Image Set-based Face Recognition: A Local Multi-Keypoint Descriptor-based Approach

Na Liu[1], Meng-Hui Lim[2], Pong C. Yuen[2], and Jian-Huang Lai[1]

[1]School of Maths and Computational Science
Sun Yat-sen University
Guangzhou, China
lindaliumail@gmail.com, stsljh@mail.sysu.edu.cn

[2]Department of Computer Science
Hong Kong Baptist University
Kowloon, Hong Kong
{mhlim, pcyuen}@comp.hkbu.edu.hk

Abstract

Image-set-based face recognition has recently attracted much attention due to widespread of surveillance and video retrieval applications. Extraction of partial and misaligned face images from a video is relatively common in unconstrained scenarios and in the presence of detection/localization error, respectively. However, existing face recognition techniques that consider holistic image-set representation would not perform well under such conditions. In this paper, we introduce a local image-set-based face recognition approach to address this issue, where each image set is represented by a cluster set of keypoint descriptors and similarity between image sets is measured by the distance between the corresponding sets of clusters. Our representation is robust to misalignment because the extraction of descriptors is carried out without respect to the absolute face position. Additionally, our approach is robust to partial face occlusion due to that (1) descriptors corresponding to non-occluded keypoints are not affected by the occluded keypoints; (2) matching decision is contributed only by distances between the matched cluster pairs corresponding to the non-occluded facial parts. Extensive experiment evaluation shows that our approach is able to achieve very promising recognition rates.

1. Introduction

With the recent widespread of surveillance and video retrieval applications, image set-based face recognition has attracted enormous research interest throughout the last decade [1][2][3][5][6][8][17][18]. Since face images in image set-based face recognition are collected from video sequences, both training and test examples are comprised of sets of an individual's face images and the final recognition decision is made based on comparisons of such image sets.

In practice, face images captured from a surveillance video for example are often obtained without user cooperation and knowledge. Frequently, the face of an individual captured in the video could be partially occluded [11]. Furthermore, since faces are usually extracted from a video frame sequence by using a face detector/tracker, faces in the video frames are rarely perfectly aligned over the set of extracted images due to potential detection or localization error of the imperfect face detector/tracker. When holistic face representation is adopted in image-set-based face recognition [1][2][5][6][8][17][18], simultaneous occurrence of face occlusion and misalignment deteriorates face recognition performance.

A straightforward way to tackle these challenges is to apply an existing single-probe-image-based face recognition method that is occlusion- and misalignment-robust in the image-set-based setting. Recently, a local face recognition approach, namely the multi-keypoint descriptor (MKD)-based approach [11] has been proposed to address the occlusion and misalignment problems in the single probe image-based setting. This approach (1) extracts a number of salient facial keypoints and a descriptor per keypoint from a face image without requiring the face to be pre-aligned with that in the other face images; and (2) performs recognition via applying Sparse Representation-based Classification (SRC) [19] on a large dictionary of keypoint descriptors [11]. It is worth to note, however, that adopting it straightforwardly in the image set-based setting is not appropriate due to suboptimal discrimination power and unacceptably-low efficiency.

To avoid these drawbacks, it is the objective of this paper to develop a simple and effective local approach for image-set-based face recognition under uncontrolled conditions. We first detect keypoints in each image, extract an alignment-free descriptor per keypoint and pool these descriptors over a set of images in a common feature space. To derive a robust representation from consistent keypoints in the spatial domain (which leads to dense descriptors in the feature domain), we adopt a density-based clustering approach to select dense descriptors corresponding to consistent keypoints and to group descriptors according to their facial parts into a number of clusters. With this cluster representation of an image set, we devise a series of occlusion-robust matching procedures to evaluate the

978-0-7695-4990-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPRW.2013.31

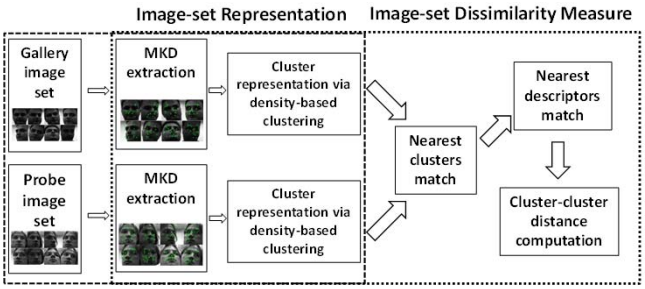

Figure 1: Our local MKD-based approach to image-set-based face recognition.

similarity between probe and gallery sets of clustered descriptors. The algorithm of our approach is shown in the block diagram of Figure 1.

