FaceMap: Towards Unsupervised Face Clustering via Map Equation

Anonymous ECCV submission

Paper ID 2630

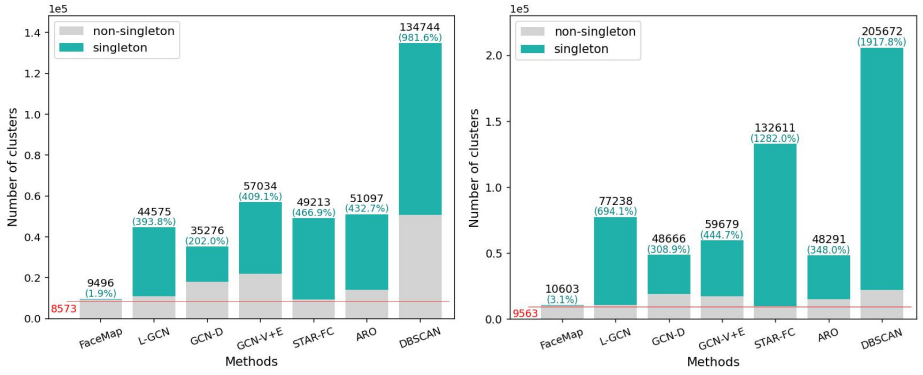
Abstract. Face clustering is an essential task in computer vision due to the explosion of related applications such as augmented reality or photo album management. The main challenge of this task lies in the imperfectness of similarities among image feature representations. Given an existing feature extraction model, it is still an unresolved problem that how can the inherent characteristics of similarities of unlabelled images be leveraged to improve the clustering performance. Motivated by answering the question, we develop an effective unsupervised method, named as FaceMap, by formulating face clustering as a process of non-overlapping community detection under information flows governed by the map equation. Inspired by observations on the ranked transition probabilities in the affinity graph constructed from facial images, we develop an outlier detection strategy to adaptively adjust transition probabilities among images. Experiments with ablation studies demonstrate that FaceMap significantly outperforms existing methods and achieves new state-of-the-arts on three popular large-scale datasets for face clustering, *e.g.*, an absolute improvement of more than 10% and 4% comparing with prior unsupervised and supervised methods respectively in terms of average of Pairwise F-score.

Keywords: face clustering, map equation, graph partitioning

1 Introduction

Face clustering has received considerable attention over the last decade due to the advancement of deep learning in face recognition. It has vast applications for personal photo management and entertainment such as augmented reality effects created by face morphing from multiple photos [27], face tagging in albums [40], video analysis [6], *etc.* It can also be utilized to effectively annotate large-scale facial image datasets [32, 33] or perform label noise cleansing [39].

Face clustering aims at grouping facial images with the same identity into one cluster, while simultaneously discriminating different identities via different cluster labels [29, 33, 34, 36, 37, 39]. The task is easy if we have a perfect face feature extractor, *i.e.*, similarities of face features belonging to the same identity are always much higher than those from different identities. However, it is always not true for real world scenarios. Given a pre-trained feature model and a set of images, a common practice of face clustering is to firstly construct an affinity



(a) Part1(0.58M) in MS1M (b) CASIA

Fig. 1: The comparison on the number of predicted clusters among different clustering methods for datasets of MS1M [11] and CASIA [38]. The red line with number denotes the true number of identities in the dataset. For each method, the total number of predicted clusters and the ratio of singleton cluster are shown. Singleton cluster means that the size of a predicted cluster is one. We clearly observe that, for prior methods, the number of clusters exceeds a lot the true number of identities, and the ratio of singleton cluster is large. Our FaceMap shows superiority in terms of the number of predicted clusters.

graph based on image similarities with k nearest neighbors (kNN), where an image is a node and the cosine similarity between image features is the weight of an edge. The resulting graph is typically noisy with incorrect connections or missing edges, which led to performance degradation for subsequent clustering tasks through graph partitioning, *etc.*

Prior studies with supervised methods train a model with annotated samples to reduce the noise in the affinity graph, and achieve the state-of-the-arts in face clustering [29]. However, the expensive cost of label annotations and the huge hyper-parameter tuning hinder the applications of supervised methods in real-world scenarios. Thus it is worth inventing an unsupervised method that is capable of working with imperfect facial feature representations and achieves outstanding clustering performance.

