On Recognizing Faces in Videos using Clustering Based Re-ranking and Fusion

Himanshu S. Bhatt, *Student Member, IEEE,* Richa Singh, *Member, IEEE,* Mayank Vatsa, *Member, IEEE.*

***Abstract*—Due to widespread applications, availability of large intra-personal variations in video and limited information content in still images, video based face recognition has gained significant attention. Unlike still face images, videos provide abundant information that can be leveraged to address variations in pose, illumination, and expression as well as enhance the face recognition performance. This paper presents a video based face recognition algorithm that computes a discriminative video signature as an ordered list of still face images from a large dictionary. A three stage approach is proposed for optimizing ranked lists across multiple video frames and fusing them into a single composite ordered list to compute the video signature. This signature embeds diverse intra-personal variations and facilitates in matching two videos with large variations. For matching two videos, a discounted cumulative gain measure is utilized which uses the ranking of images in the video signature as well as the usefulness of images in characterizing the individual in a video. The efficacy of the proposed algorithm is evaluated under different video based face recognition scenarios such as matching still face images with videos and matching videos with videos. The efficacy of the proposed algorithm is demonstrated on the YouTube faces database and the MBGC v2 video challenge database that comprise different types of video based face recognition challenges such as matching still face images with videos and matching videos with videos. Performance comparison with the benchmark results on both the databases and a commercial face recognition system shows the efficiency of the proposed algorithm for video based face recognition.**

I. Introduction

With the increase in usage of camera technology in both surveillance and personal applications, enormous amount of video feed is being captured everyday. For instance, almost 100 hours of video are being uploaded every minute on YouTube alone[1] and it is increasing rapidly. Surveillance cameras are also capturing significant amount of data across the globe. In terms of face recognition, the amount of data collected by surveillance cameras every day is probably more than the size of all the publicly available face image databases combined. One primary purpose of collecting these data from surveillance cameras is to detect any unwanted activity during the act or at least enable to analyze the events and may be determine the persons of interest after the act. Therefore, widespread use of video cameras for surveillance and security applications have stirred extensive research interest in video based face recognition.

While face recognition is a well-studied problem and several algorithms have been proposed [27], [46], a majority of the literature is on matching still images and face recognition from videos is relatively less explored. Recognizing the individuals appearing in videos has both advantages and disadvantages compared to still face matching. Since the acquisition in videos is unconstrained, the presence of covariates such as pose, illumination, and expression is significantly more but at the same time, the information available in a video is generally more than the information available for matching two still images. As shown in Fig. 1, videos provide several cues in the form of multiple frames and temporal information as compared to still images. These cues can be used for improving the performance of face recognition and provide robustness to large variations in facial pose, expression, and lighting conditions.

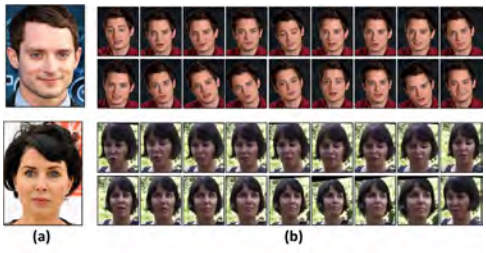

Fig. 1. Illustrates the abundant information present in videos. Compared to (a) still face images, (b) video frames represent large intra-personal and temporal variations useful for face recognition.

Video based face recognition includes (1) matching video-to-still face images (or still-to-videos) and (2) matching two videos. In video-to-still face recognition, the probe (query) is a video sequence and the gallery is composed of still face images whereas in still-to-video face matching, the gallery and probe are switched. As proposed by Zhang *et al.* [45], video-to-still/still-to-video face recognition techniques can be broadly categorized into frame selection and multi-frame fusion approaches. In frame selection, one or more optimal frames are selected from a video sequence and used to compute the similarity between the video and still images. On the other hand, in multi-frame fusion approaches, recognition results of multiple frames are fused together. In video-to-video face recognition, both gallery and probe (query) are videos of individuals to be matched. Poh *et al.* [34] evaluated several existing approaches for video-to-video face recognition and their analysis suggests that existing techniques do not efficiently utilize the abundant information in videos for enhancing face recognition performance. They also suggest that (1) part-based approaches generally out-perform holistic approaches

H.S. Bhatt, R. Singh, and M. Vatsa are with the Indraprastha Institute of Information Technology (IIIT) Delhi, India, e-mail: {himanshub, rsingh, mayank}@iiitd.ac.in.

[1]http://www.youtube.com/yt/press/statistics.html

and (2) selecting frames based on the image quality improves the recognition performance. To further evaluate existing algorithms for video-to-video face recognition, the Multiple Biometric Grand Challenge (MBGC) [33] also featured a problem on face recognition from unconstrained videos. The results from the challenge suggest that there is a huge gap in the performance of state-of-the-art algorithms from still image to video based face recognition. Observations and analysis from these evaluations elicit further research in video based face recognition.

A. Related Research

The survey on video based face recognition by Barr *et al.* [4] categorizes existing approaches as set based and sequence based approaches. Table I summarizes the existing video based face recognition algorithms. Set based approaches [40], [42] utilize the abundance and variety of observations in a video to achieve resilience to sub-optimal capture conditions. The approaches [2], [38] that model image sets as distributions use the between-distribution similarity to match two image sets. However, the performance of such approaches depends on the parameter estimation of the underlying distribution. Modeling image sets as linear sub-spaces [1], [12], [31] and manifolds [2], [15], [16], [41] is also proposed where matching between two image sets is performed by measuring similarity between the input and reference subspaces/manifolds. However, the performance of a subspace/manifold based approach depends on maintaining the image set correspondences. To address these limitations, Cui *et al.* [11] proposed to align two image sets using a common reference set before matching. Lee *et al.* [24] proposed a connected manifold approach that utilizes the likelihood and transition probability of the nearest previous manifold for recognition. Hu *et al.* [18] proposed to represent an image set using sample images, their mean, and an affine hull model. A sparse approximated nearest point method was proposed to compute the between-set distance as a pair of nearest points on the sets that are sparsely approximated by sample images. On the other hand, sequence based approaches explicitly utilize the temporal information for improved face recognition. To utilize the temporal information, Zhou *et al.* [47] proposed to use a joint posterior probability distribution of motion vector and identity variable estimated using sequence importance sampling. Several approaches that model the temporal information with Hidden Markov Models (HMM) [21], [28] are also proposed for improving video based face recognition.

Recently, the research focus has shifted and advancements in face recognition have led to a new paradigm of matching face images using a large dictionary. Patel *et al.* [32] proposed a sparse approximation based approach where test images were projected onto a span of elements in learned dictionaries and the resulting residual vectors were used for classification. Chen *et al.* [8] proposed a generative approach for video based face recognition where a video sequence was first partitioned into sub-sequences and then sequence-specific dictionaries were learned. The frames from every query video were projected onto the span of atoms in every sequence-specific dictionary and the residuals were utilized for recognition. Their approach has a computational overhead of creating multiple sequence-specific dictionaries for specific pose and illumination variations. Chen *et al.* [9] proposed a multi-variate sparse representation that simultaneously takes correlation as well as coupling information between frames. Different sub-directories were trained for multiple partitions which represents a particular viewing condition and a joint sparse representation was used for face recognition using minimum class reconstruction error criteria. Recently, Bhatt *et al.* [7] proposed to compute a video signature as an ordered list of still face images from a large dictionary. In their approach, temporal and wide intra-personal variations from multiple frames were combined using Markov chain based rank aggregation approach.

