

# A Proximity-Aware Hierarchical Clustering of Faces

Wei-An Lin, Jun-Cheng Chen, and Rama Chellappa

University of Maryland, College Park

walin@terpmail.umd.edu, pullpull@cs.umd.edu, rama@umiacs.umd.edu

**Abstract**—In this paper, we propose an unsupervised face clustering algorithm called “Proximity-Aware Hierarchical Clustering” (PAHC) that exploits the local structure of deep representations. In the proposed method, a similarity measure between deep features is computed by evaluating linear SVM margins. SVMs are trained using nearest neighbors of sample data, and thus do not require any external training data. Clusters are then formed by thresholding the similarity scores. We evaluate the clustering performance using three challenging unconstrained face datasets, including Celebrity in Frontal-Profile (CFP), IARPA JANUS Benchmark A (IJB-A), and **JANUS Challenge Set 3 (JANUS CS3)** datasets. Experimental results demonstrate that the proposed approach can achieve significant improvements over state-of-the-art methods. Moreover, we also show that the proposed clustering algorithm can be applied to curate a set of large-scale and noisy training dataset while maintaining sufficient amount of images and their variations due to nuisance factors. The face verification performance on JANUS CS3 improves significantly by finetuning a DCNN model with the curated MS-Celeb-1M dataset which contains over three million face images.

## I. INTRODUCTION

In this paper, we address the problem of face clustering, especially for the scenario of grouping a set of face images without knowing the exact number of clusters. Face clustering algorithms provide meaningful partitions for given face image sets by combining faces with similar appearances while separating dissimilar ones. Ideally, face images in a partition should belong to the same identity, while images from different partitions should not. Identity-sensitive face clustering algorithms is an active research in computer vision and has several applications, including but not limited to organizing personal pictures, summarizing imagery from social media, and homeland security camera during investigation. Clustering is also important when we need large amount of data to train a deep convolutional neural network (DCNN) for face verification, classification, or detection tasks. Recently, Microsoft Research released the MS-Celeb-1M dataset [8], which contains 1M celebrity names and over 8 million face images. Due to its diversity, this very-large dataset has the potential to improve the performance of face recognition systems. However, since the MS-Celeb-1M dataset has been built from the outputs of search engines, labeling errors could adversely affect the training of deep networks. An effective approach to tackle this problem is to apply a reliable clustering algorithm on the MS-Celeb-1M training dataset to harvest sufficient number of face images that can be used for training a DCNN.

Despite extensive studies on general clustering algorithms over the past few decades, face image clustering remains a difficult task. The difficulties are mainly two-fold. Since face images of a person may have variations in illumination, facial expressions, occlusion, age, and pose, it is challenging to measure the similarity between two face images. Another issue is that without knowing the actual number of clusters, many well-established clustering algorithms, such as  $k$ -means, may not be effective.

Recent advances on DCNNs have brought about impressive improvements for image classification and verification tasks [23], [11], which can be attributed to its ability to extract discriminative information from each image and represent it compactly. Inspired by this progress, we apply a DCNN to extract deep features from the given face and define a similarity measure to separate one face from another. Traditional methods define pairwise similarity based on monotonically decreasing functions of distance, *e.g.* the Gaussian kernel  $\exp(-d(\mathbf{x}_i, \mathbf{x}_j)^2/\sigma^2)$ . Recently, Zhu *et al.* [30] proposed Rank-Order clustering where pairwise similarity is measured based on the ranking of shared nearest neighbors. Otto *et al.* [17] improved the scalability and accuracy of Rank-Order clustering by considering only the presence or absence of nearest neighbors. We hypothesize, based on the works by Zhu and Otto, that neighborhood geometry should be considered to achieve improved clustering performance. However, Rank-Order clustering computes similarities based on shared nearest neighbors in a domain where geometrical information may be lost (*i.e.*, ‘rank’ only contains the ordering information). Our approach measures the similarity between neighborhoods directly in the feature space: neighborhood geometries are first transferred to an evaluation hyperplane, pairwise similarity is then defined by evaluating the points on the hyperplanes.

The rest of the paper is organized as follows. We present our algorithm in Section III. Qualitative and quantitative evaluations are conducted in Section IV. Finally, conclusions are given in Section V.

## II. RELATED WORK

**General Clustering Algorithms** Clustering algorithms can be generally categorized into partitional and agglomerative approaches. Both approaches build upon a similarity graph  $G(V, E)$  defined for the given dataset. The graph can be either fully connected, in  $\epsilon$ -neighborhood or in  $k$ -nearest neighbor. For partitional approaches, given the number of clusters,  $k$ -means [15] iteratively updates the group cen-

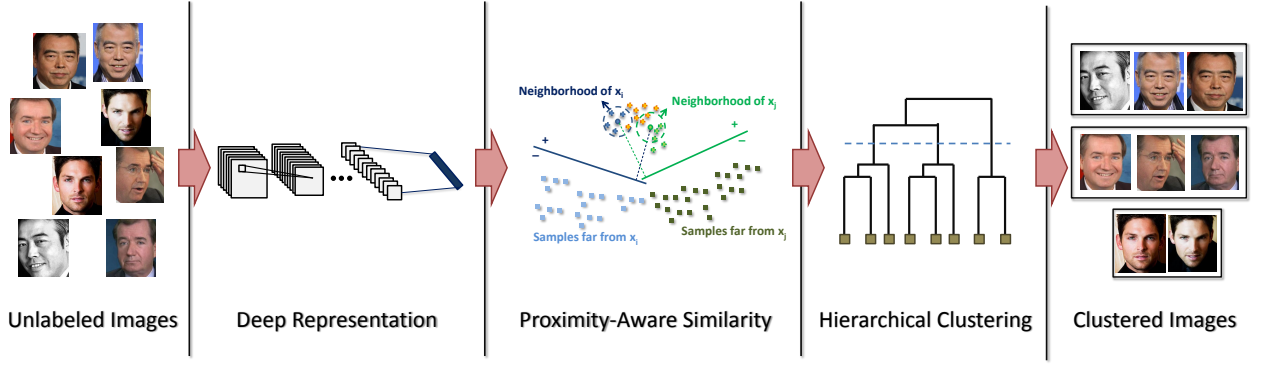


Fig. 1. Overall pipeline for the proposed PAHC algorithm. Unlabeled face images are preprocessed and passed through a DCNN to obtain deep features. The Proximity-Aware similarity between each pair of features is then computed. Based on the Proximity-Aware similarity, hierarchical clustering is applied to yield the final image clusters.