The significance of our contribution can be summarized in the following two aspects:

1) We propose a novel alignment-free representation for image set – a set of clustered descriptors associated with consistent keypoints detected at various facial parts. Our approach represents an image set belonging to a *specific* individual with a set of clustered descriptors that are extracted solely based on the appearance consistency of the keypoints detected in the image set.
2) Based on our clustering representation of the image set, we devise a series of occlusion-robust cluster-to-cluster matching procedures. The key to achieve such occlusion-robust matching is to take into account the distance between the matched cluster pairs only, when considering the matching between probe and gallery cluster sets. This is to ensure that missing clusters of descriptors due to frequently-undetected keypoints will not contribute to the final matching decision.

2. Literature Review

Existing image-set-based classification approaches differ in the ways the sets are modelled and the set similarity is computed. Set modelling approaches can be broadly divided into parametric and non-parametric representations; while the similarity function is defined mostly based on the set modelling methods. Parametric modelling approaches represent each image set with a parametric distribution function and estimate relevant parameters from the training set data [1][9]; whereas the more favourable non-parametric modelling-based approaches relax the assumptions on data distributions.

Linear subspace-based approach is a non-parametric modelling-based approach that models each image set as a linear subspace. For instance, a linear discriminant function [8] was developed to maximize and minimize the canonical correlations of within-class sets and between-class sets, respectively. The similarity between two sets is measured in terms of the canonical correlations.

Manifold-based approach is another popular non-parametric approach that expresses an image set as a collection of local linear models using a manifold. A notable work in this direction measures similarity between manifolds in terms of Manifold-Manifold Distance (MMD) [17]. Covariance Discriminative Learning (CDL) [18] is another manifold-based approach that models each image set using its covariance matrix. Since such nonsingular covariance matrices naturally lie on a Riemannian manifold, a kernel function is derived to log-map the covariance matrices from the manifold to a Euclidean space for enabling the use of classical learning algorithms in the vector space.

The final class of non-parametric modelling-based approach represents each image as a point in a linear or affine feature space and characterizes each image set by a convex geometric region known as affine or convex hull [2]. The similarity between sets is measured by geometric distances between convex models. Another approach built upon the affine hull model uses a more effective between-set distance called Sparse Approximated Nearest Point (SANP) distance [6]. For similarity measure, this approach takes the distance between the nearest points of two sets, which is sparsely approximated from the image samples of their respective set.

While these non-parametric modelling-based approaches are able to perform impressively on typical face image sets, we found that none of these is able to tackle the partial face occlusion and misalignment problems. On the contrary, our approach adopts a local approach that represents each image set as a cluster set of alignment-free descriptors and measures set similarity by occlusion-robust cluster-to-cluster distance to overcome the partial face occlusion and misalignment problems.

3. Methodology

3.1. A Cluster Representation for Image Sets

Reliable image-set-based face recognition requires robust feature representation that can be derived from consistently-detected local interest points. In our setting where each set of face images belongs to a specific individual, it is, however, difficult to tell whether a keypoint detected in different images within the set are corresponding to the same facial point although they could be closely located to each other, since characteristics of keypoints may vary in accordance with variations in facial expression, pose, age, illumination and occlusion.

Hence, instead of seeking for a consistent keypoint detector, we approach the problem of finding consistent keypoints by seeking for “dense regions” in the feature domain. The feasibility of this approach relies on the fact that (1) descriptors extracted from the same (resp. different) keypoints are similar to (resp. dissimilar from) each other

due to the robust repeatability property of keypoint descriptors [11][13]; (2) descriptors are likely to vary from individual to individual, although the keypoints are detected at common facial parts of these individuals. By pooling the descriptors from an image set in a common feature space, density of the descriptor distribution thus reflect the consistency of keypoint appearance. By performing a fast density-based clustering, sparsely distributed descriptors (inconsistent keypoints) can be excluded and the remaining dense descriptors (consistent keypoints) are categorized into multiple clusters, thus forming our final cluster representation with each cluster representing keypoints at a single facial part.