B. Research Contributions

This research proposes a novel algorithm for video based face recognition that computes the signature of a video as an ordered list of still face images from a dictionary. Fig. 2 shows the outline of the proposed algorithm which starts by computing a ranked list for every frame in the video to utilize the abundant information and capture the wide intra-personal variations. It utilizes the taxonomy of facial features [23] to efficiently compute the video signature. Taxonomy of facial features [23] groups the salient information available in face images into different feature categories: level-1, level-2, and level-3. Out of these three, level-1 features capture the holistic nature of face such as skin color, gender, and facial appearance. These features are highly discriminative in differentiating an image from other images that have largely different facial appearances. These features being computationally efficient, are generally used for indexing or reducing the search space. Therefore, level-1 features are used to generate a ranked list by congregating images from the dictionary that are similar to the input frame. A ranked list is an ordered list of face images retrieved from the dictionary where the face image with the highest similarity is positioned at the top of the list. To characterize an individual in a video, ranked lists from multiple frames are combined using a three stage process that involves clustering, re-ranking, and fusion. It produces the final composite ranked list for a video which represents the discriminative video signature. Combining multiple ranked lists into a single optimized ranked list that minimizes the overall distance from all ranked lists is a well studied problem in information retrieval domain. However, to the best of our knowledge, this paper presents the first approach to combine ranked lists pertaining to individual frames to generate a composite video signature. It transforms the problem of video based face recognition into matching two ordered lists (ranked lists). Further, a relevance score is computed for images in the final composite ranked list using the discriminative level-2 features. These are locally derived features and describe structures in a face that are pertinent for face recognition. As compared to level-1 features, these features are more discriminative and are predominantly used for face recognition. Relevance scores are computed using level-2 features that capture the discriminative

Category	Authors	Technique	Database	Recognition Rate (%)
Set Based	Arandjelovic *et al.* [2]	Manifold density divergence	Private	93.6 (avg)
	Wang *et al.* [41]	Manifold-manifold distance	Honda/UCSD [25]	96.9
			CMU MoBo [14]	93.6
	Aggarwal *et al.* [1]	Linear dynamic modeling	Private	93.7
			Honda/UCSD [25]	90.0
	Fukui & Yamaguchi [12]	Kernel orthogonal mutual subspace	Private	97.42/EER=3.5
	Nishiyama *et al.* [31]	Hierarchical image-set matching	Private	97.4/EER=2.3
	Harandi *et al.* [16]	Grassmannian manifolds	CMU PIE [39]	65.2
			BANCA [3]	64.5
			CMU MoBo [14]	64.9
	Cui *et al.* [11]	Image set alignmnet	Honda/UCSD [25]	98.9
			CMU MoBo [14]	95.0
			YouTube celebrity [21]	74.6
	Lee *et al.* [24]	Probabilistic appearance manifolds	Private	93.2
	Hu *et al.* [18]	Sparse Approximated Nearest Point	Honda UCSD [25]	92.3
			CMU MoBo [14]	97
			YouTube Celebrity [21]	65.0
	Wolf *et al.* [42]	Set-to-set similarity	YouTube Faces [42]	72.6 verification accuracy at EER
Sequence Based	Zhou *et al.* [47]	Sequential importance sampling	Private	100
			Private	~93
			CMU MoBo [14]	~56
	Liu & Chen [28]	Adaptive Hidden Markov models	Private	1.2 EER
			CMU MoBo [14]	4.0 EER
	Kim *et al.* [21]	Visual constraints using generative and discriminative models	Honda/UCSD [25]	100
			YouTube celebrity [21]	~70
Dictionary Based	Chen *et al.* [8]	Video-dictionaries	MBGC v1 [33]	~59 verification accuracy at EER (WW)
				~55 verification accuracy at EER (AW)
				~51 verification accuracy at EER (AA)
	Bhatt *et al.* [7]	Rank aggregation	YouTube Faces [42]	78.3 verification accuracy at EER
	Proposed	Clustering based re-ranking and fusion	YouTube Faces [42]	80.7 verification accuracy at EER
			MBGC v2 [33]	62.2 verification accuracy at EER (WW)
				57.3 verification accuracy at EER (AW)
				54.1 verification accuracy at EER (AA)

TABLE I
CATEGORIZATION OF EXISTING APPROACHES OF VIDEO BASED FACE RECOGNITION.

power of an image in characterizing the individual in a video. Finally, to match two videos, their composite ranked lists (video signatures) are compared using a discounted cumulative gain (DCG) measure [20]. The major contributions of this research can be summarized as follows:

- It utilizes the taxonomy of facial features for efficient video based face recognition. Computationally efficient level-1 features are used for computing multiple ranked lists pertaining to multiple video frames and discriminative level-2 features are used to compute the relevance of images in the final composite ranked list.
- Existing dictionary based face recognition algorithms [37] compute the signature of a still face image as an ordered list of images from dictionary. In this research, a new paradigm is introduced using a three-stage technique for generating video signatures as an ordered list of still face images from the dictionary.
- Existing approaches discard the characteristics embedded in the ranked lists and only consider the overlap between two lists as the final similarity. In this research, the DCG measure seamlessly utilizes rank and relevance scores of images to compute the final similarity between two lists.

II. DICTIONARY BASED VIDEO FACE RECOGNITION

Recent studies in face recognition [32], [37], [43] have shown that generating image signatures based on a dictionary is more efficient for matching images across large variations than direct comparison between two images or some of its features. In this research, video based face recognition is addressed by computing a discriminative video signature using a dictionary of still face images. The proposed algorithm congregates abundant information present in multiple video frames to generate a discriminative video signature. It facilitates in characterizing an individual as it embeds the information in

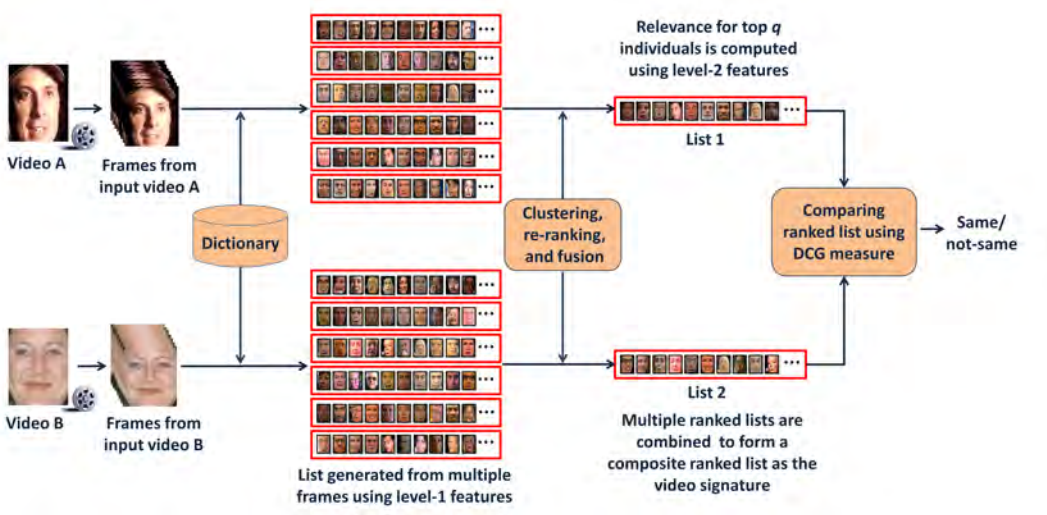

Fig. 2. Illustrates the block diagram of the proposed algorithm for matching two videos.

the form of a ranked list of images under similar intra-personal settings from the dictionary. Fig. 2 shows different stages of the proposed algorithm which are elaborated in the following subsections.

A. Dictionary

Dictionary is a large collection of still face images where every individual has multiple images capturing a wide range of intra-personal variations i.e. pose, illumination, and expression variations. Our definition of dictionary is different from the dictionary in sparse representation based approaches [8], [18]. They represent a dictionary as a collection of atoms such that the number of atoms exceeds the dimension of the signal space, so that any signal can be represented by more than one combination of different atoms. In this research, the dictionary comprises $38,488$ face images pertaining to 337 individuals from the CMU Multi-PIE [13] database. OpenCV's boosted cascade of Haar-like features provide the face boundaries and eye-coordinates. These boundaries are used to detect and crop faces from the dictionary images and eye-coordinates are used to normalize the detected image with respect to rotation. The normalized face images are resized to 196×224 pixels with inter-eye distance of 100 pixels.

B. Computing Ranked List

Let V be the video of an individual comprising n frames where each frame depicts the temporal variations of the individual. Face region from each frame is detected[2] and pre-processed[3]. Face regions corresponding to different frames across a video are represented as $\{F_1, F_2, ..., F_n\}$. To generate ranked lists, each frame is compared with all the images in the dictionary. Since the dictionary consists of a large number of images and each video has multiple frames; it is essential to compute the ranked list in a computationally efficient manner. Linear discriminant analysis (LDA), a level-1 feature, is therefore used to generate a ranked list by congregating images from the dictionary that are similar to the input frame. A linear discriminant function [5] is learnt from the dictionary images that captures the variations in pose, illumination, and expression. The linear discriminant function learns these variations and retrieves images from the dictionary that are similar to the input video frame i.e. images with similar pose, illumination, and expression. The ranking of retrieved images from such a dictionary is found to be more discriminative for face recognition than that of a signature based on the pixel intensities or some image features [37]. Each column of the projection matrix W represents a projection direction in the subspace and the projection of an image onto the subspace is computed as:

$$Y = W^T X \tag{1}$$

where X is the input image and Y is its subspace representation. The input frame F_i and all images in the dictionary are projected onto the subspace. The Euclidean distance is computed between the subspace representations of the input frame F_i and each of the dictionary images. An ordered list of images is retrieved from the dictionary based on their similarity[4] to the input frame. To generate a ranked list $\mathbf{R}_i$ corresponding to the input frame F_i, retrieved dictionary images are positioned based on their similarity to F_i with the most similar image positioned at the top of the list. For a video V, the proposed algorithm computes a set of ranked list $\{\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_n\}$ corresponding to the n frames of the video.