ters and corresponding members until convergence. Spectral clustering finds the underlying structure based on graph Laplacian [22], [16], [28]. For agglomerative approaches [7], [12], groups of data points are merged whenever the linkage between them is above some threshold. Finding the proper similarity measure is one of the major tasks in designing clustering algorithms. Traditional approaches use non-increasing functions of pairwise distance as the similarity measure, *e.g.*  $\exp(-d(\mathbf{x}_i, \mathbf{x}_j)^2/\sigma^2)$ . Recently, sparse subspace clustering (SSC) [3], [4] and low-rank subspace clustering (LRSC) [24], [14], which exploit the subspace structures in a dataset, have gained some attention. Both methods assume data points are self-expressive. By minimizing the reconstruction error under the sparsity/low-rank criterion, the similarity matrix can be obtained from the corresponding sparse/low-rank representation. However, SSC and LRSC are computationally expensive and hard to scale. In [18], dimensionality reduction and subspace clustering are simultaneously learned to achieve improved performance and efficiency. Another category, known as supervised clustering, learns appropriate distance metric from additional datasets [9], [11], [6], [13].

**Image Clustering Algorithms** Yang *et al.* [26] proposed learning deep representations and image clusters jointly in a recurrent framework. Each image is treated as separate clusters at the beginning, and a deep network is trained using this initial grouping. Deep representation and cluster members are then iteratively refined until the number of clusters reached the predefined value. Zhang *et al.* [29] proposed to cluster face images in videos by alternating between deep representation adaption and clustering. Temporal and spatial information between and within video frames is exploited to achieve high purity face image clusters. Zhu *et al.* [30] measured pairwise similarity by considering the ranks of shared nearest neighbors, and transitively merged images into clusters when the similarity is above some threshold. Otto *et al.* [17] modified the algorithm by (i) using deep representations of images (ii) considering only the absence and presence of the shared nearest neighbors

and (iii) transitively merging only once. Superior clustering results and computational time are achieved from the modifications. Sankaranarayanan *et al.* [20] proposed learning a low-dimensional discriminative embedding for deep features and applied hierarchical clustering to realize state-of-the-art precision-recall clustering performance on the LFW dataset.

Different from these studies, we propose a clustering algorithm that does not require (i) training a deep network iteratively [26] (ii) partial identity information [29] and (iii) additional training data [20]. Our approach focuses on exploiting the neighborhood structure between samples and implicitly performing domain adaptation to achieve improved clustering performance.

### III. PROPOSED APPROACH

In this section, we introduce our clustering algorithm, illustrated in Fig. 1. The face images first pass through a pre-trained face DCNN model to extract the deep features. Then, we compute the Proximity-Aware similarity scores using linear SVMs trained with corresponding neighborhoods of the samples. Finally, the agglomerative hierarchical clustering method is applied on the similarity scores to determine the cluster labels to each sample. The details of each component are described in the following subsections.

#### A. Notation

We denote the set of face images as  $I = \{I_1, \dots, I_{n_s}\}$ . Our goal is to assign labels  $L = \{l_1, \dots, l_{n_s}\}$  for each image to indicate the cluster it belongs to. The images are first passed through a pre-trained DCNN model to extract the deep features, which are then normalized to unit length. Specifically, let  $f_\theta : \mathcal{I} \rightarrow \mathcal{X}$  be the DCNN network parameterized by  $\theta$ , and  $g : \mathcal{X} \rightarrow \mathcal{X}$  be the normalization. The corresponding deep representations for the face images are given by  $X = g \circ f_\theta(I) = \{\mathbf{x}_1, \dots, \mathbf{x}_{n_s}\}$ . For each representation  $\mathbf{x}_i$ , we define  $\mathcal{N}_K(\mathbf{x}_i)$  as the set of  $K$ -nearest neighbors of  $\mathbf{x}_i$ , including  $\mathbf{x}_i$  itself.

### B. Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering [7], [12] initializes all samples as separate clusters. Based on the pairwise distance matrix  $\mathbf{D}$  measured from the features, clusters are iteratively merged whenever the cluster-to-cluster distance is below some threshold  $\eta$ . The hierarchical clustering algorithm, denoted as  $\text{Hierarchical}(\mathbf{D}, \eta)$ , generates the cluster assignments  $L$  for all the faces in  $I$ . In our work, we use average linkage as a measure of cluster-to-cluster distance.

Recent advances in DCNN have yielded great improvements for face verification task, which uses cosine distance as the similarity measure to decide whether two faces belong to the same subject. Given two features  $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$  on the unit hypersphere  $\{\mathbf{x} : \|\mathbf{x}\| = 1\}$ , the similarity measure between them is computed by

$$s(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j. \quad (1)$$

The pairwise distance matrix  $\mathbf{D}$  in this case is simply

$$[\mathbf{D}]_{i,j} = 1 - s(\mathbf{x}_i, \mathbf{x}_j). \quad (2)$$

Since DCNNs trained on large datasets extract discriminative features for images,  $\text{Hierarchical}(\mathbf{D}, \eta)$ , where  $\mathbf{D}$  is from (2), can perform well on datasets that have similar distribution as the training dataset. However, the difference in distribution encountered in many real-world applications degrades the performance significantly. Inspired by previous works [30], [17], we aim at measuring similarity based on the neighborhood structure.

