A Review of Scholastic Examination and Models for Face Recognition
and Retrieval in Video

Varsha Sachdeva¹, Junaid Baber², Maheen Bakhtyar², Muzamil Bokhari³, Imran Ali⁴

¹Department of Computer Science, SBK Women’s University, Quetta, Balochistan
²Department of CS and IT, University of Balochistan, Quetta
³Department of Physics, University of Balochistan, Quetta
⁴Institute of Biochemistry, University of Balochistan, Quetta

Abstract: - This survey report overviews recent frameworks on face recognition and retrieval solutions for videos search and surveillance applications. We introduce an engineering of a bland framework for facial detection, retrieval and recognition in video and audit scholarly frameworks reported in the scholastic writing suitable for videos search and surveillance applications. Suggestions on the choice of frameworks for benchmarking and the investigation of future patterns is introduced. Methods displayed in this report are those that give great results on understood reference video information sets and can be utilized to give establishments to build up a better retrieval system.

Keywords—: video-surveillance, video retrieval, face recognition and retrieval in video, instant face recognition, watch-list screening, reliability, performance evaluation

INTRODUCTION

Shots retrieval of the particular person by giving his face image as a query becoming more and more interesting now a days. As the volume of videos are increasing exponentially on the internet, so dealing with that huge database is becoming challenging task of computer vision. If we have given a shot, and wanted to know that who is this person appearing in this video, or is there any individual present in this video clip or not, and If the required person is present in this video or not. Such innovation can be used to gather the information related to the specific person, or detecting the shot of the particular person. Various algorithms have been introduced in literature for detecting, recognizing and retrieval human faces in videos. Previously work has been done on the videos which have straightforward content like news video. Some authors have introduced methodologies that work on to distinguish individuals in videos. for example, clothing comparability (P. Wang, 2003), voice acknowledgment (P. Wang, 2003) Taskiran, voice detection (Ikizler & Duygulu, 2005) and so forth...

With this motivation and interest and advancement in face detection, the standard methodologies have now been centered on face based innovation in videos. In (Le, 2006.), have introduced reporters face grouping and clustering in news videos (T. L. Berg, 2004) introduced a technique to group

The faces from the images in news websites. (Arandjelovic & Zisserman, 2005) Cipolla proposed a method for clustering of faces appearing in films of complex model. (Ozkan & Duygulu, 2006) proposed a method to detect oftentimes showing up countenances in news videos taking into account a face similitude chart. (I. Naseem & Bennamoun, 2009) associated small representation game plan towards the issue of common recognition of face under controlled circumstances. Nevertheless, these past techniques encountered the going with limitations. To begin with, most by far of them were compelled to controlled circumstances. Case in point, a few approaches rely on the supposition that the environment to be break down have clear establishments and the target persons don’t have tremendous developments (for example news columnists). In unbounded environment where people have optional position, lighting up and hindrance assortment, such methodologies will miss an extensive measure of face events. Second, a large portion of face grouping procedures wore down appearances in isolated video plots and did not consider the common relationship in a constant face course of action (a social occasion of back to back video diagrams with the same face).

VIDEO-BASED FACE DETECTION, RECOGNITION AND RETRIEVAL

In past work that improvement social orders see front face postures, particularly when structural picture
quality is less. The customary detection algorithms are all considering still pictures however videos based face detection and recognition have been a component inspect point for a long time. It is sorted in two frameworks and these are i) Procedures based on set ii) successive based or approaches

Set-based procedures consider videos as disordered social occasions of images and attempt the unlimited number of recognition, whereas successive based or sequential based frameworks unequivocally utilize transient data to gather gain fullness off course empower detection and recognition in poor review conditions. Video based face detection recognition and retrieval frame-works incorporate three modules.

FACE DETECTION

Human face detection is the prerequisite or first step of a face recognition framework. The framework takes a bunch of video movement and performs some image processing techniques on it recalling the completed target to discover finds contender confront area. Framework can wear down still images, where this theory of face detection and administration in video methods is known as face tracking. The theory behind face detection from the foundation on still images can be performed by using few parameters and these parameters are skin color or surfaces, facial/head shapes, 3D face shapes or blend of these parameters. In face detection, face recognition is acted like asking for the delineation in the window that it is a face or non-face window. The most acclaimed utilized algorithm for face recognizing evidence is Voila & Jones Algorithm which is described underneath.

a) Voila Jones Algorithm

This segment depicts the work did, in this section we describe the use of voila jones face detection algorithm. The most important part clarifies the theory behind the algorithm. Instead of keeping the copy of primary voila jones paper, this section is kept little short but at the same time all the important things are elucidated. Additionally, entrancing parts of the genuine use are pushed.

The important step of the Voila and Jones algorithm is to analyze the sub window which is prepared for the recognition of face in the given input image. The processing technique of image would be to rescale the locator to different variant sizes and after that run the adjusted size discoverer through these edges. The philosophy winds up being to some degree repetitive in view of the check of the different size pictures.

Regardless of the monetary standard approach Viola Jones rescale or establish on a new scale the locator as opposed to the data casing and run the discoverer customarily through the edge, each time with a substitute size. At introductory one may analyze both techniques to manage be pretty much as repetitive, however Viola and Jones have created a consistent scale indicator that need the same number of computational efforts whatever the size. This identifier is produced using an assumed crucial or integral image and some essential rectangular components reminiscent of Haar like features or Haar wavelets. The accompanying zone explains this pointer.