3.1.1 Keypoint Detection and Descriptor Extraction

To obtain an alignment-free and occlusion-robust facial feature representation, the two important requirements are that (1) the features extracted from the non-aligned face images of an individual should be similar; (2) the entire set of extracted features should not be affected when partial face occlusion occurs so that the remaining non-distorted features can still be used for recognition.

The MKD representation is able to satisfy these requirements, not only because the extracted descriptors are robustly repeatable, but also because the descriptors are independent to one another. Hence, in the presence of misalignment and occlusion problems, the descriptors representing the non-occluded parts can be consistently extracted as long as the corresponding keypoints can be detected. In general, MKD can be computed via keypoint detection and local feature descriptor extraction.

3.1.2 Fast Density-based Clustering

When a facial keypoint is detected consistently over an image set and a descriptor is extracted per keypoint, pooling these descriptors together would result in a high density cluster of descriptors in the feature space. Vice versa, inconsistent-appearing keypoints would lead to sparsely distributed descriptors. By making use of this observation, we develop a density-based clustering method to (1) exclude sparsely-distributed descriptors from contributing to the recognition decision; (2) classify descriptors associated with different types of keypoint in order to make our subsequent cluster-to-cluster distance evaluation meaningful. We desire tight clusters, such that only "sufficiently dense" descriptors are selected for face representation, where such sufficiency is later characterized by the term "saliency". To ensure that only descriptors associated with a single type of keypoint are enclosed within a cluster, we further exploit spatial information (offset in 2D coordinate) of the keypoints to facilitate exclusion of minor descriptors that do not correspond to the same type of keypoint.

Let D_s and D_f be the spatial space and the feature space, respectively. By pooling the detected keypoints and extracted descriptors in the c^{th} image set, we obtain a total of K descriptors $\delta^{(c)} = \left\{\delta_1^{(c)}, \delta_2^{(c)}, \dots, \delta_K^{(c)}\right\}$ associated with the K keypoints $P^{(c)} = \left\{p_1^{(c)}, p_2^{(c)}, \dots, p_K^{(c)}\right\}$, where $\delta_k^{(c)} \in D_f$ represents the feature vector of the k^{th} descriptor while $p_k^{(c)} \in D_s$ represents the 2D coordinate vector of the k^{th} keypoint.

Our clustering approach can be regarded as a fast variant of DBSCAN [4], which (a) considers only core descriptors in clustering and (b) incorporates the use of spatial information in cluster formation. Before describing our algorithm, we give the notions of ε-neighborhood, saliency and core/border descriptors:

Definition 1 [4]**:** (ε -neighbourhood of a point) The *ε-neighborhood* of a point ρ_k in a space D is defined by $N_\varepsilon(\rho_k) = \{\rho_{k'} | k \neq k';\ \|\rho_k - \rho_{k'}\| \leq \varepsilon;\ \rho_k, \rho_{k'} \in D\}$.

Point ρ_k in Definition 1 could be a keypoint ($\rho_k = p_k^{(c)}$) or a descriptor $\left(\rho_k = \delta_k^{(c)}\right)$. Informally, a neighborhood $N_\varepsilon(\rho_k)$ is defined as an area of radius ε around ρ_k that contains points that bear similar characteristics to ρ_k.

Definition 2: (Saliency of a descriptor) The saliency σ of a descriptor $\delta_k^{(c)}$ is defined as the discrete density around $\delta_k^{(c)}$ and its corresponding keypoint $p_k^{(c)}$, which is given by $\sigma\left(\delta_k^{(c)}, p_k^{(c)}\right) = \left|\varphi\left(N_{\varepsilon_{desc}}\left(\delta_k^{(c)}\right)\right) \cap \varphi\left(N_{\varepsilon_{keyp}}\left(p_k^{(c)}\right)\right)\right|$, where $\varphi(N_\varepsilon)$ represents the indices of the neighboring points within N_ε.