### C. Proximity-Aware Similarity

To have a formulation that is able to take neighborhoods  $\mathcal{N}_K(\mathbf{x}_i), \mathcal{N}_K(\mathbf{x}_j)$  into account when measuring the similarity between  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , we rewrite the inner product as

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^T \mathbf{x}_j + \mathbf{x}_j^T \mathbf{x}_i}{2}. \quad (3)$$

In (3), the similarity between  $\mathbf{x}_i$  and  $\mathbf{x}_j$  is evaluated by averaging two asymmetric measures: How similar is  $\mathbf{x}_j$  from the view of  $\mathbf{x}_i$  and how similar is  $\mathbf{x}_i$  from the view of  $\mathbf{x}_j$ . Specifically,  $\mathbf{x}_i^T \mathbf{x}_j$  can be interpreted as evaluating  $\mathbf{x}_j$  on hyperplane  $H_i = \{\mathbf{x} : \mathbf{x}_i^T \mathbf{x} = 0\}$  and  $\mathbf{x}_j^T \mathbf{x}_i$  can be interpreted as evaluating  $\mathbf{x}_i$  on hyperplane  $H_j = \{\mathbf{x} : \mathbf{x}_j^T \mathbf{x} = 0\}$ . This observation allows us to generalize the asymmetric measure as follows.

Given a hyperplane  $H_{\mathbf{w}_i, b_i} = \{\mathbf{x} : \mathbf{w}_i^T \mathbf{x} + b_i = 0\}$  which contains information about  $\mathcal{N}_K(\mathbf{x}_i)$ , the asymmetric similarity from  $H_{\mathbf{w}_i, b_i}$  to some set  $S$  is defined as

$$H_{\mathbf{w}_i, b_i}(S) = \frac{1}{|S|} \sum_{\mathbf{x} \in S} [\mathbf{w}_i^T \mathbf{x} + b_i]. \quad (4)$$

Following (3), the generalized similarity measure, which we call “Proximity-Aware similarity”, is the average of two asymmetric measures from  $H_{\mathbf{w}_i, b_i}$  to  $\mathcal{N}_K(\mathbf{x}_j)$  and from  $H_{\mathbf{w}_j, b_j}$  to  $\mathcal{N}_K(\mathbf{x}_i)$ :

$$s_{PA}(\mathbf{x}_i, \mathbf{x}_j) = \frac{H_{\mathbf{w}_i, b_i}(\mathcal{N}_K(\mathbf{x}_j)) + H_{\mathbf{w}_j, b_j}(\mathcal{N}_K(\mathbf{x}_i))}{2}. \quad (5)$$

Unlike cosine similarity,  $s_{PA}$  is not bounded. We introduce a nonlinear transformation to define the Proximity-Aware pairwise distance

$$[\mathbf{D}_{PA}]_{ij} = 1 - \frac{2}{\pi} \arctan [s_{PA}(\mathbf{x}_i, \mathbf{x}_j)]. \quad (6)$$

This choice of nonlinearity is for experimental simplicity. One can also consider  $[\mathbf{D}_{PA}]_{i,j} = \exp(-s_{PA}(\mathbf{x}_i, \mathbf{x}_j))$ . The Proximity-Aware Hierarchical clustering is then characterized by the following algorithm:

$$L_{PA} \leftarrow \text{Hierarchical}(\mathbf{D}_{PA}, \eta). \quad (7)$$

The above construction helps us to cast the problem of defining the similarity function between neighborhoods into finding hyperplanes  $H_{\mathbf{w}_i, b_i}$ . Our ultimate goal is to find a similarity measure for each pair of feature vectors that reflects whether they belong to the same class. We conjecture that  $H_{\mathbf{w}_i, b_i}$  and  $H_{\mathbf{w}_j, b_j}$  should have the following property:

*$H_{\mathbf{w}_i, b_i}(\cdot)$  has a large value when evaluating on sets that are near  $\mathcal{N}_K(\mathbf{x}_i)$ , and has a small value otherwise.*

The constraint not only forces the similarity measure to be locally geometry-sensitive (proximity-aware) but also adaptive to the data domain. This motivates the use of linear classifiers to separate positive samples  $\mathcal{N}_K(\mathbf{x}_i)$  from their corresponding negative samples. Fig. 2 shows a demonstrative example. This approach is analogous to the one-shot similarity technique [25]. In this work, we use the linear SVM as our candidate algorithm for finding hyperplanes. Specifically, we solve

$$\min_{\mathbf{u}} \frac{1}{2} \mathbf{u}^T \mathbf{u} + C_p \sum_{k=1}^{N_p} \max[0, 1 - y_k \mathbf{u}^T \mathbf{z}_k]^2 + C_n \sum_{k=1}^{N_n} \max[0, 1 - y_k \mathbf{u}^T \mathbf{z}_k]^2, \quad (8)$$

where  $\mathbf{u} = [\mathbf{w}^T \ b]^T$  and  $\mathbf{z}_k = [\mathbf{x}_k^T \ 1]^T$ . We treat  $\mathcal{N}_K(\mathbf{x}_i)$  as positive samples with cardinality  $N_p$ , and a subset of  $X \setminus \mathcal{N}_K(\mathbf{x}_i)$  as negative samples with cardinality  $N_n$ .  $y_k = +1$  for positive samples and  $y_k = -1$  for negative samples. The regularization constants  $C_p$  and  $C_n$  are given by  $C_p = C \frac{N_p + N_n}{N_p}$  and  $C_n = C \frac{N_p + N_n}{N_n}$ .