[2]OpenCV's boosted cascade of haar-like features is used for face detection in near-frontal videos. For profile-face videos, a tracking technique [36] is used to track and detect faces. The detected faces from videos are extracted by a combination of automatic and manual tasks where the tracker for the face region in the first frame each time is located.

[3]A multi-scale retinex with wavelet based de-noising technique [6] is utilized to enhance the quality of poor quality video frames before computing the ranked list.

[4]The distance scores computed using level-1 features are normalized in range {0-1} using min-max normalization and then converted into similarity scores.

C. Clustering, Re-ranking, and Fusion

Multiple ranked lists computed across n frames of a video have significant amount of overlap in terms of positioning of the dictionary images. Due to this redundant information, it is computationally expensive and inefficient to compare multiple ranked lists across two videos. Therefore, multiple ranked lists of a video are combined to generate a single composite ranked list, denoted as $\mathbf{R}'$. As shown in Fig. 3, the proposed algorithm generates a composite ranked list in three steps. First, each ranked list corresponding to a video frame is partitioned into different clusters and reliability of each cluster is calculated. Secondly, the similarity scores of images within a cluster are adjusted based on the reliability of that cluster [44]. Finally, multiple ranked lists of a video are fused based on the adjusted similarity scores of images to generate a composite ranked list as the video signature. The video signature thus obtained minimizes the distance from all the constituent ranked lists. These stages are described in Algorithm 1 and are elaborated below.

Algorithm 1 Fusing ranked lists with clustering and re-ranking.

Input: A set of ranked lists $\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_n$ from multiple frames in a video V.
Iterate: i= 1 **to** n (number of ranked lists)
Clustering: Partition ranked list $\mathbf{R}_i$ into different clusters $C_{i,1}, C_{i,2}, ..., C_{i,k}$, where k is the number of clusters.
end iterate.
Iterate: i= 1 **to** n, j= 1 **to** k.
Reliability: Compute reliability of cluster $r(C_{i,j})$.
Re-ranking: Adjust the similarity score of each image d based on the reliability of the cluster it belongs.
$Sim_i^*(d)$=$Sim_i(d) \times (1 + r(C_{i,j}))$, $d \in C_{i,j}$.
end iterate.
Fusion: Compute an ordered composite ranked list $\mathbf{R}'$ where similarity score of an image d is given as:
SS_d= $\frac{\sum_{i=1}^{n} Sim_i^*(d)}{n}$.
Output: Final composite ranked list $\mathbf{R}'$ for video V.

1) Clustering: Multiple frames in a video exhibit different intra-personal variations; therefore, each ranked list positions dictionary images based on the similarity to the input frame. Images in the ranked list are further partitioned into different clusters such that if an image in a cluster has high similarity to the input frame, then all images in that cluster tend to be more similar to the input frame. The main idea behind clustering is to congregate images in a ranked list into different clusters where each cluster represents a particular viewing condition i.e. a specific pose, illumination or expression. Let $\mathbf{R}_i$ be the i^{th} ranked list of a video corresponding to frame F_i, then $\{C_{i,1}, C_{i,2}, ..., C_{i,k}\}$ form k clusters of $\mathbf{R}_i$. In this research, K-means clustering [17] which is an unsupervised, non-deterministic technique for generating a number of disjoint and flat (non-hierarchical) clusters is used to cluster similar images with an equal cardinality constraint. To guarantee that all clusters have equal number of data points, k centroids are initially selected at random. For each point, similarity to the nearest cluster is computed and a heap is build. Similarity is measured using the Euclidean distance in LDA projection space, as described in Eq. 1. A data point is drawn from the heap and assigned to the nearest cluster, unless that cluster is already full. If the nearest cluster is full, distance to the next nearest cluster is computed and the data is re-inserted into the heap. The process is repeated till the heap is empty i.e. all the data points are assigned to a cluster. It guarantees that all the clusters contain equal number of data points (± 1 data points per cluster). K-means clustering is used as it is computationally faster and produces tighter clusters than hierarchical clustering techniques. After clustering, each ranked list $\mathbf{R}_i$ has a set of clusters $C_{i,1}$, $C_{i,2}$,..., $C_{i,k}$, where k is the number of clusters. K-means clustering is affected by the initialization of initial centroid points; however, we start with five different random initializations of k clusters. Finally, clusters which minimize the overall sum of square distances are selected.

2) Re-ranking: Clusters across multiple ranked lists overlap in terms of common dictionary images. Since the overlap between the clusters depends on the size of each cluster, it is required that all the clusters should be of equal size. Higher the overlap between the clusters, more likely that they contain images with similar appearances (i.e. with similar pose, illumination, and expression). Based on this hypothesis, the reliability of each cluster is computed as the weighted sum of similarities between the cluster and other clusters across multiple ranked lists [44]. The $reliability$ $r(C_{l,j})$ of a cluster $C_{l,j}$ in ranked list q is computed as shown in Eq. 2.

$$r(C_{l,j}) = \sum_{i=1, i\neq l}^{n} \sum_{p=1}^{k} \left[\frac{Sim_F C(F_i, C_{i,p})}{norm_l} Sim(C_{l,j}, C_{i,p}) \right] \quad (2)$$

where

$$norm_l = \sum_{i=1, i\neq A}^{n} \sum_{p=1}^{k} [Sim_F C(F_i, C_{i,p})] \quad (3)$$

$$Sim_F C(F_l, C_{l,j}) = \frac{\sum_{d \in C_{l,j}} ||F_l - d||^2}{|C_{l,j}|} \quad (4)$$

$$Sim(C_{l,j}, C_{m,j}) = |C_{l,j} \cap C_{m,j}| \quad (5)$$

where d is an image from the dictionary, $norm_l$ is a normalization factor for clusters in the ranked list $\mathbf{R}_l$, $|C_{l,j}|$ is the number of images in cluster $C_{l,j}$, F_l is the current frame of the video, and $||F_l - d||^2$ represents the similarity between the input frame and a dictionary image computed using the Euclidean distance[5] between their subspace representations. The similarity between frame F_i and cluster $C_{i,j}$ is measured as the average similarity score of all images in that cluster to the input frame F_i, as shown in Eq. 4. The similarity between two clusters is measured by the number of common images as shown in Eq. 5. If the reliability of a cluster is large, the images in the cluster have high contribution towards the overall similarity scores. Therefore, the similarity scores of images in a cluster are adjusted based on the reliability of the cluster.

[5]The distance scores computed using level-1 features are normalized in the range {0-1} using min-max normalization and then converted into similarity scores.

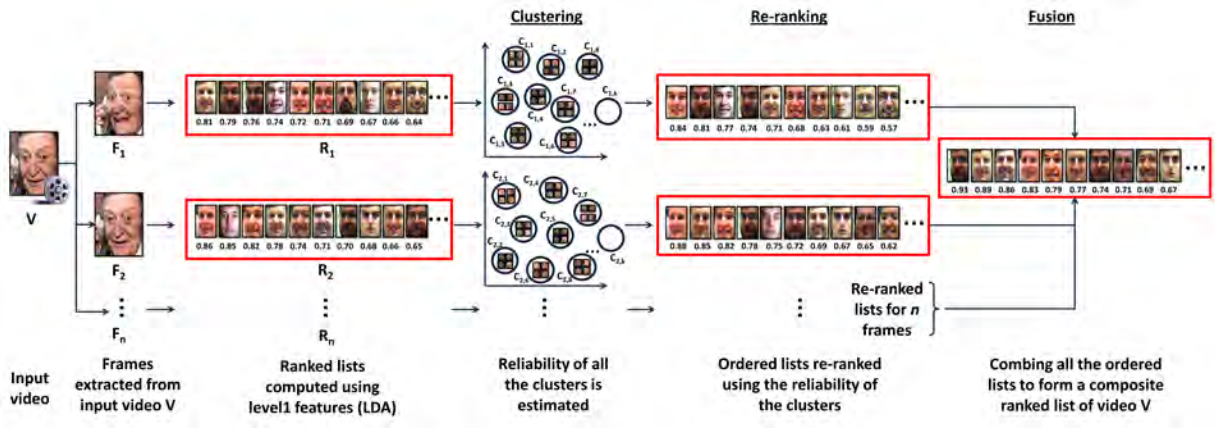

Fig. 3. Illustrates clustering based re-ranking and fusion to form the video signature. Clustering based re-ranking associates dictionary images to different clusters and adjusts their similarity scores. It facilitates to bring images similar to the query frame towards the top of the ranked list. The lists are then re-ranked using the adjusted scores and are finally combined to generate the video signature.

