ON RANK AGGREGATION FOR FACE RECOGNITION FROM VIDEOS

Himanshu S. Bhatt, Richa Singh and Mayank Vatsa

IIIT-Delhi, India

ABSTRACT

Face recognition from still face images suffers due to intra-personal variations caused by pose, illumination, and expression that degrade the performance. On the other hand, videos provide abundant information that can be leveraged to compensate the limitations of still face images and enhance 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. The video signature embeds diverse intra-personal and temporal variations across multiple frames, thus facilitates matching two videos with large variations. Two videos are matched by comparing their discriminative signatures using the Kendall tau similarity distance measure. Performance comparison with the benchmark results and a commercial face recognition system on the publicly available YouTube faces database show the efficacy of the proposed video based face recognition algorithm.

Index Terms— Video based face recognition, Rank aggregation, Dictionary based face recognition

1. INTRODUCTION

Face recognition from still images is a well-studied problem and several algorithms have been proposed to address different covariates such as pose, illumination, expression, aging, and disguise [1]. However, the performance of face recognition algorithms is affected by large intra-personal variations in unconstrained scenario. Though significant amount of research has been done in matching still face images under different conditions, use of videos for face recognition is relatively less explored. The challenges and limitations of still face recognition drive the research in video based face recognition. Moreover, widespread use of video cameras for surveillance and security applications, improvements in quality, and reduction in price of sensors (video cameras) have stirred extensive research interest in video based face recognition. Videos cover wide intra-personal variations with multiple frames capturing different pose, illumination, and expression variations. This diverse information can be aggregated together for efficient face recognition across large variations.

Survey on video based face recognition by Barr *et al.* [2] categorizes different approaches as set-based and sequence-based approaches. Set-based approaches [3, 4] utilize abundance and variety of information in a video to achieve resilience to sub-optimal capture conditions. Approaches that model image sets as distributions use between-distribution similarity to match two image sets [5, 6]. However, the performance of such approaches is dependent on the parameter estimation of the underlying distribution. Image sets are often modeled as linear sub-spaces [7, 8] and manifolds [5, 9, 10] where matching is performed by measuring the similarity between the input and reference subspaces/manifolds. The performance of a subspace/manifold based approach is dependent on maintaining the image set correspondences. To address this limitation, Cui *et al.* [11] propose to align two image sets using a common reference set before matching. On the other hand, sequence-based approaches explicitly utilize the temporal information, such as modeling it with Hidden Markov Models (HMM) [12], for improved face recognition performance.

This research proposes a video based face recognition algorithm where the discriminative signature of a video is generated as an ordered list of still face images from a dictionary. A dictionary is a large collection of face images where every individual has multiple images captured under different pose, illumination, and expression variations. Further, two videos are matched by comparing their video signatures using Kendall tau similarity distance [13].

2. PROPOSED ALGORITHM

Recent studies in face recognition [14, 15] show that generating image signature based on a dictionary is more robust for matching images across large variations than directly comparing two images or its features. Patel *et al.* [14] proposed a sparse approximation based approach where test images are projected onto a span of elements in learned dictionaries and resulting residual vectors are used for classification. Primarily, dictionary based face recognition approaches are limited to still face images except for a video based face recognition technique proposed by Chen *et al.* [16] using video-dictionaries. Moreover, existing approaches discard the characteristics embedded in ranked lists and only consider the overlap between two ranked lists as final similarity. As shown in Fig. 1, the proposed algorithm starts by extracting multi-

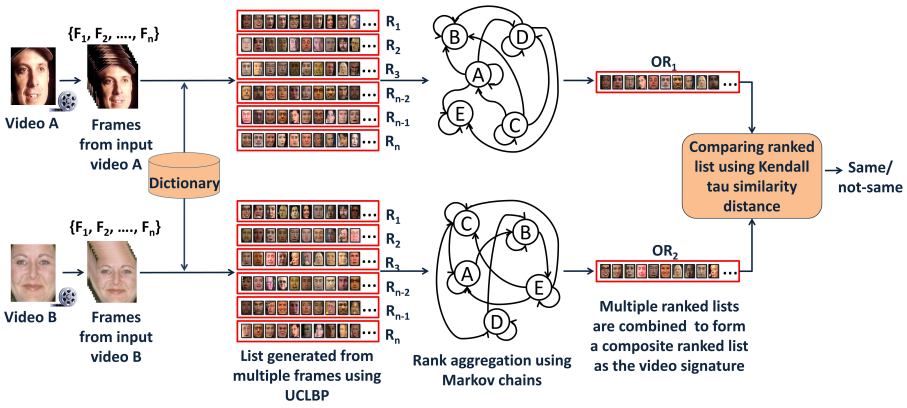

Fig. 1. Illustrates the block diagram of the proposed approach for matching two videos.