In [25], Linear Discriminant Analysis (LDA) is used as the classifier to evaluate one-shot similarity score. However, we do not consider LDA as our candidate because the bimodal Gaussian prior assumption is not always satisfied for the positive and negative samples drawn from real-world datasets. In the proposed method, negative samples often consist of features from different identities with variations from nuisance factors, which do not obey a single Gaussian distribution.

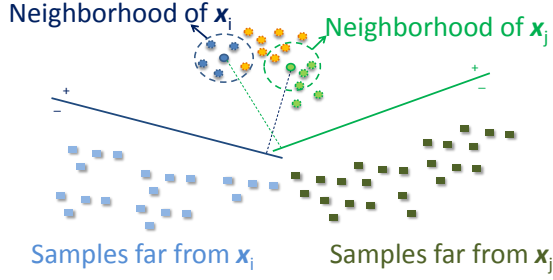


Fig. 2. Proximity-Aware similarity. Circles in blue, yellow, green represent samples with different identities. Blue dashed circles delineate the neighborhood of  $\mathbf{x}_i$  (or  $\mathcal{N}_G(\mathbf{x}_i)$ ) while green dashed circles delineate the neighborhood of  $\mathbf{x}_j$  (or  $\mathcal{N}_G(\mathbf{x}_j)$ ). The blue hyperplane is obtained by solving (8), treating  $\mathcal{N}_K(\mathbf{x}_i)$  as positive samples, and a subset of  $X \setminus \mathcal{N}_K(\mathbf{x}_i)$ , which are blue squares in this case, as negative samples. The green hyperplane is obtained in the same way. The Proximity-Aware similarity between  $\mathbf{x}_i$  and  $\mathbf{x}_j$  is evaluated using (5). The length of the blue dashed line and the green dashed line reflects how similar are the two neighborhoods.

#### D. Choice of Positive and Negative Sets

Since the hyperplane is chosen based on large-margin classification between positive samples  $\mathcal{N}_K(\mathbf{x}_i)$  and negative samples, the choice of them would be crucial. In this paper, we first construct the nearest neighbor list  $\text{NNList}_{\mathbf{x}_i}$  for each data sample  $\mathbf{x}_i$ , where  $\text{NNList}_{\mathbf{x}_i}[1] = \mathbf{x}_i$ .  $\mathcal{N}_K(\mathbf{x}_i)$  corresponds to  $\text{NNList}_{\mathbf{x}_i}[1 : K]$ , and we choose  $\text{NNList}_{\mathbf{x}_i}[N_1 : N_2]$  as the negative samples. In Section IV, we show how parameters  $(K, N_1, N_2)$  affect the clustering performance in detail.

### IV. EXPERIMENTAL RESULTS

In this section, we evaluate our clustering algorithm qualitatively on the recently released MS-Celeb-1M [8] dataset and quantitatively on the IARPA Janus Benchmark-A (IJB-A), JANUS Challenge Set 3 (CS3), and Celebrities in Frontal-Profile (CFP) datasets. To compute Proximity-Aware similarity, we use the LIBLINEAR library [5] with L2-regularized L2-loss primal SVM. The parameter  $C$  is set at 10 throughout this section.

#### MS-Celeb-1M [8]:

Microsoft Research recently released this very large face image dataset, consisting of 1M identities. The training dataset of MS-Celeb-1M is prepared by selecting top 99,892 identities from the 1M celebrity list. There are 8,456,240 images in total, roughly 85 images per identity. This dataset is designed by leveraging a knowledge base called “freebase”. Since face images are created using a search engine, labeling noise may be a problem when this dataset is used in supervised learning tasks. We demonstrate the effectiveness of the proposed clustering algorithm in curating large-scale noisy dataset in Section IV-C and Section IV-F. In this paper, we directly use the aligned images provided along with the dataset.

#### Celebrities in Frontal-Profile (CFP) [21]:

This dataset contains 500 subjects and 7,000 face images. Of the 7,000 faces, 5,000 are in frontal view, and the remaining 2,000 are in profile views where each subject contains 10 frontal and 4 profile images. Unlike the IJB-A dataset, the CFP dataset aims at isolating the factor of pose variation in order to facilitate research in frontal-profile face verification. Extreme variations in poses can be seen in Fig. 3. In this work, we apply our clustering algorithm on all 7,000 face images.

#### IARPA Janus Benchmark A (IJB-A) [10] and JANUS Challenge Set 3 (CS3):

The IJB-A dataset contains 500 subjects with a total of 25,813 images taken from photos and video frames (5,399 still images and 20,414 video frames). Extreme variations in illumination, resolution, viewpoint, pose and occlusion make it a very challenging dataset. In this work, we cluster the templates corresponding to the query set for each split in IJB-A 1:1 verification protocol where a template is composed of a combination of still images and video frames. The CS3 dataset is a superset of IJB-A dataset which contains 11,876 still images and 55,372 video frames sampled from 7,094 videos. In this paper, we use images and video frames provided in CS3 1:1 verification protocol. There are totally 1,871 subjects and 12,590 templates. Fig. 4 shows sample images from different templates.

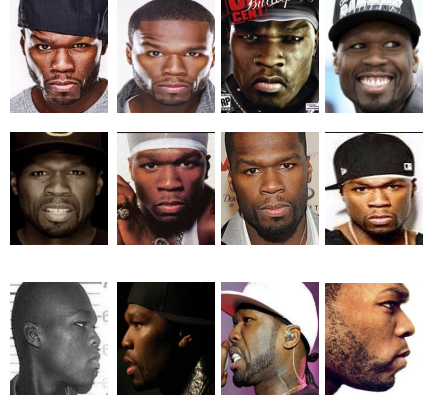


Fig. 3. Sample images in CFP dataset. The first two rows are frontal face images and the last row consists of profile face images.