It enhances the similarity scores of images from a cluster that exhibits similar settings as the input video frame and reduces the similarity scores of images from clusters which exhibit different settings i.e. pose, illumination, and expression variations. The reliability score of a cluster is then used to adjust the similarity scores of all images belonging to that cluster, as shown in Eq. 6.

$$Sim_i^*(d) = Sim_i(d) \times [1 + r(C_{i,j})], \forall\ d \in C_{i,j} \qquad (6)$$

$Sim_i(d)$ is the similarity score of an image d in ranked list $\mathbf{R_i}$ computed using level-1 features and $r(C_{i,j})$ is the reliability of the j^{th} cluster of the i^{th} ranked list, $C_{i,j}$, such that $d \in C_{i,j}$.

3) Fusion: The ranked lists across multiple frames have redundant information and matching such ranked lists across two videos can be computationally inefficient. Therefore, it is imperative to compute a composite ranked list as the video signature. Once the similarity scores of images are adjusted across all the ranked lists, multiple ranked lists are fused into a final composite ranked list, R'. The final similarity score of an image d (denoted as SS_d) is the average of adjusted similarity scores of image d across all the ranked lists, as shown in Eq. 7.

$$SS_d = \frac{\sum_{i=1}^{n} Sim_i^*(d)}{n} \qquad (7)$$

where n is the number of frames in a video. There are different types of fusion methods proposed in the literature [22], [35] such as sensor level, feature level, score level, and decision level fusion. Chen *et al.* [9] proposed to concatenate n sub-dictionaries using a joint sparsity coefficient approach to make a combined decision. However, in the proposed algorithm, adjusted similarity scores of all images in the dictionary are averaged across multiple ranked lists. The final composite ranked list $\mathbf{R}'$ of a video is generated by ordering all the images in dictionary such that the image with maximum adjusted similarity score (SS) is positioned at the top of the list.

D. Matching the Composite Ranked Lists

To match two videos, their composite ranked lists obtained after clustering based re-ranking and fusion are compared. The discounted cumulative gain (DCG) [20] measure is used to compare two ranked lists. DCG measure is widely used in information retrieval domain [29] to compare the lists of documents. Each document in the ranked list is arranged based on its similarity to the input query and also has a relevance score provided by a domain expert (or the user). It uses both these attributes (i.e. rank and relevance) to compare two ranked lists. The relevance in our context is the usefulness of a dictionary image in characterizing the individual in a video. The relevance rel_d of a dictionary image d is computed as the maximum similarity score of the image across multiple frames of the video, as shown in Eq. 8.

$$rel_d = \arg\max_{1 \le i \le n} \{Sim_{level2}(d, F_i)\} \qquad (8)$$

where n is the number of frames in a video, $Sim_{level2}(d, F_i)$ is the similarity score of a dictionary image d with the frame F_i computed using level-2 features (LBP) and χ^2 distance measure. It is observed that the similarity between a video frame and images in the ranked list drop after a particular rank and the order of images is less discriminative beyond that point. Therefore, images retrieved till rank q are considered in the video signature and their relevance is computed. Now, the images in the composite ranked list $\mathbf{R}'$ are positioned based on level-1 features and have a relevance score computed using level-2 features.

The DCG measure is based on the observation that relevant images are more useful when appearing earlier in the ranked list. If an image is ranked farther from the top of the list, its contribution towards the overall similarity score is low. Similarity is accumulated from top of the ranked list to the bottom by discounting the relevance score of an image by its position. Therefore, DCG measure is more efficient in matching two ranked list than just comparing the overlap between two lists (later shown in results). As shown in Eq. 9, DCG measure discounts the relevance of an image by the

logarithm of its rank.

$$DCG_q = \sum_{i=1}^{<b} rel_i + \sum_{i=b}^{q} \frac{rel_i}{log_b(i)} \tag{9}$$

where rel_i is the relevance score of an image at rank i and the DCG is computed till rank q. In our experiments, $q = 100$ and logarithm to the base $b = 2$ are empirically set to yield the best performance. Further, the DCG value is normalized by dividing it with ideal discounted cumulative gain ($IDCG$) to obtain normalized discounted cumulative gain $nDCG$, as shown in Eq. 10.

$$nDCG_q = \frac{DCG_q}{IDCG_q} \tag{10}$$

$IDCG$ at rank q is obtained by calculating DCG values when the images in the ranked list are positioned based on their relevance instead of similarity scores computed using level-1 features (i.e. the image with maximum relevance is positioned at the top of the list). To compute the similarity between two ranked lists, a two sided $nDCG$ measure is used. For two ranked lists $\mathbf{R'_1}$ and $\mathbf{R'_2}$, $nDCG_q$ for $\mathbf{R'_1}$ with respect to $\mathbf{R'_2}$ at rank q is computed by considering $\mathbf{R'_2}$ as the ideal ranking of images. Similarly, $nDCG_q$ for $\mathbf{R'_2}$ with respect to $\mathbf{R'_1}$ is computed by considering $\mathbf{R'_1}$ as the ideal ranking of images. The final similarity K_{sim} between two lists $\mathbf{R'_1}$ and $\mathbf{R'_2}$ is the average of the two $nDCG$ values.

$$K_{sim}(R'_1, R'_2) = \frac{1}{2}\{nDCG_q(\mathbf{R'_1}, \mathbf{R'_2}) + nDCG_q(\mathbf{R'_2}, \mathbf{R'_1})\} \tag{11}$$

E. Dictionary Based Video Face Recognition Algorithm

The proposed algorithm for computing the video signatures and matching is summarized below:

Algorithm 2 Summarization of the proposed dictionary based video face recognition algorithm.

Step-1: For a given video pair, frames from each video are extracted and pre-processed. Face region from each frame is detected and resized to 196×224 pixels.

Step-2: For each frame in the video, a ranked list of still face images from the dictionary is computed using level-1 features. The retrieved dictionary images are arranged in a ranked list such that the image with the maximum similarity score is positioned at the top of the list.

Step-3: Ranked list across multiple frames of a video are combined to form a video signature using clustering based re-ranking and fusion as elaborated in Algorithm 1.

Step-4: To match two videos, their video signatures are compared using the $nDCG$ measure that incorporates scores computed using both level-1 (rank) and level-2 (relevance) features.

The proposed video based face recognition algorithm efficiently computes the video signature and transforms the problem of video based face recognition into matching two ranked lists. Generally, in face recognition applications, level-1 and level-2 features are sufficient for efficiently matching face images. In some law enforcement applications such as matching identical twins or look-alikes, level-3 features are widely used as an additional layer of discrimination over level-1 and level-2 features. However, level-3 features are extracted from good quality high resolution face images which are generally not available in the application focused in this paper i.e. face recognition from unconstrained videos. Therefore, only level-1 and level-2 features are used in this research for computing a discriminative video signature.

III. Experimental Results

The efficacy of the proposed algorithm is evaluated on multiple databases under different scenarios such as video-to-still, still-to-video, and video-to-video. For a thorough analysis, the performance of individual components of the proposed algorithm is evaluated along with comparing it with the min-max normalization and sum rule fusion [35], referred to as MNF, for combining multiple ranked lists across the video frames. The performance is also compared with FaceVACS which is a commercial off-the-shelf face recognition system (denoted as COTS). Section III-A explains the databases used in this research, Section III-B elaborates the experimental protocol, and finally Section III-C lists the key observations and analysis.