Saliency in Definition 2 is a density indicator that computes the quantity of neighbouring descriptors of $\delta_k^{(c)}$, provided that the neighboring descriptors are also neighbors of $p_k^{(c)}$ in the spatial domain. Saliency $\sigma(\cdot)$ of a descriptor ranges from 0 to $K-1$. The main intent of leveraging ε-neighborhood of keypoints ε_{keyp} in the computation of $\sigma(\cdot)$ is to penalize minor (similar) descriptors that represent dissimilar facial part from $p_k^{(c)}$. This is important to ensure that (1) most descriptors in a dense set associated with a common facial part are assigned with a large saliency value; and (2) descriptors of the same type are matched correctly in the subsequent stage of similarity evaluation between two image sets.

Definition 3: (Core and border descriptor) A keypoint descriptor $\delta_k^{(c)}$ is defined either as a *core descriptor* if $\left(\delta_k^{(c)}, p_k^{(c)}\right) \geq MinPts$, or a *border descriptor* otherwise.

An ε-neighbourhood of a core descriptor contains more neighboring descriptors than an ε-neighbourhood of a border descriptor. Since the saliency of a core descriptor is greater than that of a border descriptor, adjacent dense sets of descriptors with non-zero core descriptors are always separated by some border descriptors. Given a saliency threshold $MinPts$, core and border descriptors can be distinguished. To ensure formation of tight cluster, our

density-based algorithm considers only the core descriptors in the clustering process to eradicate potential local outlier descriptors (border descriptors) within each cluster that may affect the computation of distance between clusters in the next matching step.

Clustering Algorithm. Given parameters ε_{keyp}, ε_{desc} and $MinPts$, the following steps are carried out to obtain a clustering representation for an image set:

Step 1: Compute the saliency value (discrete density) of all K descriptors $\delta^{(c)}$ based on ε_{keyp} and ε_{desc}.

Step 2: Identify core and border descriptors based on ε_{desc} and $MinPts$ and discard border descriptors (sparsely distributed descriptors).

Step 3: Connect neighboring core descriptors together, such that each dense set of descriptors is grouped into a single cluster if the descriptors in each set are also neighbors of one another in the spatial domain.

With this, the final representation of n_{clus} clusters of core descriptors can be obtained.

3.2. Set Similarity: Cluster-to-Cluster Distance

Once each image set is represented by a set of clustered descriptors, similarity between two image sets can be measured by the distance between their corresponding cluster representations, namely *cluster-to-cluster distance*. The corresponding clusters in the two image sets are identified so that local matching can be performed. To ensure robustness of our approach to partial face occlusion in probe image set, our approach does not allow missing clusters (occluded part) to have any influence in the computation of the cluster-to-cluster distance, since the corresponding gallery cluster of the missing probe cluster will less probably be paired up with another non-corresponding gallery cluster for distance computation. Hence, the robustness of our approach to partial occlusion can be said to rely on the robustness of the matching criteria used in finding matched cluster pairs in (5).

Assume that the gallery and probe sets of clusters are denoted by $C^{(G)} = \{C_1^{(G)}, C_2^{(G)}, \dots, C_i^{(G)}, \dots, C_m^{(G)}\}$ and $C^{(P)} = \{C_1^{(P)}, C_2^{(P)}, \dots, C_j^{(P)}, \dots, C_n^{(P)}\}$, respectively, for $C_i^{(G)} = \left\{\delta_{i1}^{(G)}, \dots, \delta_{ix}^{(G)}, \dots, \delta_{in_i}^{(G)}\right\}$ and $C_j^{(P)} = \left\{\delta_{j1}^{(P)}, \dots, \delta_{jy}^{(P)}, \dots, \delta_{jn_j}^{(P)}\right\}$.

To identify which of the gallery clusters in $C^{(G)}$ corresponds to each probe cluster $C_j^{(P)}$, the nearest and second nearest cluster indices of the j^{th} probe cluster $(n_p(j), sn_p(j))$ and the i^{th} gallery cluster $(n_g(i), sn_g(i))$ are computed by

$$n_p(j) = i' = \arg\min_i \left\| \overline{C_J^{(P)}} - \overline{C_\iota^{(G)}} \right\| \quad (1)$$

$$n_g(i) = j' = \arg\min_j \left\| \overline{C_\iota^{(G)}} - \overline{C_J^{(P)}} \right\| \quad (2)$$

$$sn_p(j) = i'' = \arg\min_{i, i \neq n_p(j)} \left\| \overline{C_J^{(P)}} - \overline{C_\iota^{(G)}} \right\| \quad (3)$$

$$sn_g(i) = j'' = \arg\min_{j, j \neq n_g(i)} \left\| \overline{C_\iota^{(G)}} - \overline{C_J^{(P)}} \right\|. \quad (4)$$