#### A. Preprocessing

All the face bounding boxes and fiducial points for the IJB-A, and JANUS CS3 datasets are extracted using the multi-task face and fiducial detector of Hyperface [19]. For the MS-Celeb-1M and CFP datasets, we use the fiducial points or aligned images provided with the dataset. Each face is aligned into the canonical coordinate with the similarity transform using seven landmark points: two left eye corners, two right eye corners, nose tip, and two mouth corners. For the CFP dataset, since fiducial points are only available for half face, we use two eye corners of one of eyes, nose tip, and



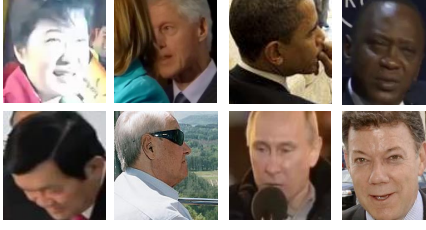


Fig. 4. Sample images in CS3 dataset. The faces contain extreme illumination, viewpoint, pose, and occlusion changes.

one of mouth corners. In training phase, data augmentation is performed by randomly cropping and horizontally flipping face images.

### B. Deep Network and Image Representation

We implement the network architecture presented in [2] and train it using the CASIA-WebFace dataset [27]. We preprocess this dataset using the steps presented in Section IV-A. We denote this pretrained network as ‘DCNN<sub>face</sub> (CASIA)’. This network is further finetuned with curated MS-Celeb-1M dataset [8], which we denote as ‘DCNN<sub>face</sub> (CASIA+MSCeleb)’. The process of removing mislabeled images in MS-Celeb-1M will be introduced in the next section. DCNN<sub>face</sub> (CASIA) is trained using SGD for 780K iterations with a standard batch size 128 and momentum 0.9. The learning rate is set to 1e-2 initially and is halved every 100K iterations. The weight decay rates of all the convolutional layers are set to 0, and the weight decay of the final fully connected layer is set to 5e-4. Pretraining the DCNN network with the CASIA dataset not only provides good initialization for the model parameters but also greatly reduces the training time on the curated MS-Celeb-1M dataset. Then, we finetune the pretrained network to obtain DCNN<sub>face</sub> (CASIA+MSCeleb) for improved face representation. We use the learning rate 1e-4 for all the convolutional layers, and 1e-2 for the fully connected layers. The network is trained for 240K iterations. In the training phase of DCNN<sub>face</sub> (CASIA) and DCNN<sub>face</sub> (CASIA+MSCeleb), the dropout ratio is set as 0.4 to regularize fc6 due to the large number of parameters (*i.e.*  $320 \times 10503$  for the CASIA dataset and  $320 \times 58207$  for the curated MS-Celeb-1M dataset.). Note that we manually remove the overlapping subjects with the IJB-A and JANUS CS3 datasets from the CASIA-WebFace and the MS-Celeb-1M datasets.

The inputs to the networks are  $100 \times 100 \times 3$  RGB images. Given a face image, deep representation is extracted from the pool5 layer with dimension 320. In the case of IJB-A and CS3 datasets, if there are multiple images and frames in one template, we perform media average pooling to produce the final representation.

### C. Qualitative Study on MS-Celeb-1M

As a qualitative study, we apply the clustering algorithm to remove face images with noisy labels in

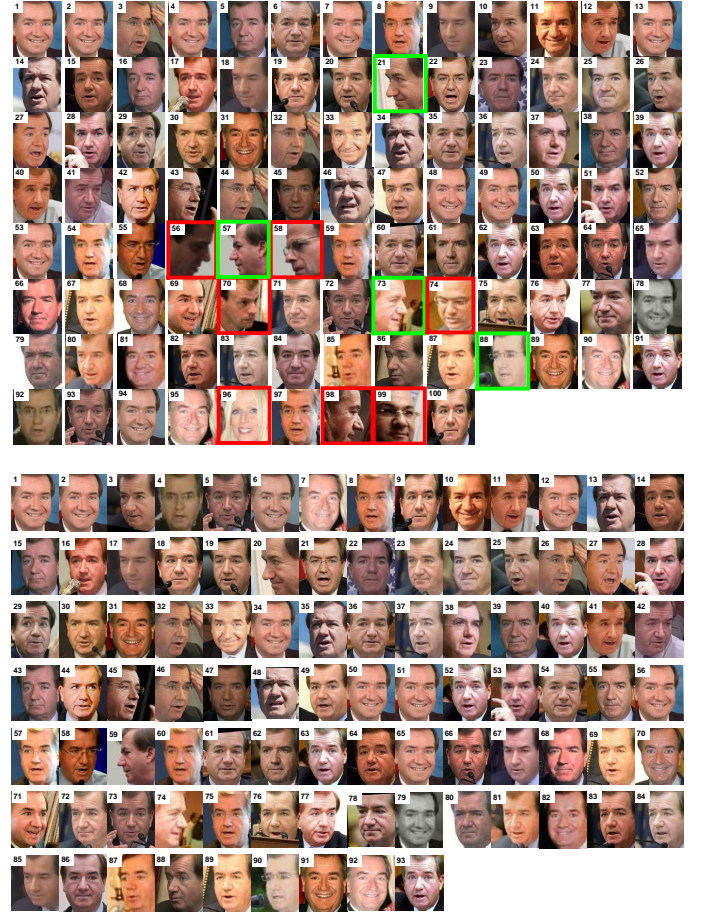


Fig. 5. Sample face images in the MS-Celeb-1M dataset with improved purity after applying the PAHC. Upper-half of the figure shows original face images having machine identifier m.024xcy in MS-Celeb-1M dataset. The lower half of the figure is obtained following the process described in Fig. IV-C. The red boxes are the face images removed by our algorithm. The green boxes are face images that are retained by our algorithm. The variation in extreme pose (*e.g.* 21, 57, 73) and resolution (*e.g.* 88) will assist the DCNN to learn improved representation.