A. Databases

The experiments are performed on two publicly available video databases: The YouTube faces database [42] and MBGC v2 video challenge database [33]. The YouTube faces database [42] is the largest available unconstrained video database comprising $3,425$ videos of $1,595$ different individuals downloaded from YouTube where each video has ~180 frames on average. The database provides 10-fold pair-wise matching ('same'/'not-same') test benchmark protocol for comparison with existing algorithms. $5,000$ video pairs are randomly selected from the database, half of which are pairs of videos of the same individual and half of different individuals. As per the given protocol [42], these pairs are further divided into 10 splits where each split contains 250 'same' and 250 'not-same' pairs. Further details about the database are available in [42].

The MBGC v2 video challenge database [33] comprises videos in standard (720×480) and high definition (1440×1080) formats pertaining to individuals either walking or performing some activity. From the MBGC v2 video challenge database, experiments are performed on the data collected from the University of Notre Dame. The experiments are performed for matching videos of individuals under three settings, 1) walking vs walking (WW), 2) walking vs activity (WA), and 3) activity vs activity (AA). Further, to evaluate the performance of the proposed algorithm for still-to-video and video-to-still matching, face images pertaining to 147 individuals from the MBGC v2 still portal are utilized which contains good quality still face images and their corresponding videos. Fig. 4 shows still face images along with samples from activity and walking video frames.

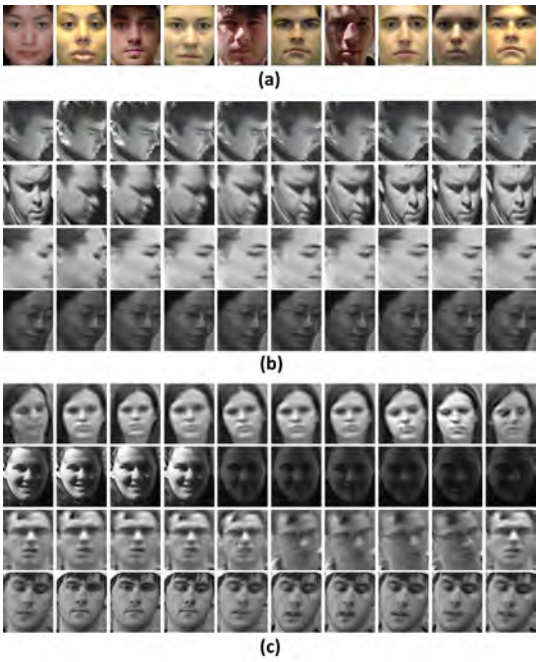

Fig. 4. Sample images from the MBGC v2 database (a) still face images, (b) frames from activity video, and (c) frames from walking video.

B. Protocol

The efficacy of the proposed algorithm for video based face recognition is evaluated in verification mode (1:1 matching). The performance of the proposed algorithm is compared with existing video based face recognition algorithms using the experimental protocol defined in [42] where the verification accuracy is reported at equal error rate (EER) along with area under the curve (AUC). For matching two videos using COTS, set-to-set matching is used where each frame in the first video is matched to all the frames in the second video. The mean score obtained corresponding to all the frames of the second video is assigned as the similarity score of the frame in the first video. The final similarity score of the first video is the average score of all the frames in that video. In MNF, similarity scores across multiple ranked lists are normalized using min-max score normalization [19]. The score for each dictionary image is then re-computed as the average score across all the ranked lists. Finally, the combined ranked list is generated based on the averaged similarity scores where the dictionary image with the largest similarity score is positioned at the top of the list. The experimental protocol for the two databases are further elaborated below:

1) YouTube Faces Database: The performance of the proposed algorithm is evaluated using the experimental protocol defined by Wolf *et al.* [42]. In this experiment both gallery and probe consist of videos and training is performed as two class problem with 'same'/'not-same' labels. In our experiments, ten splits provided along with the database are used. Training is performed on nine splits and the performance is computed on the tenth split. The final performance is reported as an average of 10 folds. In this protocol, the information about the subject's label associated with the video is discarded and only the information about whether a pair is 'same' or 'not-same' is retained.

On YouTube face database, the performance is compared with benchmark results of several algorithms already provided with the database [42] namely, LBP descriptor with matched background similarity (MBGS (mean) LBP), minimum distance (mindst LBP), maximum correlation measures ($||U1'U2||$ LBP) and FPLBP descriptor with matched background similarity (MBGS (mean) FPLBP), minimum distance (mindst FPLBP), maximum correlation measures ($||U1'U2||$ FPLBP), APEM+FUSION [26], STFRD+PMML [10], VSOF+OSS [30]. The performance is also compared with a commercial-off-the-shelf system and one recently proposed algorithm, referred to as Bhatt *et al.* [7].

The experiments are also performed to evaluate the performance enhancement due to different stages of the proposed algorithm on the YouTube faces database. Firstly, to evaluate the performance gain due to clustering based re-ranking and fusion steps, the performance is compared when ranked list across multiple frames are combined using the MNF approach. Secondly, to evaluate the effectiveness of the $nDCG$ measure, the performance is evaluated when two ranked lists are compared using the distance measure proposed by Schroff *et al.* [37]. Their distance measure only considers the overlap between two ranked lists and ignores other information such as relevance of images in the ranked list. It should be noted that while evaluating the gain in performance due to an individual step, all other steps in the proposed algorithm remain the same.

2) Multi Biometric Grand Challenge v2 Database: Multiple experiments are performed on this database to evaluate the efficacy of the proposed algorithm. Specifically, the algorithm is evaluated for two different scenarios: (1) matching still face images with videos and (2) matching videos with videos.

Matching still face images with videos: In many real world applications, such as surveillance, it is required to match still face images with videos for authenticating the identity of individuals. In this experiment, still face images from the MBGC v2 still portal and videos (comprising both walking and activity videos) from the MBGC v2 video challenge database [33] pertaining to 147 subjects are used. To evaluate the efficacy of the proposed algorithm, experiments are performed with 10 times repeated random sub-sampling (cross validations). In each experiment, training is performed on 47 subjects and the performance in reported on the remaining 100 subjects. This experiment further comprises two different subsets:

- *Matching video probe with still gallery images*: In this experiment, the probe is a video of an individual whose identity is to be matched against a gallery of still face images. The ranked list of an image in the gallery is computed by positioning the images retrieved from the dictionary based on their level-1 similarity scores. The composite ranked list of a probe video is then compared with the ranked list computed for each of the gallery images. The experiment is further divided as: 1) probe comprises 618 walking videos pertaining to 100 subjects and 2) probe comprises 513 activity videos pertaining to 100 subjects. In both the cases the gallery consists of 100 still face images, one image per subject.

- *Matching still probe with video gallery*: In this experiment, the probe is a still face image and the gallery comprises videos. The ranked list of a still probe image is compared with the composite ranked list of each video in the gallery. The experiment is divided as: 1) gallery comprises 100 walking videos and 2) gallery comprises 100 activity videos. In both the cases, the probe comprises 1543 still face images pertaining to 100 subjects.

Matching videos with videos: The proposed algorithm is evaluated for matching video-to-video face information where both gallery and probe comprise videos of individuals. The performance of the proposed algorithm on the MBGC v2 video challenge database is evaluated under three different scenarios, 1) walking vs walking (WW), 2) walking vs activity (WA), and 3) activity vs activity (AA). In the MBGC v2 video challenge protocol, verification experiments are specified by two sets: target and query. The protocol requires the algorithm to match each target (gallery) sequence with all the query (probe) sequences. In this experiment, the composite ranked list of a probe video is compared with the composite ranked lists of the gallery videos.

C. Results and Analysis

The proposed algorithm utilizes the observation that a discriminative video signature can be computed using a dictionary of still face images. Key results and observations from the experiments are summarized below:

- For both still images and videos, a dictionary of non-overlapping individuals is used to generate discriminative signatures represented as ranked lists of images. The results suggest that the representation based on dictionary is very efficient for matching individuals across large intra-personal variations in videos.

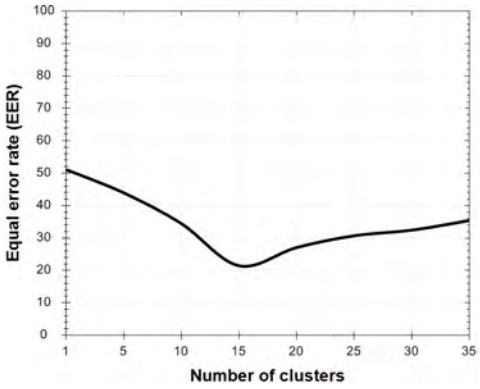

Fig. 5. Illustrates the variations in equal error rate by varying the number of clusters.