ple frames from a video. A ranked list of images from the dictionary is computed for each frame. A ranked list is an ordered list of still face images where each image is positioned based on its similarity to the input frame with the most similar image positioned at the top of list. Since each video has several frames, the algorithm computes multiple ranked lists across all video frames. These ranked lists are then combined into a single list using Markov chain based rank aggregation technique [13]. The aggregated ranked list forms the discriminative video signature. Finally for matching two videos, their aggregated ranked lists are compared using Kendall tau similarity distance that incorporates the rankings as well as the similarity among images to compute the final similarity between two lists. The proposed algorithm draws its motivation from rank aggregation techniques in information retrieval. In biometrics, rank aggregation techniques such as Borda count [17] have been explored for rank-level fusion; however to the best of our knowledge, it is the first approach that proposes rank aggregation for combining multiple ranked lists to characterize an individual in a video.

2.1. Dictionary

Dictionary is a large collection of still face images where each individual has multiple images capturing a wide range of intra-personal variations. In our research, the dictionary comprises $38,488$ images pertaining to 337 individuals from the CMU Multi-PIE [18] database captured in four sessions.

2.2. Computing the Ranked List

Let U be the set of all images in the dictionary and V be a video of an individual comprising n frames. Face region from each frame is detected[1] and represented as $\{F_1, F_2, ..., F_n\}$. To generate the ranked lists, face region from each frame is compared with all images in the dictionary using uniform circular local binary patterns (UCLBP) [19, 20]. In this research, UCLBP is used because of its robustness to gray-level intensity changes and high computational efficiency. For computing UCLBP descriptor, the face image is first tessellated into non-overlapping local patches of size 32×32. For each local patch, the UCLBP descriptor is computed based on 8 neighboring pixels uniformly sampled on a circle of radius size 2. The concatenation of descriptors from each local patch constitutes the UCLBP descriptor of an image and two UCLBP descriptors are matched using χ^2 distance. To generate a ranked list R_i corresponding to the input frame F_i, the retrieved dictionary images are positioned based on their distance to F_i with the least distinct image positioned at the top of the list. For a video V, the proposed algorithm computes a set of ranked lists $\{R_1, R_2, ..., R_n\}$ corresponding to n frames of the video. Multiple ranked lists across different video frames facilitate to capture large temporal and intra-personal variations of an individual.

2.3. Aggregating Ranked List using Markov Chains

Multiple ranked lists computed across n frames of a video have significant amount of overlap. It is computationally expensive and inefficient to compare multiple ranked lists across two videos because of the redundant information. Therefore, multiple ranked lists of a video are aggregated into a single optimized ranked list, denoted as OR, to form the video signature. Rank aggregation is aimed at computing an aggregate ranking that minimizes the distance from each of the input ranked lists. To compute the video signature, multiple ranked lists are mapped onto a Markov chain. A Markov chain is represented by a set of states (nodes) S=$\{1, ..., s\}$. The process starts in one of these states and moves successively from one state to another with a probability denoted by $p_{i,j}$. The probability depends only upon the current state and not on any previous states. The probabilities $p_{i,j}$ are called transi-

[1]OpenCV's boosted cascade of haar-like features is used for face detection and detected faces are resized to 192×224 pixels.

tion probabilities and are represented by a transition probability matrix P of dimension $s \times s$. The transition probabilities are defined by relative rankings of images in the ranked lists across multiple frames of the video.

2.3.1. Mapping Ranked Lists onto a Markov Chain

Dwork *et al.* [21] proposed different schemes to map multiple ranked lists onto a Markov chain. According to the MCS4 mapping scheme, a transition from the state $S_k = I$ to S_{k+1} is performed by choosing a state J randomly from U. I and J are the images retrieved from dictionary that form states in the markov chain. If $r(J) < r(I)$, where $r(\cdot)$ represents the rank, for a majority of ranked lists that ranked both I and J, then $S_{k+1} = J$, otherwise similarity transition with a probability γ ($\gamma = 1$) is executed. A similarity transition is defined based on the similarity among the nodes. A similarity transition from $S_k = I$ is executed by selecting J from U randomly from the weighted distribution,

$$Pr(I \rightarrow J) = \frac{Sim(I, J)}{\sum_{l \in U} Sim(I, l)} \tag{1}$$

where, $Sim(I, J)$ is the similarity between two images I and J computed using UCLBP. It facilitates to utilize the similarity among the images along with their rankings across multiple ranked lists of a video. If no similarity transition is executed, then an epsilon transition with a probability ϵ ($\epsilon = 0.1$) is executed. Epsilon transition is executed at S_k by choosing an item J randomly from U and setting $S_{k+1} = J$. Epsilon transitions eliminate the possibility of sink nodes in the Markov chain and ensure a smooth ranking of all items in U. If no epsilon transition is executed, then $S_{k+1} = I$. Using these mapping rules, multiple ranked lists for a video are transformed into a Markov chain.