The scale invariant detector the methodology of voila jones face detection algorithm is to change the input image into an integral image and this can be done by making each single pixel proportional to the all pixels above and to the left and right side of the concerned pixel. This is appeared in Figure 1

```
1 1 1
1 1 1
1 1 1
```

```
1 2 3
2 4 6
3 6 9
```

Fig. 1. Representation of an Integral Image. Reprinted from (J. SUNEETHA (2014))

This reflect over the number of the total of all the pixels in any given rectangle by using only four values. This is appeared in Figure 2

```
A

D
```

Sum of Gray Scale = D - (B + C) + A

Since the values of both the rectangles B and C are fuse to A, and the whole of the necessity is added to the calculation. It has been now that how the pixels within rectangles of different sizes can be found in reliable time. The Viola and Jones face detector algorithm separates a given sub-window using points
containing two or more rectangles. The differing sorts of features are showed up in Figure 3.

![Figure 3](image.png)

**Fig. 3.** Types of features. Reprinted from (J. SUNEETHA (2014))

Each single feature of an image gives about a value which is learned by subtracting all the white rectangles from the dark rectangles. Voila and Jones have analyzed that a pointer with determination of 24*24 pixels give fine results. While thinking about each and every possible size and location of feature in above figure 4, total number of 160,000 feature can be created. In this way, the measure of possible features unfathomably predominate the 576 pixels contained in the finder at base determination.

These segments may look, by all accounts, to be unreasonably simple to perform such a moved assignment as face identification, yet what the parts require in capriciousness they without a doubt have in computational adequacy.

**b) The Altered Ada-Boost Algorithm**

As explained above that there can be 160,000 feature values within the base determination. Between each of the feature there are almost very few features are to give high values when on the top of the face. Voila Jones used an altered form of the algorithm made by the following author in 1995 (Freund and Schapire). Adaboost is basically a machine learning algorithm prepared for a strong classifier via mix-up of strong and weak classifiers. A feeble classifier is numerically depicted as

\[ h(x,f,p) = \begin{cases} 1 & \text{if } pf(x) > p \\ 0 & \text{otherwise} \end{cases} \]

Where in above equation \( x \) is 24x24-pixel window, \( f \) is the concerned feature, \( p \) is the extremity and is the limit of whether the \( x \) is positive that is it contains a face or else it is negative that is it is a non-face window. The Voila Jones algorithm is changed to pick only the best parts of the image and ignores the useless parts. The complete voila jones algorithm is shown in pseudo code in figure 4.

![Figure 4](image.png)

**Fig. 4.** The altered AdaBoost algorithm. Reprinted from (Lu Sheng-nan (2015))

The important portion of the Adaboost algorithm is the determination of the best features or edges. Before there was no better technique for this thing but voila and Jones propose a novel technique. This suggests that the determination of the new weak classifier incorporate and evaluating each part on all the preparation test with a particular deciding objective that is to find the best representative features. This is the slowest and dull part of the preparation technique. The best feature is then picked in perspective of the weighted error it delivered. This weighted screw up is a segment of the weights fitting in with the preparation samples. As found in Figure 5 segment 4) the heavi ness of an adequately masterminded case is lessening and the greatness of an unclassified test is kept unfaltering. Likewise, it is all the more “unreasonable” for the second element (in the last classifier) to misclassify a representation moreover misclassified by the essential segment, than a case portrayed successfully.

Another choice is that the second feature is constrained to concentrate on the specimens misclassified by the first. The truth being that the weights are a fundamental part of the mechanics of the AdaBoost calculation.

With an integral or crucial image, the successful computational parts and the balanced voila Jones algorithm set up it seems like the face locator is arranged for utilization, however Viola and Jones have one all the more master up the sleeve.

**c) The cascade classifier**

The crucial step of voila and Jones algorithm is to check the detector for a face through the same input image each time with another size. Instead of detecting a face each time with different window sizes
the algorithm ought to detect the non-face appearances because of the issue of the gigantic measure of surveyed sub-windows would be negative at present. The advantage behind the above mentioned technique is that it is speedier to discard a non-face region than to detect a face region. Thus a necessity for a cascade classifier is developed. The cascade classifier is made by the stages, each containing a strong classifier. The task of each stage is to make sure that the window is surely not a face or conceivably a face. At whatever point sub-window is placed if there is a non-face region then it is a flash discarded. Then again a window named a maybe face is gone ahead to the accompanying stage in the course. It takes after that the more stages a given window passes, the higher the risk the window truly contains a face. The thought is spoken to with two stages in Figure 5.

The extraction of the dominant features is the important issue in face detection problems. Subsequent getting an image contain a face, then the step is to think for a facial part for the detection of the face in an image. There are two kinds of features are being extracted for the identification of the facial parts and these are 1) Geometric Features 2) Appearance-Based Features. Geometric Features addresses the 3D shapes and zone of facial parts, for example eyebrows, eyes, nose. The Appearance-based approaches use the skin surface of the face for face detection in still images or videos.

As said before, because of the quick extension in storage and computation assets, appearance-based techniques have commanded the late advances in face detection. The general practice is to gather a huge set of face and non-face illustrations and receive certain machine learning algorithms to take in a face model to perform classification.