In addition, performance of face clustering methods is usually evaluated based on Pairwise F-score (F_{P}) [30] and BCubed F-score (F_{B}) [2]. The two traditional metrics are biased toward large-size clusters [2, 19], which grossly neglect the negative impact of incorrect partitions on small-size clusters. Those clusters create lots of burdens for subsequent applications because they misinformed the true number of clusters. Moreover, F_{P} and F_{B} do not have sufficient characterization on the difference between the number of predicted clusters and the true number of identities. However, the corresponding information could be critical for applications that need to understand the true number of unique persons.

For further illustrations, we show the comparison on the number of predicted clusters among different methods in Fig. 1. In prior methods, we clearly observe that the number of resulting clusters could be 10× larger than the true number of identities, and singleton clusters could be the dominant portion. It had been pointed out in [33] that singleton clusters usually contain hard samples, which means the illustrated algorithms actually turn hard examples into singleton clusters to improve performance.

To this end, we study the face clustering problem from two perspectives. On one hand, we formulate unsupervised face clustering as a process of non-overlapping community detection based on the map equation [24], where each identity is an underlying community and the map equation characterizes the entropy of paths by a random walker travelling over the affinity graph. To address the challenge of imperfect feature extractors, we develop a strategy of outlier detection (OD) to adaptively adjust transition probability of a given affinity graph. Our proposed FaceMap method, equipped with the module OD, minimizes the map equation and performs high quality face clustering without using any labeled data. On the other hand, we introduce three new metrics for evaluating face clustering performance which contains key implications about clustering quality in real applications. The new metrics measure identity-level quality, which take incorrect partitions of small-size clusters into account and measure the discrepancy between the number of predicted clusters and the true identity numbers. We also present corresponding studies of state-of-the-art methods with the new metrics.

In this paper, we show that a dedicatedly-designed unsupervised method has the capability of outperforming all the existing state-of-the-art methods. To our best knowledge, face clustering has not been investigated from the viewpoint of non-overlapping community detection. Moreover, there is little study that the map equation is applied into computer vision with significant performance. In summary, we make the following contributions.

- To the best of our knowledge, we are the first to formulate face clustering as community detection with the map equation. We propose an effective method that adaptively adjusts the distribution of transition probability in the affinity graph for face clustering.
- We illustrate the limitations of the traditional metrics for face clustering. For a comprehensive comparison among methods in clustering facial images, we design three metrics for evaluating the identity-level quality.
- The unsupervised FaceMap significantly outperforms the prior unsupervised and supervised methods, and achieves new state-of-the-arts in light of traditional metrics on three large-scale datasets. We also show the superiority of our method in terms of identity-level quality via new metrics.

2 Problem Definition, Metrics and Related Works

In this section, we first present the problem definition of face clustering. Then, we show the limitations of traditional metrics (F_{P} and F_{B}) in this task. We

thereby design three new metrics as complementary evaluations from identity-level quality perspective. Finally, we present related works on this topic.

2.1 Problem Definition

In this paper, we study the problem of face clustering with images. In particular, the input of face clustering is a set of images $X = \{x_i\}_{i=1}^{S}$, where $x_i \in R^d$ denotes the facial feature of an image and d is the dimension of the feature. Face clustering requires an algorithm to produce a set of predicted labels $Y = \{y_i\}_{i=1}^{S}$, where y_i is the predicted label to each image x_i with $y_i \in \{1, 2, \cdots, N\}$, and N denoting the number of predicted clusters. Note that N is unknown and should be determined during clustering. Images with the same predicted label form a cluster. The set of true labels is defined as $Y_i^* = \{y_i^*\}_{i=1}^{S}$ where $y_i^* = \{1, \cdots, N^*\}$ with N^* denoting the true number of identities. Predicted clusters are denoted by $C = \{C_j\}_{j=1}^{N}$ with $C_j = \{x_i | y_i = j, \forall x_i \in X\}$, and the true identities are denoted by a set $T = \{T_l\}_{l=1}^{N^*}$ with $T_l = \{x_i | y_i^* = l, \forall x_i \in X\}$. Singleton clusters should be a subset of C, and we denote the number of incorrectly predicted singleton clusters as N_{S}.