2.3.2. Ordering Nodes using Stationary Distribution

The stationary distribution represents the proportion of time that a Markov chain is in any particular state. A row vector π is called stationary distribution over a set of states S if π is a probability distribution such that $\pi = \pi P$, where P is the transition probability matrix. It represents the equilibrium state of a Markov chain and can be approximated using power method on the transition probability matrix to find the dominant eigenvector of the matrix. Dwork *et al* [21] observed that stationary distribution implies an ordering on states when the transition probabilities in a Markov chain are represented by relative rankings. In this research, it is utilized for combining ranked lists across multiple frames of a video to generate a combined ranked list as the video signature. The states in a Markov chain are ranked based on the stationary distribution where the state with the highest value in π is positioned at the top of the ranked list. It is observed that the similarity between a video frame and images in the dictionary drops after a particular rank (say rank q) and the order of images is less discriminative beyond that point. Therefore, in the final aggregated ranked list, images till rank q ($q = 100$ in our case) are considered as the video signature. The algorithm to compute the discriminative video signature based on still face images from the dictionary is described in Algorithm 1.

Algorithm 1 Algorithm for computing the video signature

Input: A set of dictionary images U and a video V.
Process: Decompose video V into n frames and detect facial regions as $F_1, F_2, ..., F_n$ for all frames.
Compute Ranked Lists
Iterate: i= 1 **to** n (number of frames)
Compute R_i from U using UCLBP and χ^2 distance.
end iterate.
Rank Aggregation
Step-1: Map ranked lists $R_1, R_2, ..., R_n$ onto a Markov chain M : states in $M \in U$.
Step-2: Compute the stationary distribution π on M.
Step-3: Sort states in M with decreasing values in π.
Output: Aggregated ranked list OR as the video signature.

2.4. Matching the Ranked Lists

Rank aggregation across multiple ranked lists yields similar rankings to similar items. Therefore, the distance measure to compare two ranked lists should penalize the score if two similar items are not placed nearby in the aggregated ranked lists. The standard Kendall tau distance does not account for similarity among items in the list. Therefore, to incorporate the similarity among the dictionary images while computing the number of pairwise disagreements in the ranked lists, the Kendall tau similarity distance proposed by Sculley [13] is used. It incorporates the similarity among dictionary images by formulating an aggregate similarity position function (g) and an aggregate similarity list (R_g). The aggregate similarity position of an image, I, for a list R_i with similarity function $Sim(\cdot, \cdot)$ is defined as:

$$g(I, R_i, Sim(\cdot, \cdot)) = \frac{\sum_{J \in R_i} Sim(I, J) R_i(J)}{\sum_{J \in R_i} Sim(I, J)} \tag{2}$$

where J is an image in the ranked list R_i and $Sim(\cdot, \cdot)$ is the similarity computed using UCLBP. Aggregate similarity list, R_g, is composed from lists OR_1 and OR_2 such that for every image $I \in OR_1$, $R_g(I) = g(I, OR_2, Sim(\cdot, \cdot))$. The Kendall tau similarity distance between two ranked lists OR_1 and OR_2 is given as:

$$\begin{aligned} K_{sim}(OR_1, OR_2, Sim(\cdot, \cdot)) = \\ \frac{1}{2}\{K(OR_1, R_g(OR_1, OR_2, Sim(\cdot, \cdot))) + \\ K(OR_2, R_g(OR_2, OR_1, Sim(\cdot, \cdot)))\} \end{aligned} \tag{3}$$

Thus, $K_{sim}(OR_1, OR_2)$ is a distance score computed as an average of the Kendall tau distance between (1) OR_1 and its aggregate similarity list drawn from OR_2 and (2) OR_2 and its aggregate similarity list drawn from OR_1.

3. EXPERIMENTAL EVALUATION

The efficacy of the proposed video based face recognition algorithm is evaluated on the YouTube faces database [4] comprising $3,425$ videos of $1,595$ individuals downloaded from YouTube. The database provides ten-fold pair-wise matching ('same'/'not-same') test benchmark protocol. In our experiments, training is performed on nine splits and the performance is computed on the tenth split. The final performance is reported as an average of the 10 folds. The performance of the proposed algorithm is compared with the benchmark results provided by Wolf *et al.* [4] and an off-the-shelf commercial face recognition system, Neurotechnology VeriLook (referred to as COTS). For matching two videos using COTS, set-to-set matching is used where each frame in the first video is matched to all frames in the second video. The mean score obtained corresponding to all 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.