- Fig. 5 shows that the performance is dependent on the number of clusters. In our experiments, the number of clusters k is varied from 1 to 35. It is observed that all the variations in dictionary images can be broadly grouped into 15 different categories of pose, illumination, and expression. This observation also corroborates with our experiment to empirically determine the number of clusters as shown in Fig. 5 where $k = 15$ yields the lowest EER. If the number of clusters is less, images are not segregated in the cluster representing the exact viewing conditions. It results in erroneously updating the similarity scores of images based on the reliability of the cluster which increases the error rate. On the other hand, large number of clusters also increases the error rate and computational cost.
- The proposed algorithm utilizes the taxonomy of facial features to compute the initial ranked lists using computationally efficient level-1 features and more discriminative level-2 features to compute the relevance of images in the final composite ranked list. This selection of features for computing the ranked lists and relevance makes the proposed algorithm discriminative and computationally efficient.

Algorithm	**Verification Accuracy at EER (%)**	**SD (%)**	**AUC (%)**	**EER (%)**
mindst LBP	65.7	1.7	70.7	35.2
mindst FPLBP	65.6	1.8	70.0	35.6
$\|\|U1'U2\|\|$ LBP	65.4	2.0	69.8	36.0
$\|\|U1'U2\|\|$ FPLBP	64.3	1.6	69.4	35.8
MBGS (mean) FPLBP	72.6	2.0	80.1	27.7
MBGS (mean) LBP	76.4	1.8	82.6	25.3
APEM-FUSION [26]	79.1	1.5	86.6	21.4
STFRD+PMML [10]	79.5	2.5	86.6	19.9
VSOF+OSS [30]	79.7	1.8	89.4	20.0
Bhatt *et al.* [7]	78.3	1.7	85.8	21.6
COTS	67.9	2.3	74.1	33.1
MNF	76.4	2.1	81.6	24.3
Schroff *et al.* [37]	77.5	1.6	83.8	23.6
Proposed	**80.7**	**1.4**	**90.5**	**19.4**

TABLE II
COMPARING THE PROPOSED ALGORITHM WITH THE BENCHMARK TEST RESULTS AND COTS ON THE YOUTUBE FACES DATABASE [42].

1) Results on YouTube database:

- The results in Table II and receiver operating characteristic (ROC) curves in Fig. 6 demonstrate the performance of the proposed algorithm with benchmark results on the YouTube faces database [42]. The proposed algorithm outperforms existing algorithms and COTS for video-to-video face recognition. It achieves an average accuracy of 80.7% at EER of 19.4%. The proposed algorithm also achieves a higher area under the curve (AUC) of 90.5% as compared to other algorithms.
- To evaluate the gain in performance due to clustering, re-ranking, and fusion, the performance of the proposed algorithm is compared when multiple ranked lists are combined using min-max normalization and sum-rule fusion (referred to as MNF). Table II shows that clustering based re-ranking and fusion reduces the EER by ~9%. This gain can be attributed to the observation that images with similar appearances are clustered together and similarity scores of images are adjusted based on the reliability of the clusters.
- To match two video signatures, a two-sided $nDCG$ measure is used that seamlessly utilizes both level-1 (ranks) and level-2 (relevance) features. The performance gain due to two sided $nDCG$ measure is evaluated by

comparing the performance of the proposed algorithm when two signatures are matched using the similarity measure used by Schroff *et al.* [37]. Existing approaches only compute the overlap between two lists while discarding other information embedded in the lists, whereas, the results in Table II show that the two sided $nDCG$ measure reduces the EER by ~7%.

- Generally, existing approaches that use set-to-set similarities [1], [15], [40], [42] do not consider that multiple frames capture different intra-personal variations. Matching such diverse image sets independently leads to sub-optimal performance. However, the proposed algorithm combines the diverse information from multiple frames to form a composite video signature to match two videos. Fig. 7 shows some successful and unsuccessful verification results of the proposed algorithm.

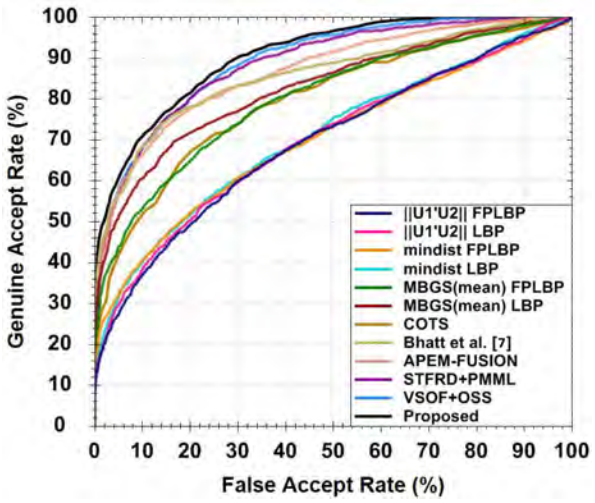

Fig. 6. ROC curves comparing the performance of the proposed algorithm with benchmark results on the YouTube faces database [42]. (Best viewed in color). The results from the YouTube database website are as of March 7, 2014.

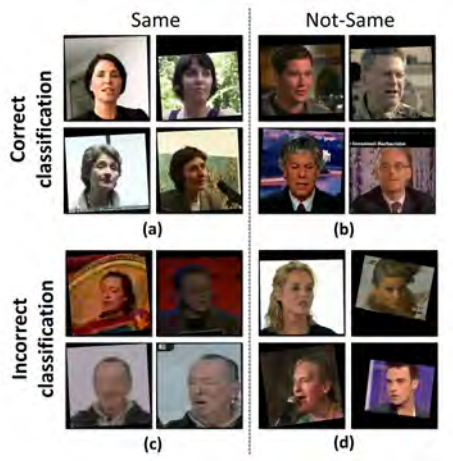

Fig. 7. Illustrating examples when the proposed algorithm correctly classified (a) 'same', (b) 'not-same' video pairs from the YouTube faces database [42]. Similarly, examples when the proposed algorithm incorrectly classified (c) 'same' and (d) 'not-same' video pairs.

- The proposed algorithm has different stages such as computing ranked lists for each video frame, clustering, re-ranking and fusion for combining multiple ranked lists into a discriminative video signature. Finally, two video signatures are matched using two sided $nDCG$ measure. The algorithm takes about 0.06 seconds to compute the ranked list for a single frame, 0.04 seconds to cluster a ranked list, 0.04 seconds for re-ranking the similarity scores within a ranked list. Further, for computing the signature for a video with 100 frames and fusing 100 ranked lists takes around 1.3 seconds. Therefore, total time to compute a composite ranked list for a video with 100 frames is $100 \times (0.06 + 0.04 + 0.04) + 1.3 = 15.3$ seconds. The time is reported on 2 GHz Intel Duo Core processor with 4 GB RAM under C# programming environment.

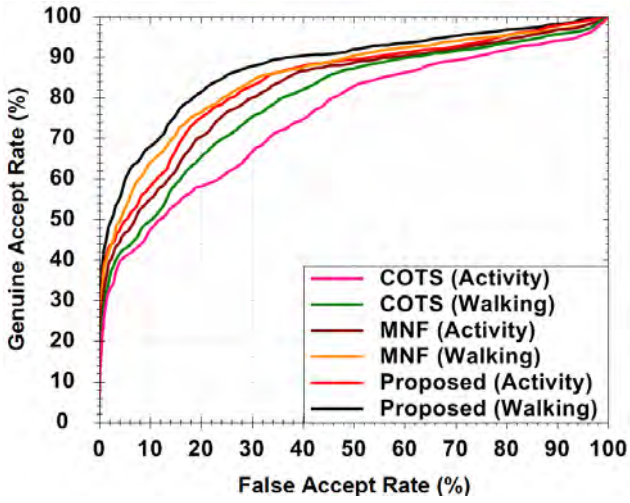

Fig. 8. ROC curves comparing the performance of the proposed algorithm with COTS and MNF on the MBGC v2 database [33] for matching activity and walking videos with the gallery comprising still face images.