Here, $\overline{C_\iota^{(G)}} = \frac{1}{n_i}\sum_{x=1}^{n_i} \delta_{ix}^{(G)}$ and $\overline{C_J^{(P)}} = \frac{1}{n_j}\sum_{y=1}^{n_j} \delta_{jy}^{(P)}$ are the centers of $C_i^{(G)}$ and $C_j^{(P)}$, respectively. A positive match with the closest gallery cluster is considered to be found if

$$n_p\left(n_g(i)\right) = i \quad (5)$$

when both matching criteria [13]: $\frac{\left\|\overline{C_J^{(P)}} - \overline{C_{n_p(J)}^{(G)}}\right\|}{\left\|\overline{C_J^{(P)}} - \overline{C_{sn_p(J)}^{(G)}}\right\|} \leq \tau$ and $\frac{\left\|\overline{C_\iota^{(G)}} - \overline{C_{n_g(\iota)}^{(P)}}\right\|}{\left\|\overline{C_\iota^{(G)}} - \overline{C_{sn_g(\iota)}^{(P)}}\right\|} \leq \tau$ are satisfied. The same comparison with the gallery clusters is carried out for every probe cluster $C_j^{(P)}$ and only the matched cluster pairs are selected for subsequent distance computation. This rules out the possibility of having (1) clusters representing different facial parts erroneously matched against each other; (2) cluster in gallery set corresponding to missing cluster in probe set (due to undetected keypoint) erroneously matched against other cluster in probe set. However, if no matched cluster pair that can be found, then $C^{(P)}$ and $C^{(G)}$ are considered to be belong to different individuals.

For each matched pair of clusters, say $C_{i^*}^{(G)}$ and $C_{j^*}^{(P)}$ for $j^* \in S_{match}$, the closest descriptor $\delta_{i^*x}^{(G)} \in C_{i^*}^{(G)}$ to each descriptor of $\delta_{j^*y}^{(P)} \in C_{j^*}^{(P)}$ is identified, such that

$$d\left(\delta_{j^*y}^{(P)}, C_{i^*}^{(G)}\right) = \min_{\delta_{i^*x}^{(G)} \in C_{i^*}^{(G)}} \left\|\delta_{j^*y}^{(P)} - \delta_{i^*x}^{(G)}\right\|. \quad (6)$$

By letting $d_h\left(\delta_{j^*y}^{(P)}, C_{i^*}^{(G)}\right)$ to be the h^{th} minimum distance among the n_{j^*} distance comparisons, we obtain the distance between clusters for a matched pair as

$$d\left(C_{j^*}^{(P)}, C_{i^*}^{(G)}\right) = \sum_{h=1}^{\gamma_{j^*}} d_h\left(\delta_{j^*y}^{(P)}, C_{i^*}^{(G)}\right) \quad (7)$$

where $\gamma_{j^*} = \lceil R n_{j^*} \rceil$ for $0 < R < 1$. The idea of considering only the first γ_{j^*} instead of all minimum distances is to further prevent the matching results from being negatively affected by potential outliers within the clusters. Finally, the cluster-to-cluster distance of all matched pairs can be expressed by

$$d\left(C^{(P)}, C^{(G)}\right) = \frac{1}{N}\sum_{j^* \in S_{match}} d\left(C_{j^*}^{(P)}, C_{i^*}^{(G)}\right) \quad (8)$$

where $N = \sum_{j^* \in S_{match}} \gamma_{j^*}$ is the total number of descriptor comparisons. The distance measure in Eq. (8) thus serves as the robust set similarity measure of our approach. Once the cluster-to-cluster distance is computed for every gallery set with reference to a probe set, only those gallery sets with at least M matched clusters are taken into consideration and a nearest neighbour classifier is used for identification.