MS-Celeb-1M training dataset. Feature representation is first obtained by passing the whole dataset through DCNN<sub>face</sub>(CASIA) described in Section IV-B. We divide the total 99,892 identities into batches with size 50. For each batch, we apply Hierarchical( $\mathbf{D}_{PA}, 2, 3$ ), with  $[\mathbf{D}_{PA}]_{i,j} = \exp(-s_{PA}(\mathbf{x}_i, \mathbf{x}_j))$ . Clusters whose majority identity have less than 30 images are discarded. The number of the curated dataset is about 3.5 millions face images of 58,207 subjects. Fig. 5 shows one example of the clustering results. Since the PAHC exploits local property, face images with extreme pose are not discarded. In addition, our approach does not require external training dataset.

### D. Quantitative Study on the CFP, IJB-A, and CS3 datasets

Images in CFP, IJB-A, and CS3 datasets are first processed as described in Section IV-A. In this section, we aim to compare our clustering algorithm with traditional hierarchical clustering, *k*-means, and Approximate Rank-Order clustering [17], on the three datasets described earlier. Throughout our experiments, ‘Approximate Rank-Order clustering’ refers

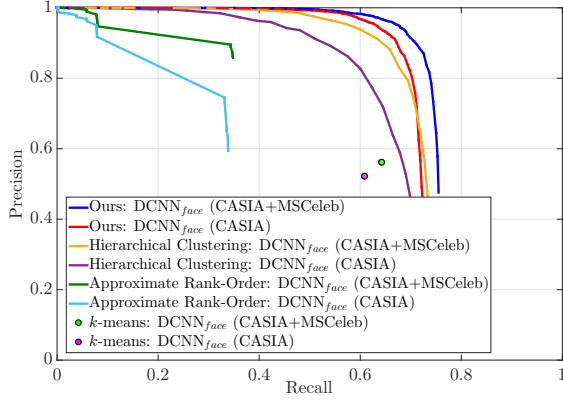


Fig. 6. Precision-Recall curve evaluated on the CFP dataset.  $(K, N_1, N_2) = (5, 50, 100)$  for the PAHC algorithm, where the parameters are as defined in Section III-D

to our implementation of the algorithm proposed in [17]. We use the precision-recall curve defined in [17] as the performance metric to compare different algorithms at all operation points. *Pairwise precision* is the fraction of the number of pairs within the same cluster which are of the same class, over the total number of same-cluster pairs. *Pairwise recall* is the fraction of the number of pairs within a class which are placed in the same cluster, over the total number of same-class pairs.

For the CFP dataset, we cluster 7,000 images. For the IJB-A dataset, we cluster the query set provided in the 1:1 verification protocol for each split, and compute the average performance over 10 splits. For the CS3 dataset, we apply the clustering algorithms on 10,718 probe templates. We use the standard MATLAB implementation for hierarchical clustering and  $k$ -means, where we choose  $k$  as the true number of identities. Fig. 6, Fig. 7, and Fig. 8 show the precision-recall performance comparisons, and Fig. 9 and Fig. 10 show some sample clustering results.

Compared to hierarchical clustering based on cosine distance, the PAHC attains significant gains by exploiting the neighborhood structure for each sample. The Approximate Rank-Order clustering cannot reach recall greater than 0.7 because it *only computes distances between samples which share a nearest neighbor*, which means there are some samples which will not be merged for any choice of thresholding.

#### E. Parameter and Negative Set Study

Fig. 11, 12, and 13 show results for different parameters settings  $(K, N_1, N_2)$  of the proposed algorithm on the CFP, IJB-A, and CS3 datasets. For smaller values of neighborhood size  $K$ , e.g.  $K = 5$ , the choice of negative sets have little effect on the performance. This is because  $\mathcal{N}_K(\mathbf{x}_i)$  may not be able to represent the ‘local’ structure for large  $K$ . In this case, the similarity between  $\mathcal{N}_K(\mathbf{x}_i)$  and  $\mathcal{N}_K(\mathbf{x}_j)$  would deviate significantly from the similarity between  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . According to our experiments, setting  $K = 5$  gives the best performance.

Since we claim that by choosing  $\text{NNList}_{\mathbf{x}_i}[N_1 : N_2]$  as negative samples, the deep representation can be adapted to

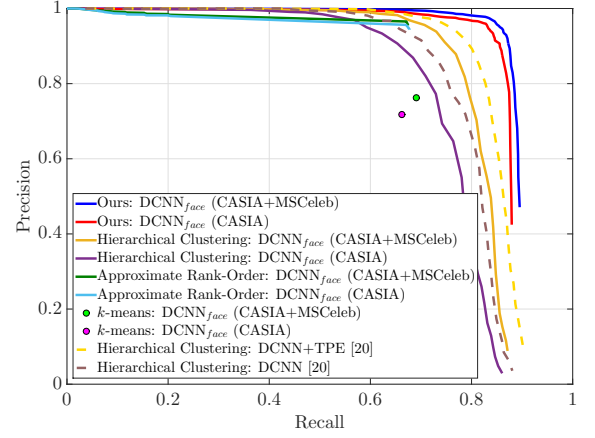


Fig. 7. Precision-Recall curve evaluated on the IJB-A dataset.  $(K, N_1, N_2) = (5, 50, 100)$  for the PAHC algorithm, where the parameters are as defined in Section III-D. It can be observed that the PAHC algorithm outperforms TPE in [20] by a large margin without the need of external training dataset.