There are two key issues in this procedure: what features to extract, and which algorithm to apply. In this segment, we first survey the late advances in feature extraction.

The Haar-like rectangular features as shown in above Fig 2 and 3 are exceptionally proficient to process because of the integral image method, furthermore, give great execution to building frontal face detection. In various subsequent works, analysts developed the clear features with more varieties in the ways rectangle features are consolidated. The Haar-Like Features used in voila jones algorithm has good computational efficiency and perform good in frontal face detection. After that many researchers modified this feature for multi view face detection. In (Li et al., 2002) proposed three kinds of rectangular features with variant sizes and distances. As shown in Figure 6.

**d) Progressions in Feature Set and Boosting Algorithm:**

One would consistently recognize the false negative in single stage classifiers by remembering the finished objectives to reduce the false positive rates. In any case, for the primary stages in the arranged classifier false positives are not considered to be an issue resulting to the succeeding stages are depended upon to manage them. Thusly Viola-Jones embrace the affirmation of various false reassuring focuses in the hidden stages. Consequently, the measure of false negatives in the last masterminded classifier is depended upon to be little. Viola-Jones furthermore allude to the cascade classifier as an attention course. This name derives that more thought (computing power) is composed towards the ranges of the info outline suspected to contain confronts. It takes after that when preparing a given stage, say n, the negative examples should clearly be false negatives created by stage n-1.

The greater part of contemplations presented in the “Methods” section is taken from the principal Viola-Jones paper.

![Fig. 5. The Cascade Classifier. Reprinted from (J. SUNEETHA (2014))](image)

![Fig. 6. The Stages of Cascade Classifier. Reprinted from (Li et al., 2002)](image)
FACE RECOGNITION AND RETRIEVAL

Face recognition is the important step in the whole structure. Videos give much richer information than the still image. The colossal reasons for energy for utilizing videos are Firstly the likelihood of utilizing repetition contained as a part of the video strategy to enhance still images affirmation execution, second component data is open and thirdly to overhaul affirmation influences from the video movement utilizing all the more extreme representations, for occurrence, a 3-Dimensional face model. At the end face recognition based on videos ensure learning the subjected model of face in less time. Face recognition frameworks separated into two approaches 1) Frame-based approaches for face recognition 2) Sequence-based approaches.

The first one is based on concerning static images and second one is the masterminded in the light of component video frames. Sequence based utilizes the common information of the motion pictures to see the information of one or more edges or frames. The figure 7(a) and 7(b) represents the general framework of face recognition and retrieval systems respectively.

![Diagram](image)

Fig. 7 (a). General Structure of Face Recognition. Reprinted From (J.SUNEETHA, 2014). (b) General Structure of Face Retrieval System

Shot Detection

Videos are organized by diving pecking order of video clips, scenes, shots, and frames. Video structure examination goes for fragmenting a video into various auxiliary components that have semantic substance, including shot boundary detection, key frame extraction, and scene segmentation.

A shot is set of consecutive frames which have greater similarity between them, and is captured by a camera action that take place between start and stop. As frames have greater correlation between them that is why shots are considered to be the single unit of video. Shots are divided into two types that are Gradual Shots, Abrupt Shots. Gradual shots are the shots which contain fade in and out, wipes and dissolve effects. These kind of shots are combined using chromatic, spatial-chromatic effects which slowly and gradually replace one shot by another. While in Abrupt Shot Boundaries there is a sudden change from one shot to another. There is a big difference between two consecutive frames in abrupt boundaries. The second frame is the start frame of next shot. These kind of shots are also known as hard cuts or simply cuts.

Every technique for cut recognition takes a shot at a two-stage standard:

Scoring - Each pair of back to back frames of a computerized video is given a specific score that tells to the comparability/uniqueness between these two frames.

Decision - All scores computed already are assessed and a cut is distinguished if the score is viewed as high.

In the first place, in light of the fact that even minor exceeding of the edge esteem delivers a hit, it must be guaranteed that stage one disperses values broadly to expand the normal contrast between the score for cut and no-cut. Second, the edge must be picked with consideration; typically, helpful qualities can be picked up with measurable techniques. The Sum of Absolute Differences (SAD) and histogram differences are mostly used for scoring.

SAD is the simplest algorithm for finding shot detection in videos. The two back to back frames are compared pixel by pixel, summing up their absolute difference of every two comparing pixels. The outcome is positive number that is used as a score. SAD respond delicately to the minor changes inside scene, camera and this leads to false recall. However, SAD is utilized frequently to find out the hard cuts in videos because of its simplicity.

Histogram Difference computes the difference of histogram of two consecutive frames. A histogram is a table that contains the number of pixels for each color in a frame. Histogram difference is not as
sensitive as SAD. One major problem in histogram difference is that histogram of two different image can be same, e.g. the histogram of sea and beach can be same histogram. HD offers no guarantee for it.

In decision phase different approaches are usually used. Some of them are

a) Fixed Threshold:

In this approach, the threshold is compared with the score, if the threshold is less than score then the cut is detected.

b) Adaptive Threshold:

In this approach, the scores are compared to a threshold which considers various scores in the video to adapt the threshold to the properties of the current video. Like in the previous case, if the score is higher than the corresponding threshold a cut is declared.

c) Machine Learning:

Machine learning techniques can be applied also to the decision process.