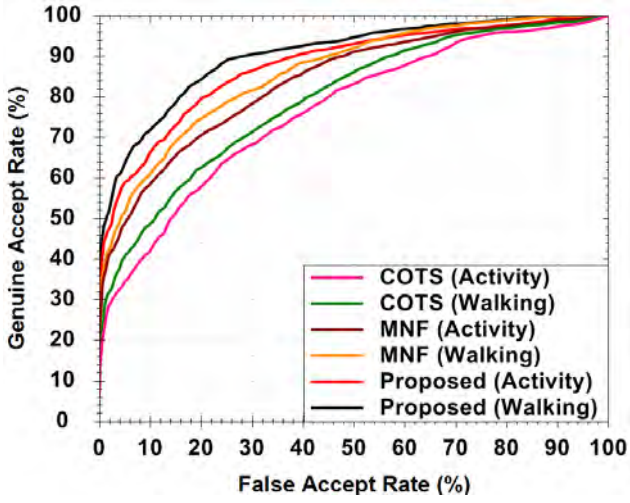

Fig. 9. ROC curves on the MBGC v2 database [33] for matching still face images with gallery comprising activity and walking videos. (Best viewed in color)

2) Results on MBGC v2 database:

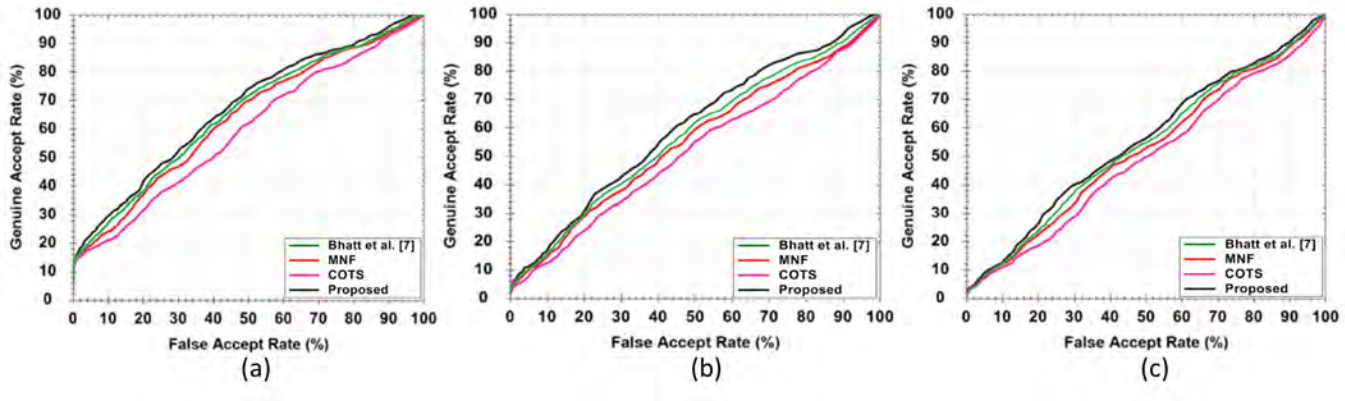

Fig. 10. ROC curves on the MBGC v2 video challenge database [33] for matching (a) walking vs walking (WW), (b) walking vs activity (WA), and (c) activity vs activity (AA) videos (Best viewed in color).

- Surveillance applications generally require matching an individual in a live-video stream with a watch-list database consisting of still face images. The proposed algorithm can efficiently represent both still face images and videos as ranked lists of still face images from the dictionary. ROC curves in Fig. 8 show the efficacy of the proposed algorithm for matching both walking and activity videos as probe with still gallery images from the MBGC v2 database. Table III demonstrates that the proposed algorithm yields at least 1.3% lower EER as compared to other algorithms for matching video probe with still gallery images.
- Matching a still probe image with video gallery also has a very important law enforcement application when a known individual has be identified at a crime scene using multiple surveillance videos of the crime scene. The results in Fig. 9 and Table III demonstrate the efficacy of the proposed algorithm for such scenarios. It yields a lower equal error rate of 17.8% and 20.1% (at least 5.2% lower than other algorithms) for matching still probe images with the gallery consisting of videos of individuals walking or performing some activity from the MBGC v2 database respectively.
- The results in Table IV and Fig. 10 show the efficacy of the proposed algorithm for matching unconstrained videos i.e. where the individual is walking or performing some activity. The proposed algorithm outperforms COTS, MNF, and existing video-to-video matching algorithm [7] for all the three matching scenarios i.e walking vs walking (WW), walking vs activity (WA), and activity vs activity (AA). The proposed algorithm yields at least 0.5% lower EER from the existing video-to-video matching algorithm (referred to as Bhatt *et al.* [7]) and at least 2.3 % lower EER from COTS and MNF on different video-to-video matching protocols of the MBGC v2 video challenge database.
- Our previously proposed video-to-video based algorithm [7] is also a rank aggregation based approach that combines multiple ranked lists for a video using Markov chain based rank aggregation. However, unlike the proposed algorithm, existing algorithm [7] does not optimize every ranked list before fusion which results in lower performance. Moreover, the proposed algorithm is computationally faster than existing algorithm [7] as it utilizes level-1 feature to compute the ranked lists and level-2 feature to compute the relevance of images in a video signature. On average, the proposed algorithm requires 7.4 seconds less than existing algorithm to compute the video signature of a video with 100 frames.
- The proposed algorithm performs better for walking vs walking experiment as compared to the other two scenarios that involve videos of individuals performing some activities. As shown in Fig. 4, activity videos are more challenging due to the presence of pose, illumination, and expression variations.
- YouTube faces database comprises videos of individuals that are captured in unconstrained environment with different types of cameras and settings whereas the MBGC v2 videos are of higher quality. Therefore, it is observed that applying the multi-scale retinex based pre-processing enhances the performance on YouTube face database by $\sim$2%. On the other hand, it has no effect on the MBGC v2 database results.
- Unlike many existing techniques that are affected by unequal number of frames in two videos [11], [21], [24], [28], [47], the proposed algorithm mitigates such limitations and can efficiently match two videos regardless of the number of frames in each video. As shown in Table III, the proposed algorithm can also match still face images, analogous to a video having a single frame with videos comprising multiple frames.

IV. Conclusion and Future Research

With advancements in technology, reduction in sensor cost (video camera), and limitations of face recognition from still images in unconstrained scenario, video based face recognition has gained significant attention from the research community. Multiple frames in a video provide temporal and intra-class variations that can be leveraged for efficient face recognition. The proposed video based face recognition algorithm is based

Gallery	Probe	Algorithm	Verification Accuracy at EER(%)	SD (%)	AUC (%)	EER (%)
Still images	Walking videos	COTS	73.2	2.1	79.6	27.1
		MNF	78.3	1.7	85.0	22.7
		Proposed	**80.6**	**1.4**	**87.6**	**19.2**
Still images	Activity videos	COTS	68.4	1.9	75.5	31.9
		MNF	76.5	1.8	82.1	24.4
		Proposed	**77.8**	**1.5**	**84.0**	**22.7**
Walking videos	Still images	COTS	70.7	2.2	79.4	29.3
		MNF	77.1	2.0	86.0	23.7
		Proposed	**82.6**	**1.7**	**90.4**	**17.8**
Activity videos	Still images	COTS	69.2	2.0	76.5	31.3
		MNF	74.3	1.7	83.8	25.3
		Proposed	**79.8**	**1.5**	**87.7**	**20.1**

TABLE III
RESULTS ON THE MBGC V2 [33] DATABASE FOR MATCHING STILL FACE IMAGES WITH VIDEOS.

Protocol	Algorithm	Verification Accuracy at EER(%)	AUC (%)	EER (%)
Walking vs walking (WW)	COTS	55.7	59.1	44.3
	MNF	59.9	63.6	40.1
	Bhatt *et al.* [7]	60.8	65.8	38.6
	Proposed	**62.2**	**67.0**	**37.8**
Activity vs walking (AW)	COTS	52.5	53.8	47.5
	MNF	54.4	55.3	45.6
	Bhatt *et al.* [7]	55.2	57.3	44.1
	Proposed	**57.4**	**59.8**	**42.7**
Activity vs activity (AA)	COTS	50.2	50.6	49.8
	MNF	51.5	52.2	48.5
	Bhatt *et al.* [7]	52.8	54.8	46.4
	Proposed	**54.1**	**55.4**	**45.9**

TABLE IV
EXPERIMENTAL RESULTS ON DIFFERENT PROTOCOLS OF THE MBGC V2 VIDEO CHALLENGE DATABASE [33].

on the observation that a discriminative video signature can be generated by combining the abundant information available across multiple frames of a video. It assimilates this information as a ranked list of still face images from a large dictionary. The algorithm starts with generating a ranked list for every frame in the video using computationally efficient level-1 features. Multiple ranked lists across the frames are then optimized using clustering based re-ranking and finally fused together to generate the video signature. Usefulness (relevance) of images in the video signature is computed using level-2 features. The video signature thus embeds large intra-personal variations across multiple frames which significantly improves the recognition performance. Finally, two video signatures (ordered ranked lists) are compared using DCG measure that seamlessly utilizes both ranking and relevance of images in the signature. This research thus transforms the problem of video based face recognition into comparing two ordered lists of images. Several experiments on YouTube and MBGC v2 video databases show that the proposed algorithm consistently outperforms existing algorithms including a commercial face recognition system.