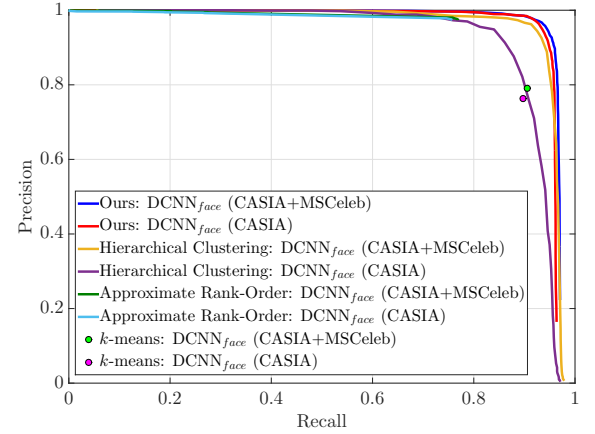


Fig. 8. Precision-Recall curve evaluated on the CS3 dataset.  $(K, N_1, N_2) = (5, 50, 100)$  for the PAHC algorithm, where the parameters are as defined in Section III-D.

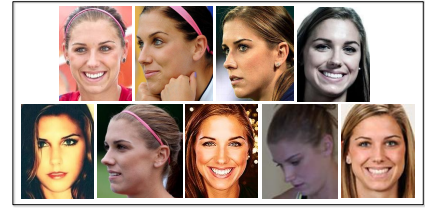


Fig. 9. One sample cluster for the CFP dataset after applying the PAHC algorithm.

the target domain, in the following, we experiment with the IJB-A dataset by sampling templates from the training data in each split and use them as negative samples when training the linear SVM. It can be observed from Fig. 14 that when properly labeled templates, which contain nonoverlapping identities with the verification dataset, are used, improved performance is achieved. However, it is not an effective approach to real-world problems since preparing data with nonoverlapping identity with the unseen dataset  $I$  is difficult.



Fig. 10. Sample clusters for the CS3 dataset after applying the PAHC algorithm. Robustness to pose variation can be seen throughout the images. Top row shows robustness to illumination changes. Middle row shows robustness to age and makeup. Bottom row shows robustness to blur and viewpoint changes.

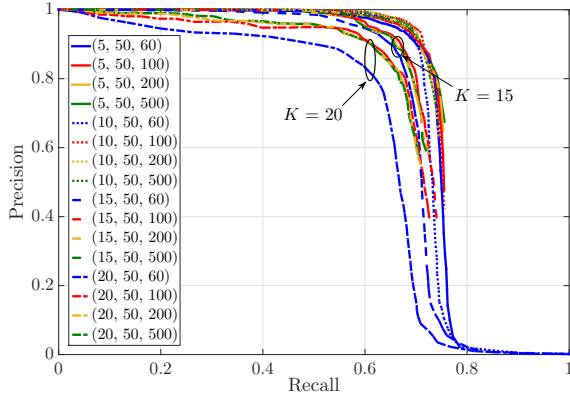


Fig. 11. Precision-Recall curves evaluated on the CFP dataset using  $DCNN_{face}$  (CASIA+MSCeleb) as the image feature extractor. Results are reported by varying  $(K, N_1, N_2)$ , where the parameters are as defined in Section III-D.

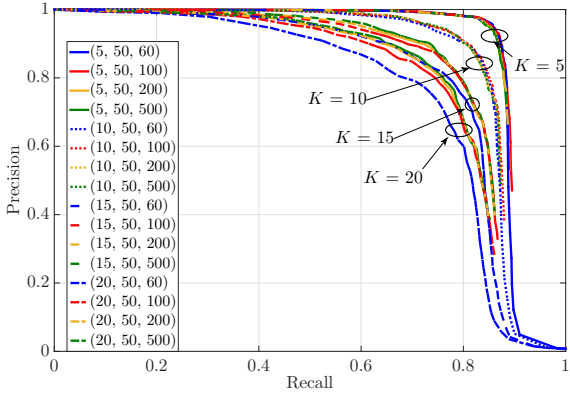


Fig. 12. Precision-Recall curves evaluated on the IJB-A dataset using  $DCNN_{face}$  (CASIA+MSCeleb) as the image feature extractor. Results are reported by varying  $(K, N_1, N_2)$ , where the parameters are as defined in Section III-D.

#### F. Finetuning DCNN using Curated MS-Celeb-1M dataset

As described in Section IV-B, we finetune the pre-trained  $DCNN_{face}$ (CASIA) model using the curated subset of MS-Celeb-1M attained by our clustering algorithm,  $DCNN_{face}$ (CASIA+MSCeleb). In contrast, if we do not

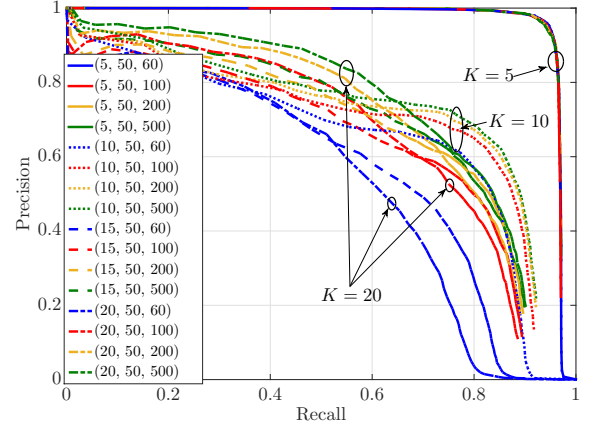


Fig. 13. Precision-Recall curves evaluated on the CS3 dataset using  $DCNN_{face}$  (CASIA+MSCeleb) as the image feature extractor. Results are reported by varying  $(K, N_1, N_2)$ , where the parameters are as defined in Section III-D.