Table 1. Identification rates on Honda/UCSD dataset

		MMD[17]	AHISD[2]	CHISD[2]	SANP[6]	Our Method
(a)	1st Test Set: 50 frames	87.18%	89.74%	89.74%	92.31%	**100%**
	1st Test Set: 100 frames	**100%**	97.44%	**100%**	**100%**	**100%**
	1st Test Set: Full length	**100%**	**100%**	**100%**	**100%**	**100%**
(b)	1st Test Set: With falsely detected objects	92.31%	97.44%	97.44%	97.44%	**100%**
(c)	2nd Test Set: Occlusion	11.11%	11.11%	11.11%	44.44%	**88.89%**

Table 2. Average identification rates and the standard deviations of different methods on Youtube Celebrities dataset for ten-fold cross validation experiments

Methods	Average Performance
MMD [17]	$56.99 \pm 2.71\%$
AHISD [2]	$58.44 \pm 3.64\%$
CHISD [2]	$59.22 \pm 3.09\%$
SANP [6]	$53.56 \pm 2.39\%$
Our method	$\mathbf{66.67 \pm 2.30\%}$

4. Experiments

4.1. Datasets

To ensure extensive evaluation of our approach, the experiments are carried out on the following three benchmark datasets:

- **Honda/USCD [9]**: This dataset contains 68 video sequences of 20 different individuals. Among the 68 sequences, 20 are used for training while 48 are used for testing (39 in the first set with variations in head pose and expression; and 9 (out of 11) in the second set with partial face occlusion). Note that in the second test set, Saito's test video and KClee's 2nd test videos are not used due to very short duration of the occlusion event and unavailability of the video, respectively.
- **YouTube Celebrities [7]**: This dataset has 1910 video sequences of 47 individuals collected from YouTube. For each individual, 3 and 6 sequences are randomly chosen for training and testing, respectively. Most of the videos are low resolution and highly compressed.

For both datasets, an image set is extracted from each video sequence by using a cascaded face detector [16], except for the second (occluded) test set in Honda dataset where an incremental tracker [14] is used. The images are grey scaled and resized to 40×40.

4.2. Experimental Settings

Existing Approaches. We compare the proposed local approach with several state-of-the-art image-set-based face recognition approaches in the literatu re. They are

1) Manifold-Manifold Distance (MMD) [17]
2) Affine Hull-based Image Set Distance (AHISD) [2]
3) Convex Hull-based Image Set Distance (CHISD) [2]
4) Sparse Approximated Nearest Point (SANP) [6]

These approaches are implemented according to the source code and optimized parameters provided by the original authors. The images are histogram-equalized to minimize illumination variations before these approaches are implemented. For MMD, PCA is applied to learn linear subspaces and to preserve 95% of energy. The difference between Geodesic distance and Euclidean distance is set to 2. For linear AHISD and CHISD, PCA is used to retain 95% energy. The error penalty in CHISD is set to C=100 for grey-scale features. For SANP, the original weight parameters [6] are adopted for convex optimization.

Our Approach. For all datasets, we have chosen Facial Sparse Descriptor (FSD) [12] to be the MKD of our approach. The parameters $\varepsilon_{desc}, \varepsilon_{keyp}, MinPts, \tau, R$ and M are selected to be 0.3, 6 pixels, $0.1 \times \max_k \sigma\left(\delta_k^{(c)}, p_k^{(c)}\right)$, 0.8, 0.2 and $\frac{N_c}{2}$, respectively, where N_c denotes the maximum number of matched clusters among the comparisons of a probe set with each gallery set.

4.3. Performance Evaluation

Experiments on the Honda dataset are divided into three parts, where the identification rates reported in Table 1 are based on (a) various numbers of frames, (b) full-length frames with the presence of falsely-detected objects, and (c) full-length frames containing partial occluded faces. These three parts of experiments correspond to three different real-world scenarios: Part (a) models the scenario where a tracker fails to track for a long video sequence and only the first part of the sequence is usable for recognition; Part (b) models the scenario where the detector incorrectly detects faces of non-target individuals or non-facial objects in addition to the target face; Part (c) models the unconstrained scenario where partial face occlusion occurs.