These all modules have most of the thing in common, which feature extraction methods and approaches, indexing and similarity measure techniques. And these common steps are described below into detail.

**FEATURE EXTRACTION TECHNIQUES**

There are several ways to extract the features. We briefly explain some famous techniques.

a) Geometric based Approaches:

In Geometric Based approaches the features are extracted using geometric information such as the position and size of the face components. In this kind of approach local features and their geometrical features are analyzed. These techniques don’t require threshold, which adversely affect the achieved performance. In template based approaches the input image is compared with the set of template images which are constructed using the tools like PCA, LDA and ICA. For the extraction of facial features energy function is used.

b) Color Segmentation Based Approach:

As we all know that most of the images have noise or have complex background which is considered as noise. So, in this kind of approach first of all the skin color is detected and then the facial components like eyes, nose etc. are extracted from the image. After getting the image from the skin color the image is converted to gray scale image. And after the gray scale it is important to apply the suitable threshold to remove the hue and saturation. This approach is limited because of the background noise.

c) Appearance Based Approaches:

Appearance Based Approaches represents the face in so many raw intensity images. It is important to keep the important feature as when to recognize the face, when we apply the feature extraction methods. Examples of this approach is PCA, LDA, or ICA. The main pros of this approach is that it keeps the important features and removes or eliminate the redundant features.

d) Template Based Approach:

In this approach the input image is compared with the template and the template are constructed using different tools like PCA, LDA, or ICA

e) PCA Based Method:

Principal component analysis (PCA) is a statistical system that uses an orthogonal variation to change over an arrangement of observation of potentially related variables into an arrangement of estimations of straightly uncorrelated variables called principal components. The quantity of principal components is not exactly or equivalent to the quantity of unique variables. This change is characterized in a manner that the primary chief segment has the biggest conceivable fluctuation, and each succeeding segment thusly has the most astounding difference conceivable under the limitation that it is orthogonal to the first parts. The subsequent vectors are an uncorrelated orthogonal premise set. The main segments are orthogonal on the grounds that they are the eigenvectors of the covariance matrix, which is symmetric. PCA is touchy to the relative scaling of the first variables.

When we give a $T$-dimensional vector presentation of a face image, the PCA is being utilized to find out the subspace which correspond to the maximum maximum-variance direction in the original space. Let $W$ represents the transformation that perform the mapping of t-dimensional into $F$-dimensional feature subspace where $F < T$. 
The new feature vector \( y_i \in R^f \) is defined as \( y_i = W^T x_i \).
Where \( i = 1, 2, 3...N \). The column of the matrix \( W \)’s the eigenvalues of \( e_i \) obtained by solving the Eigen structure decomposition \( \lambda_i e_i = Q e_i \). Where \( Q \) is the XX\(^T\) is the covariance matrix and \( \lambda_i \) is the eigenvalue associated with \( e_i \).

Before go further with eigen values firstly the vectors are normalized to make the system invariant or unchangeable, and secondly the values from images are subtracted from the normalized values, assures the eigenvalues with the maximum correlation sense. The covariance matrix \( Q \) is too large for computation of the eigen vectors. (Aleix M et al., 2001)

f) **LDA Based Approaches:**

LDA is an extension of PCA. It constructs a linear subspace that maximize the scatter between images of different classes. The dimensionality of the subspace is fixed between number of features, number of samples, and number of classes. The between-class scatter matrix \( V_b \), and within-class scatter matrix \( V_w \) and the projection matrix P can be written as follows:

\[
V_b = \sum_{i=1}^{c} N_i (m - m_i)(m - m_i)^T
\]

\[
V_w = \sum_{i=1}^{c} \sum_{j=1}^{N_i} (x_j^i - m_i)(x_j^i - m_i)^T
\]

Where \( m_i \) and \( N_i \) are the average face and sample number in each face class respectively, and \( x_j^i \) is the \( j^{th} \) sample in the \( i^{th} \) class. The purpose is to maximize the between-class difference and to minimize the with-class distance. One way to do this is to maximize the ration of \textit{ratio} = \frac{\text{Det}[V_b]}{\text{Det}[V_w]} . The advantage of using this ratio is that if the \( V_w \) is the non-singular matrix then the ratio is maximized. (Aleix M et al., 2001)

When taking class discriminatory information, it is necessary to perform dimensionality reduction at that point. When classes are separated seek to find direction. Because of the illumination and expression variation LDA is more capable of distinguishing these image variations.

g) **Independent component Analysis:**

Rather than finding the uncorrelated image decomposition and representation, ICA find independent. ICA separated the Non-Gaussian distributed features. This Method is the extension of PCA which leads to the criteria of discriminant analysis.

h) **Kernel PCA:**

As the above mentioned PCA is linear, so in order to increase the capability of PCA [author name] have introduced non-linear PCA known as Kernel-PCA. In this technique the non-linear mapping is first applied to input and then for feature space in results it uses linear PCA

i) **Local Binary Pattern:**

This is the one of the most powerful feature extraction method which describes the image as a texture and divide it into mainly three different features that are pixel level, region level and global level. This method gives label to each pixel in image by 3x3 neighborhoods of each pixel and assigns the one pixel as a center pixel and that label of histogram is used as texture. LBP is broadly used in face recognition due to its uniqueness of features and less computational time.

