n. where n is the number of persons in a video. Each value m_{ij} in the matrix represents the co-occurrence of person i and j in a video. Following figure 8 includes the graph G (V,E) representing the co-occurrence information based on the co-occurrence matrix. Each node V represents the person. Two nodes are connected if the co-occurrence value m_{ij} >0. The thickness of the edge represents the weight of co-occurrence. Highly co-occurred faces are connected using thick lines.

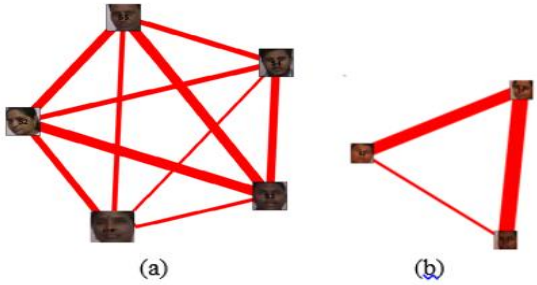

Figure 8. Co-occurrence Graph generation for two different video from Dataset

6. Conclusion

The proposed work mainly categorized into 2 sections: Video-based face recognition and video summarization. The recognition includes the identification and the authorization process. The input video is like a pre-recorded video or video data from a Surveillance camera. The front and profile faces are identified from the video frames. The identified faces are clustered together using fixed-width clustering, for efficiency improvement of the authentication process. Only top k faces are used for the authentication process and label is assigned to all the faces in cluster accordingly. The video members, their occurrence count, and the co-occurrence information is extracted from the video as a summarization process. The graph-based co-occurrence information generates the pictorial representation of video summarization. Fixed width clustering outperforms and using this technique 80% of faces are correctly identified and authenticated from the video. Though the identified faces have variation of scale and pose, expression etc. still the huge range of variations in expression, occlusion, and illumination limit the face detection process and thus approximates the extracted information.

Acknowledgement

We are very thankful to all the authors of literature used in this work. The referred techniques and concepts are really found useful and essential in the field of face recognition.

References

1. Annis Fathima,V Vaidehi,S Vasuhi, Mukund Murali,S K Parulkar (2015). *Pose Invariant Face Detection in Video* in Proceedings of the 2nd International Conference on Perception and Machine Intelligence , (pp 110–115)
2. Balasubramaniam, Vivekanandam & Dr. Babu (2017). *Adaptive Face Recognition Under Different Pose And Illumination Variation In Video Survelliance*. Journal of Advanced Research in Dynamical and Control Systems. 9,244-270.
3. Barr J. R., K. W. Bowyer, P. J. Flynn, and S. Biswas (2012). *Face recognition from video: A review*. International Journal of Pattern Recognition and Artificial Intelligence, 26(5)
4. Chai D. and K. N. Ngan, (1999). Face segmentation using skin-color map in videophone applications, IEEE Transactions on Circuits and Systems for Video Technology, 9(4), 551-564
5. Chen Y.C, V. M. Patel, S. Shekhar, R. Chellappa, and P. J. Phillips (2013). *Video-based face recognition via joint sparse representation*. In Proceedings of International Conference and Workshops on Automatic Face and Gesture Recognition, (pp 1–8)
6. Chen Y.C., V. M. Patel, P. J. Phillips, and R. Chellappa (2012) *Dictionary based face recognition from video.* In Proceedings of European Conference on Computer Vision, (pp 766–779)
7. Chu W., Yale Song and A. Jaimes (2015). *Video co-summarization: Video summarization by visual co-occurrence*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, (pp 3584-3592)
8. Cui Z., S. Shan, H. Zhang, S. Lao, and X. Chen (2012). *Image sets alignment for video-based face recognition.* In Proceedings of International Conference on Computer Vision and Pattern Recognition, (pp 2626–2633)
9. Dahake R., M. U. Kharat, Priti Lahane (2016). *Challenges and Advances in Human Face Recognition from Real Time Video*", International Journal of Advanced Trends in Computer Science and Engineering, 5(6), 114-119
10. Ding S., Ying Li, Junda Zhu, Yuan F. Zheng, Dong Xuan (2013). Robust video-based face recognition by sequential sample consensus", Advanced Video and Signal Based Surveillance 10th IEEE International Conference ,(pp 336-341)
11. Franco Annalisa, Dario Maio, F. Turroni (2014). *Spatio-temporal Keypoints for Video-Based Face Recognition*, Pattern Recognition (ICPR) 22nd International Conference ,(pp. 489-494)
12. Haq, I. U. K. Muhammad, A. Ullah and S. W. Baik, (2019) *DeepStar: Detecting Starring Characters in Movies*, IEEE Access, 7, 9265-9272
13. Harandi M. T., C. Sanderson, S. Shirazi, and B. C. Lovell (2011). *Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching*. In Proceedings of International Conference on Computer Vision and Pattern Recognition, (pp 2705–2712)
14. https://docs.opencv.org/master/d9/d52/tutorial_java_dev_intro.html
15. https://heartbeat.fritz.ai/deep-learning-based-video-summarization-a-detailed-exploration-8b1a25946404,written by Surva Remanan 24/01.**2020** retrieved on 24/02/2021
16. Kayal S. (2013). Face clustering in videos: *GMM-based hierarchical clustering using Spatio-Temporal data*, 13th UK Workshop on Computational Intelligence (UKCI), Guildford, (pp 272-278)
17. Kharat M. U. R. P. Dahake1, Kalpana V. Metre (2018). *Clustering Techniques for Content-Based Feature Extraction From Image* ,Feature Dimension Reduction for Content –Based Image Identification , IGI Global, A volume in the Advances in the Multimedia and Interactive Technologies (AMIT) Book Series,(pp 100-121)
18. Li B, Liu J. (2011). *The Connections between Principal Component Analysis and Dimensionality Reduction Methods of Manifolds*. In: Huang DS., Gan Y., Gupta P., Gromiha M.M. (eds) Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. ICIC, Lecture Notes in Computer Science, vol 6839. Springer, Berlin, Heidelberg.
19. Li, P, Patrick J. Flynn, Loreto Prieto, Domingo Mery (2019). *Face Recognition in Low Quality Images: A Survey*, ACM Comput Surv vol.1, no 1
20. Pande N, R. P. Dahake (2014). *Automatic Naming of Character using Video Streaming for Face Recognition with Graph Matching*", International Journal on Recent and Innovation Trends in Computing and Communication, 2(5)
21. Tadge S., Ranjana Dahake (2017). *Celebrity Face-Name Association in Web Videos using Unsupervised Approach*", International Journal of Advance Research and Innovative Ideas in Education 3 (4)

22. Wang R., S. Shan, X. Chen, Q. Dai, and W. Gao (2012) *Manifold-manifold distance and its application to face recognition with image sets*, IEEE Transactions on Image Processing, 21(10) 4466–4479
23. Workie, Ashenafi & Rajendran, Rajesh Sharma & Chung, Yun. (2020). *Digital Video Summarization Techniques: A Survey*, International Journal of Engineering and Technology. 09(5)
24. Y. Hu Y., A. S. Mian, and R. Owens (2011). *Sparse approximated nearest points for image set classification.* In Proceedings of International Conference on Computer Vision and Pattern Recognition, (pp 121–128)
25. Zhang L., Dmitri V. Kalashnikov, Sharad Mehrotra, (2014). *Context Assisted Face Clustering Framework with Human-in-the-Loop*, International Journal of Multimedia Information retrieval
26. Zhang X., F. Pala and B. Bhanu (2017) *Attributes co-occurrence pattern mining for video-based person re-identification*, 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (pp 1-6)