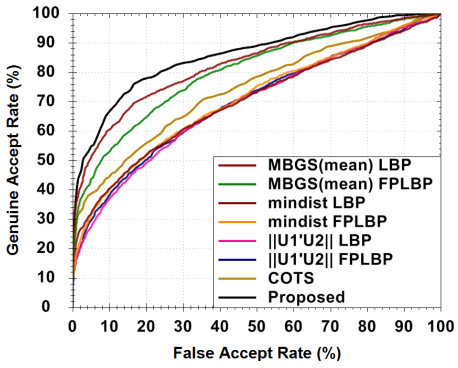

Fig. 2. ROC curves compare the performance of proposed algorithm with benchmark tests on the YouTube faces database [4].

Receiver operating characteristic (ROC) curves in Fig. 2 show the efficacy of the proposed algorithm for unconstrained video based face recognition in verification scenario. Table 1 reports that the proposed algorithm achieves an average accuracy of 78.3% at an equal error rate (EER) of 21.6%. The area under the curve (AUC) is at least 3% greater than existing approaches and COTS. This improvement in the performance is due to the fact that video signatures generated using dictionary of still face images capture wide intra-personal and temporal variations across multiple frames, and thus can be efficiently used to match videos with large variations. Markov chain based aggregation yields a list which minimizes the overall distance from multiple ranked lists across the video frames and helps characterizing an individual in the video. Existing video based approaches that use set-to-set similarities do not consider that multiple frames capture different intra-personal variations. Matching such diverse image sets independently leads to sub-optimal performance. The proposed algorithm also incorporates the similarity among the images in ranked list along with their rankings. It facilitates to compensate for noisy rankings and enhances the effectiveness of comparison between the ranked lists. Fig. 3 shows examples where the proposed algorithm successfully classified and where it failed to correctly classify the videos.

Table 1. Performance comparison on the YouTube faces database [4]. Verification accuracy is reported at EER.

Algorithm	**Accuracy $\pm$ SD**	**AUC**	**EER**
MBGS(mean) CSLBP [4]	72.4±2.0	78.9	28.7
MBGS(mean) FPLBP [4]	72.6±2.0	80.1	27.7
MBGS(mean) LBP [4]	76.4±1.8	82.6	25.3
COTS	67.9±2.3	74.1	33.1
Proposed	**78.3±1.7**	**85.8**	**21.6**

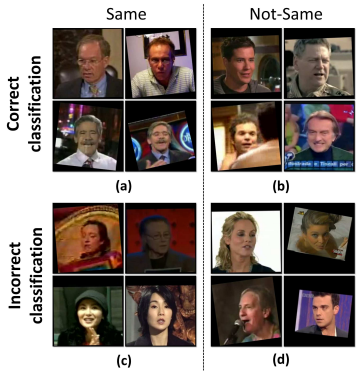

Fig. 3. Illustrating examples when the proposed algorithm correctly classified (a) same, (b) not-same video pairs (represented row-wise). The proposed algorithm incorrectly classified (c) same and (d) not-same video pairs.

4. CONCLUSION

This research transforms the problem of video based face recognition into the problem of comparing two ordered lists of images. It starts with computing a ranked list for every frame in the video using computationally efficient texture based features. Multiple ranked lists across the frames are then combined using Markov chain based rank aggregation to form the video signature. The video signature thus embed large intra-personal variations across multiple frames. Finally to match two videos, kendall tau distance measure is used to compare their video signatures. The proposed algorithm provides significant improvements in performance on the YouTube faces database.

5. REFERENCES

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

[2] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas, "Face recognition from video : A review," *Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 5, 2012.

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

[4] 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*, 2011, pp. 529–534.

[5] 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*, 2005, pp. 581–588.

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

[7] 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*, 2004, pp. 175–178.

[8] 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*, 2007, pp. 1–8.

[9] 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*, 2011, pp. 2705–2712.

[10] R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao, "Manifold-manifold distance and its application to face recognition with image sets," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4466–4479, 2012.

[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*, 2012, pp. 2626–2633.

[12] 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*, 2008, pp. 1–8.

[13] D. Sculley, "Rank aggregation for similar items," in *Proceedings of the International Conference on Data Mining*, 2007, pp. 587–592.

[14] 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*, vol. 7, no. 3, pp. 954–965, 2012.

[15] 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*, 2011, pp. 2494–2501.

[16] 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*, 2012, pp. 766–779.

[17] A. Abaza and A. Ross, "Quality based rank-level fusion in multibiometric systems," in *Proceedings of International Conference on Biometrics: Theory, Applications, and Systems*, 2009, pp. 1–6.

[18] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[19] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[20] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[21] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of International Conference on World Wide Web*, 2001, pp. 613–622.