It is noticed that our approach consistently outperforms the state-of-the-art approaches in all three settings. Our approach achieves perfect identification rate for all settings in part (a). Without manual exclusion of falsely-detected objects in part (b), it is observed that the performance of most existing approaches deteriorate a little with reference to full-length setting in part (a). On the contrary, our approach remains performing ideally when such false

detections appear inconsistently in each image set, justifying the ability of our approach in handling inconsistent-appearing "outliers". In part (c) where most face images in the probe image set are partially occluded, our approach successfully recognizes 8 out of 9 probe image sets, yielding 88.89% identification rate. This is rather significant compared to the second-ranked SANP that recognizes 4 probe image sets correctly (44.44%), since it is known that sparse-representation-based approaches are quite robust to occlusion. The remaining holistic approaches are generally found to be less capable of recognizing partially-occluded images due to their heavily-distorted image set representation/structure when occlusion occurs.

Table 2 illustrates the average identification rates over ten-fold cross validations on the challenging Youtube Celebrities dataset. Although much lower identification rates are obtained by all evaluated approaches, it is shown that our approach achieves the best performance among them, thus justifying the ability of our approach in handling low quality images captured from the real-world videos.

5. Conclusion

In this paper, we have proposed a local approach for image-set-based face recognition under uncontrolled condition, specifically to address partial face occlusion and face misalignment problems. Our approach extracts alignment-free keypoint descriptors from each image and represents an image set with a set of clustered descriptors (corresponding to consistent keypoints) via fast density-based clustering. By identifying the corresponding clusters between two cluster sets, the set similarity of our approach is measured by a cluster-to-cluster distance. To prevent matching score being significantly affected by partial face occlusion, our approach ensures that probe clusters that do not correspond to any of the gallery clusters will not be considered in our distance computation. Experimental evaluations on two benchmark datasets involving naturally misaligned and partially occluded images revealed that our approach achieves the best overall performance among the state-of-the-art approaches.

Acknowledgment

This project is partially supported by the Earmarked RGC grant HKBU 211612, Science Faculty Research Grant of Hong Kong Baptist University, National Natural Science Foundation of China grants 61172136 and 61128009, and the 12th Five-year Plan China S & T Supporting Programme (No. 2012BAK16B06).

References

[1] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. CVPR, pp. 581–588, 2005.

[2] H. Cevikalp and B. Triggs. Face recognition based on image sets. CVPR, pp. 2567–2573, 2010.

[3] Y.-C. Chen, V.M. Patel, P.J. Phillips, and R. Chellapa. Dictionary-based face recognition from video. ECCV, pp. 766-779, 2012.

[4] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD, 1996.

[5] J. Hamm and D. D. Lee. Grassmann discriminant analysis: A unifying view on subspace-based learning. ICML, 2008.

[6] Y. Hu, A.S. Mian, and R. Owens. Sparse Approximated Nearest Points for Image Set Classification. CVPR, 2011.

[7] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. CVPR, pp. 1–8, 2008.

[8] T.K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. IEEE Trans. on PAMI, 29(6):1005–1018, 2007.

[9] K.C. Lee, J. Ho, M.H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. CVPR, pp. 313–320, 2003.

[10] Z. Li, J. Imai, and M. Kaneko. Robust face recognition using block-based bag of words. ICPR, 2010.

[11] S. Liao, A.K. Jain and S.Z. Li. Partial Face Recognition: Alignment Free Approach. IEEE Transactions on PAMI, 2012 (in press).

[12] N. Liu, J.-H. Lai, and W.-S. Zheng. A facial sparse descriptor for single image based face recognition. Neurocomputing 93: 77–87, 2012.

[13] D.G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60:91–110, 2004.

[14] D.A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. IJCV, 77(1-3):125–141, 2008.

[15] D. Sun and Z. Qiu, Bag-of-words vector quantization based face identification. ISECS, vol. 2, pp. 29–33, 2009.

[16] P. Viola and M. Jones. Robust real-time face detection. IJCV, 57(2):137-154, 2004.

[17] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. CVPR, pp. 2940–2947, 2008.

[18] R. Wang, H. Guo, L. Davis and Q. Dai. Covariance Discriminative Learning: A Natural and Efficient Approach to Image Set Classification. CVPR pp. 2496–2503, 2012.

[19] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. IEEE Transactions on PAMI, 31(2):210–227, 2009.