2.2 Metrics for Face Clustering

1) Traditional metrics. F_{P} and F_{B} are two widely used metrics to evaluate clustering performance. Basically, the performance of F_{P} and F_{B} is dominated by the large-size clusters [2, 19]. Note that we omit the metric of NMI in prior works [29, 33, 34, 36, 37, 39] due to its tendency to choose the results with large number of clusters [1].

The limitations of the above two metrics are illustrated in Fig. 2, where we show four examples (*i.e.*, (A)-(D) in the figure) of clustering results with three identities. We observe that the images of the identity in blue bounding boxes are failed to group together in (B), leading to more singleton clusters compared to (A), however F_{P} and F_{B} of (B) are higher than those of (A). Similarly, F_{P} and F_{B} of (C) are higher than those of (A), but the images of two identities in (C) are heavily split. Thus higher F_{P} and F_{B} do not imply better clustering results.

2) Proposed new additional metrics. We design three new metrics, named as Ratio of Identity Number ($R_{\#\mathrm{I}}$), Ratio of Singleton Cluster Number ($R_{\#\mathrm{S}}$), and Identity F-score (F_{I}), for complementary evaluation.

$R_{\#\mathrm{I}}$ is the ratio of the number of predicted clusters to the true identity number: $R_{\#I} = N/N^* \times 100\%$. $R_{\#\mathrm{I}}$ is closer to 100%, the better.

$R_{\#\mathrm{S}}$ is the ratio of the number of incorrectly predicted singleton clusters to the true identity number: $R_{\#S} = N_S/N^* \times 100\%$. $R_{\#\mathrm{S}}$ is smaller, the better.

F_{I} is inspired from the metric IDF1 in the evaluation of multiple objects tracking [23]. To evaluate the performance of imbalanced data, F_{I} measures the degree of alignment between predicted clusters and identities. Given an associated pair of sets $APair(j, l) = (C_j, T_l)$, where $j \in \{1, \cdots, N\}$ and $l \in \{1, \cdots, N^*\}$, we can calculate precision and recall on $APair(j, l)$, *i.e.*, $Pre(j, l) = |C_j \cap T_l|/|C_j|$