**VIDEO-BASED RECOGNITION FRAMEWORKS**

There are different approaches for video based face recognition frameworks. Some of them are spatial-common information based systems, statistical model based approaches, and hybrid cues based procedures.

**a) Spatio-fleeting data based approaches**

Recent approaches use the Spatio-temporal information for face recognition approaches in videos. There also exists several algorithms for extracting 2D or 3D face structures from the videos. The distance between two videos is basically the distance between the two frames of the videos. Spatio-Temporal approaches are basically the storage of information extracted from the video images based on the pixel values and time example sequences. Spatial means the space, it means image consists of so many pixels which is need to store in memory. This is called Spatio Information. Temporal means time, video consists of image frame sequences which changes with respect to time in video.

This is called Temporal Information

The techniques considering Spatio-temporal representations for face recognition in video have a few obstacles:

- The local features extracted from the facial images is very important to facial image analysis, so it is not well exploited.
- Specific person facial feature is useful for discriminating the two different faces of...
different persons, So the intra-temporal information of specific person which is related to facial expression and emotion is also encoded and used.

- The weights are equally distributed to spatiotemporal features (Pietikinen, 2009)

b) Statistic show based approaches

(Chellappa, 2002) get the statistical model from videos by using low level features extracted from the images which is further used to perform matching between the single frames or between two video streams. The common subspace procedure in (Gregory Shakhnarovich, 2004) work with the video frames and took the video frames for each individual to calculate the individual eigenvalues, considering the angle between the image and reference subspaces which is calculated from the principal component of the image sequences. Principal component null space analysis (PCNSA) is proposed in (Vaswani & Chellappa, 2006), which is pleasing for nonwhite clutter covariance matrices. Starting late, the Auto Regressive and Moving Average (ARMA) model procedure are proposed in (S. Soatto & Wu, 2001) to exhibit a moving face as a straight dynamical challenge. (Y. Xu & Patel, 2008) S. Soatto, G. Doretto, and Y. Wu proposed dynamic surfaces for video-based face acknowledgment. (M. Kim & Rowley, 2008) Kim et al. associated HMM to settle the visual impediments issue for face taking after and recognition (M. Kim & Rowley, 2008).

c) Hybrid cues based approaches

Video give much wealthier information than the still images. In (C. Shan, n.d.) found that the combination of the face and gait at feature extraction level can gained performance by combining two cues and the author adopted the face and speaker recognition techniques for audio video biometric recognition. In (M. Balasubramanian & Ramalingama, n.d.) authors presented the approach by implementing the radial basis function neural network, which is used to recognize a person in video streams by utilizing the mouth and face models.

MAJOR CHALLENGES IN FACE RECOGNITION AND RETRIEVAL

a) Illumination:

It is difficult for the system to recognize the individual when there is big change in lightning. (Georghiades, 2001) utilize principal component analysis and achieved better results from images under different lightening conditions. This method also deals with the shadowing and multiple lightning condition, the base of which is 3D subspaces. The drawback of this method is that the training image set needs to be aligned three image per individual.

b) Pose:

Posing of the face is the basic thing for face recognition systems. Present approaches are divided into three groups, these are 1) multiple pictures based techniques 2) Hybrid based approaches, 3) single still image based approaches. A method by the 3 Dimensional model of the complete set out toward manhandling features like hairline, which dealt with tremendous stance assortments in head taking after and video-based face affirmation was shown (Everingham & Zisserman, 2005).

c) 3D investigates:

Face acknowledgment in perspective of 3D is a hot examination subject. All around, reprehensive systems can be parcelled into three essential orders, specifically,

2D pictures based, 3D picture based and multimodal structures. The qualifications among these three characterizations are according to the accompanying: the key class consolidates approaches which use 2 Dimensional pictures and 3 Dimensional non particular face model to improve the force and acknowledgment rate. Besides, the 2nd one, the procedures work clearly on 3 Dimensional data-sets. While the last assembling infers those which utilize both 2D and 3D information. Since 2000, more multimodal techniques have been proposed to upgrade face recognition execution. Dalong Jiang et al. (Dalong Jiang, 2005) proposed a capable and totally customized 2 Dimensional-to-3 Dimensional joined face recognition system in an examination via mix way.

d) Low Resolution:

It is difficult to see human appearances in the video of low determination. With the comprehensive use of a camera (perception etc.), game plans which deal with such issues fulfill progressively thought. The guideline two strategies are Super Resolution (SR) besides, Resolution-confronts approach. In (Jeremiah r. Barr, 2012), shading invariance was associated with face affirmation. Their result showed that shading invariants do have liberal discriminative power and
extend the energy and precision for low determination facial.