Real world applications of face recognition involve verifying individuals at lower FARs. However, existing and proposed algorithms yield very low verification accuracies at lower FARs. For instance, on the YouTube database, even though the proposed algorithm outperforms existing algorithms at 0.1% FAR, the verification accuracy is only 23.8%. As a future research direction, we plan to improve the performance at lower false accept rates. To yield better face recognition performance across large variations, the proposed algorithm utilizes the abundant information available in a video. Therefore, it requires more computational time as compared to still face recognition algorithms. Another future research direction is to reduce the computational time of the proposed algorithm. One possible approach to enhance the computational efficiency of the proposed algorithm is to process video frames in a parallel manner.

REFERENCES

[1] G. Aggarwal, A. K. R. Chowdhury, and R. Chellappa. A system identification approach for video-based face recognition. In *Proceedings of International Conference on Pattern Recognition*, pages 175–178, 2004.

[2] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 581–588, 2005.

[3] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Marithoz, J. Matas, K. Messer, V. Popovici, F. Pore, B. Ruiz, and J. Thiran. The BANCA database and evaluation protocol. In *Audio- and Video-Based Biometric Person Authentication*, pages 625–638. 2003.

[4] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas. Face recognition from video: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(5), 2012.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[6] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. Memetically optimized MCWLD for matching sketches with digital face images. *IEEE Transactions on Information Forensics and Security*, 7(5):1522–1535, 2012.

[7] H. S. Bhatt, R. Singh, and M. Vatsa. On rank aggregation for face recognition from videos. In *Proceedings of International Conference on Image Processing*, pages 1–5, 2013.

[8] Y. C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *Proceedings of European Conference on Computer Vision*, pages 766–779, 2012.

[9] Y. C. Chen, V. M. Patel, S. Shekhar, R. Chellappa, and P. J. Phillips. Video-based face recognition via joint sparse representation. In *Proceedings of International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–8, 2013.

[10] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 3554–3561, 2013.

[11] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for video-based face recognition. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 2626–2633, 2012.

[12] K. Fukui and O. Yamaguchi. The kernel orthogonal mutual subspace method and its application to 3D object recognition. In *Proceedings of Asian Conference on Computer Vision*, pages 467–476, 2007.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[14] R. Gross and J. Shi. The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, 2001.

[15] A. Hadid and M. Pietikäinen. Manifold learning for video-to-video face recognition. In *Proceedings of International Conference on Biometric ID Management and Multimodal Communication*, pages 9–16, 2009.

[16] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 2705–2712, 2011.

[17] J. A. Hartigan. *Clustering Algorithms*. Wiley Series in Probability and Mathematical Statistics. 1975.

[18] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 121–128, 2011.

[19] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005.

[20] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[21] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[22] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[23] B. Klare and A. K. Jain. On a taxonomy of facial features. In *Proceedings of International Conference on Biometrics: Theory Applications and Systems*, pages 1–8, 2010.

[24] K. Lee, J. Ho, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 313–320, 2003.

[25] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, 99(3):303–331, 2005.

[26] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 3499–3506, 2013.

[27] S. Z. Li and A. K. Jain. *Handbook of Face Recognition, 2nd Edition*. Springer, 2011.

[28] X. Liu and T. Chen. Video-based face recognition using adaptive Hidden Markov Models. In *Proceedings of International conference on Computer Vision and Pattern Recognition*, pages 340–345, 2003.

[29] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[30] H. Mendez-Vazquez, Y. Martinez-Diaz, and Z. Chai. Volume structured ordinal features with background similarity measure for video face recognition. In *Proceedings of International Conference on Biometrics*, pages 1–6, 2013.

[31] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara, and O. Yamaguchi. Recognizing faces of moving people by hierarchical image-set matching. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[32] V. M. Patel, T. Wu, S. Biswas, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition under variable lighting and pose. *IEEE Transactions on Information Forensics and Security*, 7(3):954–965, 2012.

[33] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O'Toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M. Lui, H. Sahibzada, J. A. Scallan, III, and S. Weimer. Overview of the multiple biometrics grand challenge. In *Proceedings of International Conference on Advances in Biometrics*, pages 705–714, 2009.

[34] N. Poh, C. H. Chan, J. Kittler, S. Marcel, C. McCool, E. A. Ruanda, J. L. A. Castro, M. Villegas, R. Paredes, V. Struc, N. Pavesic and, A. A. Salah, H. Fang, and N. Costen. An evaluation of video-to-video face verification. *IEEE Transactions on Information Forensics and Security*, 5(4):781–801, 2010.

[35] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24:2115–2125, 2003.

[36] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *International Journal on Computer Vision*, 77(1-3):125–141, 2008.

[37] F. Schroff, T. Treibitz, D. Kriegman, and S. Belongie. Pose, illumination and expression invariant pairwise face-similarity measure via doppelganger list comparison. In *Proceedings of International Conference on Computer Vision*, pages 2494–2501, 2011.

[38] G. Shakhnarovich, J. W. Fisher, III, and T. Darrell. Face recognition from long-term observations. In *Proceedings of European Conference on Computer Vision*, pages 851–868, 2002.

[39] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003.

[40] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen. Video-based face recognition on real-world data. In *Proceedings of International Conference on Computer Vision*, pages 1–8, 2007.

[41] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[42] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 529–534, 2011.

[43] Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 497–504, 2011.

[44] J. Zhang, J. Gao, M. Zhou, and J. Wang. Improving the effectiveness of information retrieval with clustering and fusion. *Journal of Computational Linguistics and Chinese Language Processing*, 6(1):109–125, 2001.

[45] Z. Zhang, C. Wang, and Y. Wang. Video-based face recognition: State of the art. In *Proceedings of Chinese Conference on Biometric Recognition*, pages 1–9, 2011.

[46] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

[47] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91(12):214–245, 2003.

Himanshu S. Bhatt received the Bachelor in Technology degree in information technology in 2009 from the Jaypee Institute of Information Technology, Noida, India. He has completed his Ph.D. degree from the Indraprastha Institute of Information Technology (IIIT) Delhi, India in 2014 and joined Xerox Research Center India. His research interests include image processing, machine learning and their applications in biometrics. He is a recipient of IBM PhD fellowship 2011-13, best poster awards in IEEE BTAS 2010 and IJCB 2011.

Richa Singh received the M.S. and Ph.D. degrees in computer science in 2005 and 2008, respectively from the West Virginia University, Morgantown, USA. She is currently an Assistant Professor at the Indraprastha Institute of Information Technology (IIIT) Delhi, India. Her research has been funded by the UIDAI and DIT, India. She is a recipient of FAST award by DST, India. Her areas of interest are biometrics, pattern recognition, and machine learning. She has more than 125 publications in refereed journals, book chapters, and conferences. She is also an editorial board member of Information Fusion, Elsevier and EURASIP Journal of Image and Video Processing, Springer. Dr. Singh is a member of the CDEFFS, IEEE, Computer Society and the Association for Computing Machinery. She is the recipient of several best paper and best poster awards in international conferences.

Mayank Vatsa received the M.S. and Ph.D. degrees in computer science in 2005 and 2008, respectively from the West Virginia University, Morgantown, USA. He is currently an Assistant Professor at the Indraprastha Institute of Information Technology (IIIT) Delhi, India. He has more than 125 publications in refereed journals, book chapters, and conferences. His research has been funded by the UIDAI and DIT. He is the recipient of FAST award by DST, India. His areas of interest are biometrics, image processing, computer vision, and information fusion. Dr. Vatsa is a member of the IEEE, Computer Society and Association for Computing Machinery. He is the recipient of several best paper and best poster awards in international conferences. He is also an area editor of IEEE Biometric Compendium, area chair of Information Fusion, Elsevier, and PC Co-Chair of ICB-2013 and IJCB-2014.