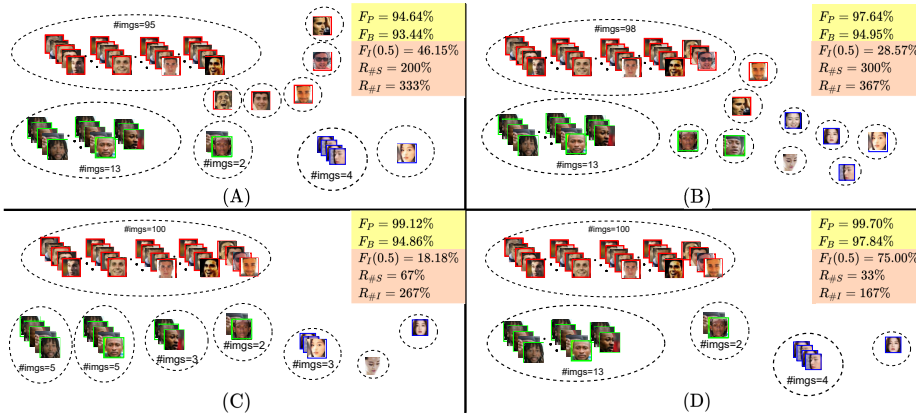


Fig. 2: We illustrate the limitations of F_{P} and F_{B} with four examples of clustering a set of images, denoted by (A)-(D). Face images with bounding boxes via the same color belong to the same identity, *i.e.*, 3 identities in examples. Each predicted cluster is represented by a dashed-line ellipse. The results by traditional metrics and new metrics are respectively shown in yellow and pink blocks. We observe that the images of the identity in blue bounding boxes are failed to group together in (B), leading to more singleton clusters compared to (A), however F_{P} and F_{B} of (B) are higher than those of (A). Similarly, F_{P} and F_{B} of (C) are higher than those of (A), but the images of two identities in (C) are heavily split. By contrast, the new metric F_{I} is sensitive to the small-size clusters with false identities, and correctly discriminates the results in (A)-(C). Similar cases are found in $R_{\#\mathrm{S}}$ and $R_{\#\mathrm{I}}$. For a complete comparison, good performance should be reported not only by F_{P} and F_{B}, but also by F_{I}, $R_{\#\mathrm{S}}$ and $R_{\#\mathrm{I}}$, as shown in (D).

and $Rec(j,l) = |C_j \cap T_l|/|T_l|$. Given a quality threshold $\theta > 0$, we define optimal associated pairs by $OAPair(j,l,\theta)$, where $Pre(j,l) > \theta$ and $Rec(j,l) > \theta$. We may evaluate the identity-level clustering by calculating F-score based on the optimal associated pairs, *i.e.*,

$$F_I(\theta) = 2 \cdot \frac{Pre(\theta) \cdot Rec(\theta)}{Pre(\theta) + Rec(\theta)}, \tag{1}$$

where $Pre(\theta) = |OAPair(j,l,\theta)|/N$ and $Rec(\theta) = |OAPair(j,l,\theta)|/N^*$. The degree of identity-level quality can be controlled by θ according to the purpose of applications in practice, and $\theta \in [0.5, 1)$.

The effectiveness of the new metrics is shown in Fig. 2. The new metric F_{I} is sensitive to the small-size clusters with false identities, and correctly discriminates the results in (A)-(C). Similar cases are found in $R_{\#\mathrm{S}}$ and $R_{\#\mathrm{I}}$. By comparing the four examples, we clearly observe that the combined evaluation metrics better illustrate the performance of face clustering.

2.3 Related Works

In this subsection, we present two streams of studies on face clustering in terms of whether annotated samples are used or not, *i.e.*, unsupervised methods and supervised methods. We also give a brief discussion on the map equation [24].

1) Unsupervised methods. Clustering methods without help from annotated samples have been investigated for facial images [10, 12, 35]. However, the performance of traditional clustering methods, such as K-means [15] and DBSCAN [9], is not satisfactory, especially on large-scale datasets, which has been reported in previous studies [4, 29, 36, 37]. The main reason is that these traditional methods resort to simplistic assumptions on data distributions. For example, K-means needs to pre-set the number of predicted clusters and implicitly assumes that the numbers of samples for different clusters are roughly balanced. Besides, certain techniques, *e.g.*, agglomerative hierarchical clustering (HAC) [31], have been developed to partition facial data with complex distributions. Lin et al. proposed proximity-aware hierarchical clustering in [17] and density-aware clustering in [16] for face clustering. Zhu et al. [40] proposed rank-order distance for clustering images, and demonstrated its ability in filtering outliers. The methods in [16, 17, 40] are hard to scale to large datasets due to complexity. An approximation version of rank-order distance is proposed in [22], which is termed as ARO and is capable of clustering faces at millions scale. However, the performance of ARO decreases rapidly when the data scale increases and is highly influenced by hyper-parameters.

2) Supervised methods. The recent advanced supervised-based methods for face clustering are built with deep learning models. The most studied branch is based on Graph Convolutional Network (GCN). GCN is trained to produce a similarity measure between two facial features with the whole graph structure and thus can be used to correct the noisy affinity graph in face clustering. Clearly, GCN-based methods require annotated samples to learn graph structures.

There had been several techniques proposed for training the GCN model. The first is link prediction, which means that GCN models are trained to predict a link between two nodes in the affinity graph [33]. The second is node confidence estimation [36]. Each node is estimated with confidence of belonging to a cluster by edge connectivity. The third is sub-graph structure learning [29, 37].