**APPLICATION SURVEYED**

Human Based Video Retrieval frameworks have wide range of applications which examined broadly. (Y. Xu & Patel, 2008) have evaluated the methodologies on video indexing and retrieval based problems. Overall the human face video retrieval systems have same structure and steps which of feature extraction, indexing similarity measure etc. (Sivic & Zisserman, 2003) Video retrieval is the object retrieval for articles and video using invariant and robust descriptors from images of videos. (Arandjelovic & Cipolla, 2006) focuses on the retrieval of characters in Television series and shows. The image query human face video retrieval has been a challenging computer vision task. (Arandjelovic & Zisserman, 2005) and (Tong Zanga, 2013) proposed a methodology for utilizing the image query to search and retrieve video shots of specific person from extensive database. They have evaluated three procedures, and exhibit proportional accuracy with different types of movement. (Le, 2006.) proposed a method to recoup the object using labels and heading and appearances, regardless, the need of parallel enrolling plot make it is hard to be used as a piece of such huge dataset. (T. L. Berg, 2004) oversaw with tremendous video data by method for video rundown, and the issue is point by point as a base inadequate revamping of the principal video. (Ikizler & Duygulu, 2005) keep their attention on adjusting to the semantic space by unmoving focuses models. Previous and existing examination concentrates in a manner of speaking on specific parts of video retrieval. On the other hand, this contribution takes into consideration the complete structure of human face video retrieval from the massive dataset

Expansive examination and testing has been done over the huge video datasets for human face video retrieval type frameworks and tackle the issues regarding the video indexing and retrieval. In past the researchers (Arandjelovic & Cipolla, 2006), (Arandjelovic & Zisserman, 2005) uses the SVM (Support Vector Machine) and HMM (Hidden Markov Model) methods to deal with the above mentioned issues. (Ikizler & Duygulu, 2005) propose particular procedures for fusing sound and visual information for video course of action of Television undertakings in perspective of HMM. In (Ikizler & Duygulu, 2005), content elements from sound and video from pictures are joined to portray show news videos utilizing meta request by method of SVM. (C. M. Taskiran, 2004) proposed the time interval multimedia event(TIME) framework as a solid technique for portrayal of semantic events in multimodal video data. (Tong Zanga, 2013) and (Sivic & Zisserman, 2003) used semantic connections among video and sound objects of variant models for cross-media indexing. In cross-media indexing and retrieval, the request cases and recuperation results need not be of the same media sort. Case in point, customers can request pictures by submitting either a sound case or a picture outline in cross media retrieval systems. In (Sivic & Zisserman, 2003), worked for the media objects of different modalities and similarity measure technique is utilized for the retrieval. In (Sivic & Zisserman, 2003) for each user query, the perfect estimation of cross-media indexing space (CMIS) is actually chosen from get ready data and the cross-media indexing is performed on a for every query premise.

In (Y. Xu & Patel, 2008), decided the issue of commonly exhibiting the content and picture parts of interactive media chronicles. Associations between the two fragments are academic using acknowledged relationship examination and reflection is proficient by addressing substance and pictures at a broader level. It is showed up in (Y. Xu & Patel, 2008) that speaking to both cross modal associations and semantic reflection push ahead recovery precision. Unlike the above papers, this paper uses a summed up measure of association, the planned information, between multimodal (sound and video) data streams to finish better gathering and retrieval execution.

We further explain two famous frameworks in detail, VideoWHO (Tong Zanga, 2013) and SurvSurf (Ding et al., 2016).

a) SurvSurf

The authors of SurvSurf (Ding et al., 2016) introduces human face retrieval framework on huge amount of surveillance videos. They introduce another term, M-Clips to represent video. By working on M-clips, they improve their performance in compressed structure. They perceive video substance that are most committed to contain humans, subsequently lessen the data volume. They affect the improvement of data to vitalize visual operation among the video substance examination. Such advancement guided speedup can in like the way be connected with other operations other than human face retrieval.

They organize the V-BigTable to store the structuralize M-Clips semantic data and retrieve the results accord-ing to user queries effectively. They understand Surv-Surf in context of hadoop (“Apache Hadoop”, n.d.) and HBase (“Apache HBase”, n.d.). They assemble 11 machines to assess SurvSurf. The results which is explored, demonstrate SurvSurf
efficiency with satisfactory results in human face retrieval accuracy.

**System design:** In this area, we first outline and represent the work process of their framework. At that point we examine the principle parts in their framework.

**System overview:** Framework Flow outline of SurvSurf structure. As showed up in Fig. 8. Their system includes four essential parts: video preprocessing, video data investigation, formalization and online processing of query.

![Fig. 8. System Overview of SurvSurf. Reprinted from (Ding et al., 2016)](image)

**Video pre-processing:** The main purpose behind the video preprocessing is to lessen down the dummy data substance and segments the video in pieces aka clips which is further used for processing. In any case, they propose using the development data contained as a part of the video to help the video segmentation into clips. In development they observe the moving things in the video and denoise the video by removing the movement values and look at the development changes. Because of the varying motion vectors, the videos are segmented into clips or shots called M-clips or M-Shots. Dividing the video into shots is beneficial for preprocessing because it makes the system speedier.

**Video data investigation and findings:** After the video preprocessing, the videos are divided into clips and these clips are further exchanged to the dispersed document system. The author used the Map reduce functions to reduce the dimensionality of the M-clips and in parallel they perform mappers to do couple of calculations like to find the human location in given frame, face features extraction and action recognition. After the analysis and investigation authors re-move some useless semantic information from the M-Clips and some feature vectors.

**Formalization/Structuralization:** After the information extraction from the M-clips the authors wanted to store the extracted information into scattered, versatile and flexible data store in perspective of big table, called V-BigTable. One line of the V-BigTable represent the information of one cut, including feature vectors, non-highlight catchphrases can be put into the V-BigTable, which can be filled amongst the Map Reduce execution. Their framework consists of three types of the data.

1) Queried by Video Clip
2) Query by Image
3) Textual Keywords.