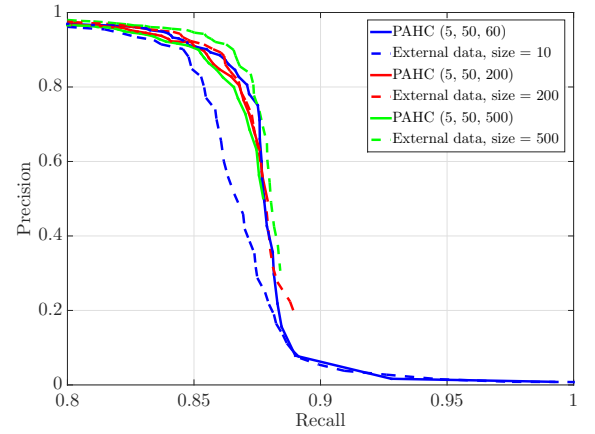


Fig. 14. Precision-Recall curves evaluated on the IJB-A dataset. Solid lines indicate the performance of the PAHC for different parameters  $(K, N_1, N_2)$ . Dashed lines indicate the performance when using a subset of training data as negative samples in a linear SVM.

perform clustering and finetune  $DCNN_{face}$ (CASIA) using all the images of MS-Celeb-1M, the model does not converge. Then, we compare the results of JANUS CS3 1:1 verification task for the two networks:  $DCNN_{face}$  (CASIA) and  $DCNN_{face}$  (CASIA+MSCeleb). From Fig. 15 and Fig. I,  $DCNN_{face}$  (CASIA+MSCeleb) outperforms  $DCNN_{face}$  (CASIA), and it demonstrates that the proposed clustering algorithm improves the quality of training data used for the DCNN.

TABLE I  
CS3 1:1 VERIFICATION PERFORMANCE

FAR	$DCNN_{face}$ (CASIA)	$DCNN_{face}$ (CASIA+MSCeleb)
1e-1	0.9703	0.9731
1e-2	0.8934	0.9184
1e-3	0.7599	0.8355
1e-4	0.5813	0.7252



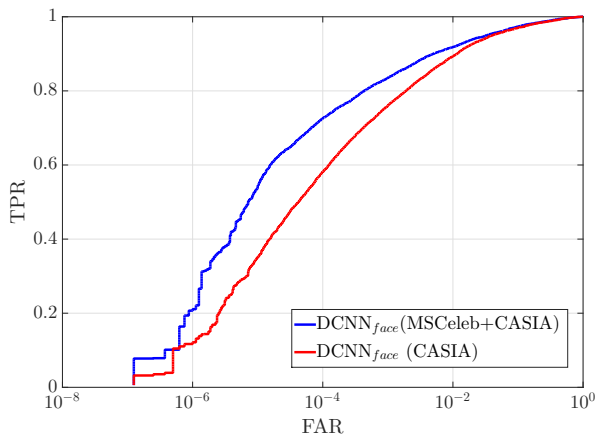


Fig. 15. CS3 1:1 verification performance using  $DCNN_{face}$  trained on CASIA and CASIA+MSCeleb.

## V. CONCLUSION

We proposed an unsupervised algorithm, namely, the PAHC algorithm, to measure the pairwise similarity between samples by exploiting their neighborhood structures along with domain adaptation. From extensive experiments, we show that our clustering algorithm achieves high precision-recall performance at all operation points when the neighborhood is properly chosen. Following this, the PAHC is applied to curate the MS-Celeb-1M training dataset. Our algorithm retains faces with variations in pose, illumination and resolution, while separating images with different identities. We further finetuned the DCNN network with the curated dataset. Significant improvement on CS3 1:1 verification task demonstrates the effectiveness of our algorithm.

## VI. ACKNOWLEDGMENTS

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

## REFERENCES

- [1] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6:937–965, 2005.
- [2] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.
- [3] E. Elhamifar and R. Vidal. Sparse subspace clustering, 2009.
- [4] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.
- [5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

- [6] T. Finley and T. Joachims. Supervised clustering with support vector machines. pages 217–224, 2005.
- [7] K. C. Gowda and G. Krishna. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition*, 10(2):105–112, 1978.
- [8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *European Conference on Computer Vision (ECCV)*, 2016.
- [9] M. I. Jordan, F. R. Bach, and F. R. Bach. Learning spectral clustering. In *Advances in Neural Information Processing Systems*, 2003.
- [10] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
- [12] T. Kurita. An efficient agglomerative clustering algorithm using a heap. *Pattern Recognition*, 24(3):205–209, 1991.
- [13] M. T. Law, Y. Yu, M. Cord, and E. P. Xing. Closed-form training of mahalanobis distance for supervised clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [14] G. Liu and S. Yan. Latent low-rank representation for subspace segmentation and feature extraction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1615–1622, 2011.
- [15] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.
- [16] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2001.
- [17] C. Otto, D. Wang, and A. K. Jain. Clustering millions of faces by identity. *CoRR*, abs/1604.00989, 2016.
- [18] V. M. Patel, H. V. Nguyen, and R. Vidal. Latent space sparse and low-rank subspace clustering. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):691–701, 2015.
- [19] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv:1603.01249*, 2016.
- [20] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016.
- [21] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.
- [22] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [24] R. Vidal and P. Favaro. Low rank subspace clustering (lrscl). *Pattern Recognition Letters*, 43:47–61, 2014.
- [25] L. Wolf, T. Hassner, and Y. Taigman. The one-shot similarity kernel. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.
- [26] J. Yang, D. Parikh, and D. Batra. Joint unsupervised learning of deep representations and image clusters. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [27] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.
- [28] L. Zelnik-manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, pages 1601–1608, 2004.
- [29] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Joint face representation adaptation and clustering in videos. In *European Conference on Computer Vision (ECCV)*, 2016.
- [30] C. Zhu, F. Wen, and J. Sun. A rank-order distance based clustering algorithm for face tagging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 481–488, 2011.