They applied same detection algorithm on the video and image query for individual appearance components or activities.

By then they look for the V-BigTable through literary visual keywords organizing and incorporate vector relationship. The last results to the customer are sorted in perspective of the feature comparability.

**Evaluation:** Around there, they lead focused investigations to survey their structure. Their appraisal generally spotlights on the time viability of the structure. They furthermore take a gander at the exactness of human face retrieval framework. The video retrieval algorithm proposed by (Araujo A, 2012) is used as the comparable technique in appraisals retrieval and time cost precision.

**Correctness of M-clips or Shots generation:** They assess the SurfSurv M-clips generation n the following things.

- They review their system on 15GB of dataset
- The ground truth of the M-Clips of that dataset is 412 among of which 403 of them is remedy.
- Accuracy equivalents to 403 (accurately created M-cuts) partitioned by 412 (all produced M-cuts), which is 97.81%.
- Review equivalents to 403 (effectively generated M-cuts) isolated by 405 (accurately generated M-cuts plus missed M-cuts), which is 99.50%.
- This implies 97.81% of the produced M-cuts really contain movement and 99.50% of all movement that ought to be found is incorporated into the created M-cuts.

**Preprocessing Time cost and information uploading:**

In the common Map Reduce structure, information is spread to different pro hubs to estimate the different pieces of video clips, each of which has level with record measure. Before processing is done on the clusters, they first process the M-Clips and find out the bounding boxes for each clip. They exchange M-Clips and pass them to different laborer hubs. They
noticed the preprocessing time and data exchanging time for data under 5000 MB in size are insignificantly more than the traditional Map Reduce size. But in any case of larger dataset their technique takes less time than estimation. This makes their system much more efficient than the original Size-based map reduce. That is why they process and upload M-clips which is shorter in size rather than the whole video. Which saves time paying little respect to the overhead of preprocessing development vectors. This time save finances augments with the data measure. This figure below 9(a) shows the time and data relationship in the system.

**Accuracy of human retrieval:** From the figure 9(a)(b) we can see that the length of the impediment is not extreme or the individual is not completely impeded, it is still conceivable to retrieve the individual effectively. Nonetheless, the impediment will be an issue when a swarm shows up in the scene.

### b) VideoWho

This framework is designed for separating face sequences in different sort of uncurbed videos like home camera video, television movies and dramas and shows, surveillance videos etc. This system is proficient by face detection and tracking system. Secondly, the authors of VideoWHO (Tong Zhang, 2013) introduced new idea of making the clusters of faces appearing in the uncontrolled videos sequences (as contrasted and clustering face images in their earlier work)

What’s more, actualize a unique algorithm to grouping faces sequences, which have critical stance, enlightenment and impediment varieties. In order to perform this demanding function, the algorithm considers not just similitude between each pair of face samples, additionally their transient setting and quality in the concerned video, bringing about a semi-regulated clustering technique particularly to cluster face successions or sequences. Test results demonstrate that the VideoWho (Tong Zhang, 2013) framework accomplishes genuinely great get back rate in discovering face examples in unconstrained videos. The precision of consolidating face groupings fitting in with the same individual is additionally agreeable. (Deng J, 2011) This report exhibits a study of the past academic frameworks that have been proposed for videoWHO and serves as the premise for the work led inside of the VideoWho venture. Procedures introduced in this report are those that give great results on surely understood reference video information sets and can be utilized to give establishments to add to a face retrieval system.

**System Diagram:** It is the fully automatic persons based video summarization and retrieval system. VideoWHO is described in Figure 10, which have four major steps to do which is also described below.
**Shot Boundary Detection:** The undertaking of any shot boundary detection technique on video frames is to recognize the visual discontinuities along the time area. Throughout the course or duration of the identification process, it is significant to remove the visual components that measure the level of likeness between successive frames in a given shot. They find out the shots in the video via method described below

(a) **Color histogram:** The color histogram-based shot boundary location is a standout amongst the most dependable variations of histogram-based detection algorithm (Le, 2006.). It considers that color content does not change quickly inside, but rather crosswise over shots. In this way, shot cuts furthermore steady moves, can be recognized as single tops in the time arrangement of the contrasts between color histograms of nonstop edges. Frequently, computerized images are spoken to in RGB color space, so it is better to utilize \(2^24\) bits/pixel images (8 bits for each color channel). The general number of conceivable hues levels is \(2^24\) bins. Because of the restricted reaction of human visual framework, we are not ready to recognize the entire levels of conceivable hues. A basic arrangement considers just the most noteworthy bits of each RGB segment (Fig2). This arrangement decreases computational overhead and expansions strength to-ward straight forward camera and item movement.

(b) **Threshold determination:** The shot Boundary Detection technique depends on the difference between histograms of edges fitting in with a video sequence. This difference is calculated utilizing Manhattan distance method \(M\) or Euclidean Distance \(E\), respectively.

\[
M(h_i - h_{i-1}) = \sum_{j=1}^{M} |h_i(j) - h_{i-1}(j)|
\]

\[
E(h_i - h_{i-1}) = \frac{\sum_{j=1}^{M} (h_i(j) - h_{i-1}(j))^2}{\sqrt{\sum_{j=1}^{M} h_i(j)^2}}
\]

Where \(h_i\) and \(h_{(i-1)}\) are the histograms for frame \(f_i\) and \(f_{(i-1)}\). This strategy utilizes Manhattan separation for color histograms and Euclidian distance for edge heading histogram.

Having these qualities, the measures can be computed. In our work we utilized the accompanying:

\[
\text{Precision} = \frac{N_t - N_d}{(N_t - N_d) + N_i}
\]

\[
\text{Recall} = \frac{N_t - N_d}{N_t}
\]

The review measure takes percentage real abrupt shots that has been recognized by the strategy, while the exactness measure is a rate demonstrating how precise the technique is at identifying just the genuine shot boundary.

**Face detection and tracking:** The human face detection and tracking step is first and foremost step of extracting the faces from the video. Keeping the end goal in the mind, viewing the event of a man from the starting point until he/she leaves the video frame. Face detection algorithm is applied on video frame by frame. The face detection and tracking algorithm which the authors in VideoWHO used is shown in Figure 11.

Fig 11. Reprinted from (Tong Zhanga, 2013)

Towards the starting, the face detection algorithm is put on frame by frame, attempting to recognize the front or close front face areas utilizing the Adaboost algorithm with Haar-like features. Once the face is recognized, another task is assigned to it and tracking procedure start on the same face. In following process, the new appearance of face in the frame is demonstrated utilizing the particle feature algorithm with HOG features. The face detection algorithm is called every M outlines (M ranges from 15000) so as to upgrade the face. In the event the face area is detected and tracked until and unless the face is lost from the consecutive video frames.

**Face feature and constraint extraction:** The information collected by step 2 contains the face images with different varieties and poses, and obviously are not suitable for clustering. Thus in this phase now representative faces from the sequences are selected and their features are extracted. As a result, feature vector for every face is created. In the clustering process the endeavor is to recognize all the face successions of the same individual and attempt to minimize the stance,
brightening and impediment varieties when directing the detection and recognition. Along these lines, there are vast majority of faces which are not suitable for measuring the face similarity between face sequences. It is important to choose representative faces from the sequence which have high quality and comparable postures. To achieve this, testing is done in the uncontrolled video, they propose a feature to assess the nature of the face images and select the representative faces among them for the clustering. Every appeared face is appraised by angles, size and eye confinement certainty. The faces which have almost front and close view and high eye recognition certainty esteem (over a threshold TE) and sensible bigger size, that is bigger than 200 pixels in zone, will be chosen for clustering. We call these type of face images as representative faces, which is blend of the Gabor feature and LBP features, with the measurement lessened through both PCA and LDA. At that point similarity matrix is acquired by calculating the distance between each pair of representative facial feature vector so that everything in the matrix results the comparability between couple of faces.

**Semi-supervised agglomerative clustering:** With the help of the resulted closeness matrix, the face tests having the place with the common people are assembled. As the quantity of faces in the video is large so the quantity of clusters is also obscure, so the clustering of face should be done in an agglomerative way. The algorithm of clustering used in this paper depends on the algorithm of the author’s past work on face based image clustering and they extend it to the video clustering issue in this paper. It begins with the allotment of clusters, and then the two comparable clusters are merged into one cluster. This process continues until the likeness esteem between two nearest neighbor groups falls beneath a halting edge or threshold. The halting edge controls the quantity of clusters got. With this technique, it is to ensure that each cluster must contain the faces of same individual. However, it is conceivable that one individual has different clusters. The last step of clustering procedure is a rundown of clustering with the complexity of 3D model is high.

The future application for this video retrieval or face recognition frameworks can be retailing, retail stores, restaurants, Movie Theater etc.

**CONCLUSION**

Rethinking the pictures in Fig. 1 in the Introduction, we now have, available to us, various face acknowledgment algorithm that can serenely handle these impressive looking images. Scarcely 10 years back, would these images have been risky for face recognition algorithm of the time. The new ideas and bits of knowledge presented in examining enlightenment displaying in the previous decade have more numerous natural products as face acknowledgment calculations that are powerful against brightening variety. From multiple points of view, we are extremely lucky in light of the fact that human faces don’t have more mind boggling geometry and reflectance. Combined with the superposition way of brightening, this permits us to use low-dimensional straight appearance models to catch an extensive parcel of image variety because of enlightenment. Linearity makes the algorithms effective and simple to execute, and the appearance models make the calculations strong. Yet the speculation to posture variety exhibited for the brightening cone technique does not scale well as the quantity of subjects increments since the test picture should be contrasted with the representation of each enlisted subject. One approach would be to utilize the generative models to make a discriminative classifier.

A part of a full recognition framework is powerful face recognition and arrangement over an extensive variety of brightening and stance varieties. Present face recognition and detection systems are not as powerful over the same scope of conditions as the introduced acknowledgment strategies. Since face tracking is an essential and key
some portion of video face recognition techniques, it is additionally a testing issue to add to a tracker that is vigorous against stance and light varieties. Expression, incomplete impediment, cosmetics, maturing, and other figures should likewise be considered working together with work on brightening and stance.

REFERENCES


J.SUNEETHA (2014), A survey on Video-based Face Recognition Approaches, ISSN 2319 - 4847

Le, D.-J. (2006.) human face processing techniques with application to large scale video indexing. Addison-Wesley.


Received on 10th Sep 2015 Accepted on 20th Nov 2015

Manuscript can be seen online at www.luawmd.edu.pk