Other supervised methods for face clustering include multi-layer perceptron and transformer. Liu et al. [18] select face pairs based on neighborhood density and train a classification model with annotated data. Nguyen et al. [20] formulate the pairwise relationships as a classification of sequences of kNN.

3) Map equation. Map equation is entropy based on information flows of a random walk on networks, and it can be solved by Infomap algorithm [24]. In map equation, information flows refer to sequential transitions among nodes in the graph, which are optimized based on transition probability matrix. The map equation and its solver Infomap are originally proposed for the analysis of complex systems, especially in physics [14, 25, 26] and biology [8]. Traditional applications of Infomap include structure learning on human brain functional networks [21].

3 Method

In this section, we first formulate face clustering as community detection. Then, we propose an effective unsupervised face clustering framework called FaceMap. Equipped with an OD module, it is able to conduct high quality face clustering.

3.1 Face Clustering as Community Detection

Non-overlapping community detection is a process that takes a graph G as input, and produces a set of communities M, which are clusters of nodes in the graph, as output by optimizing a particular objective function, *i.e.*,

$$\arg\min_{M} f(G, M), \tag{2}$$

where $M = \{m_i | m_i \cap m_j = \emptyset \text{ with } i \neq j, m_i \neq \emptyset, 1 \leq i, j \leq |M|\}$ and f is an objective function. The solution M corresponds to C in face clustering.

Given a set of facial features, we calculate a directed affinity graph A based on kNN, where $A \in R^{S \times S}$. For large-scale datasets, we usually have $S \gg k$ and thus A is highly sparse. The affinity graph characterizes the information flows within images, which are inherently quantified by similarities. The transition probability matrix denoted by P is obtained by row normalization of A.

We adopt map equation in [24], which represents the entropy of information flows, as the objective function in Eq. (2). Thus, the map equation for face clustering can be formulated as:

$$\arg\min_{N,Y} L(P, N, Y) = -q_{\curvearrowright} \sum_{i=1}^{N} \frac{q_{i\curvearrowright}}{q_{\curvearrowright}} \log \frac{q_{i\curvearrowright}}{q_{\curvearrowright}} - \sum_{i=1}^{N} p_{i\circlearrowright} \left(\sum_{i=1}^{N} \frac{q_{i\curvearrowright}}{p_{i\circlearrowright}} \log \frac{q_{i\curvearrowright}}{p_{i\circlearrowright}} + \sum_{\alpha \in i} \frac{p_\alpha}{p_{i\circlearrowright}} \log \frac{p_\alpha}{p_{i\circlearrowright}} \right), \tag{3}$$

where Y is predicted labels, and N is the predicted number of identities for facial images. $q_{\curvearrowright} = \sum_{i=1}^{N} q_{i\curvearrowright}$ and $p_{i\circlearrowright} = q_{i\curvearrowright} + \sum_{\alpha \in i} p_\alpha$ with $\alpha \in i$ denoting over all nodes α in cluster i, p_α being the ergodic node visit frequency at node α by a random walk, and $q_{i\curvearrowright} = \sum_{\alpha \in i} \sum_{\beta \notin i} p_\alpha P(\alpha, \beta)$ denoting the per step probability that the random walker exits cluster i. Here, $P(\alpha, \beta)$ is the outgoing probability from node α to β in the transition probability matrix.

In Eq. (3), we have the following intuitive understandings.

1. $q_{\curvearrowright}$ is the probability of a random walker travelling among different clusters. $q_{i\curvearrowright}$ is the probability of a random walker jumping from the cluster i to other clusters. $p_{i\circlearrowright}$ is the probability of a random walker travelling in the cluster i. p_α is the probability of a random walker visiting the node α.
2. The vanilla design of the map equation applies coding theory into network discovery, which compresses the description of information flows on networks. For face clustering, by minimizing the entropy in Eq. (3), we encode paths among images with N clusters via minimal length of descriptions.
3. The above map equation has two terms. The first term describes the entropy of random walk travelling among different clusters, and the second term characterizes the entropy of random walk within a cluster. Besides, the number of clusters is also optimized based on the map equation.

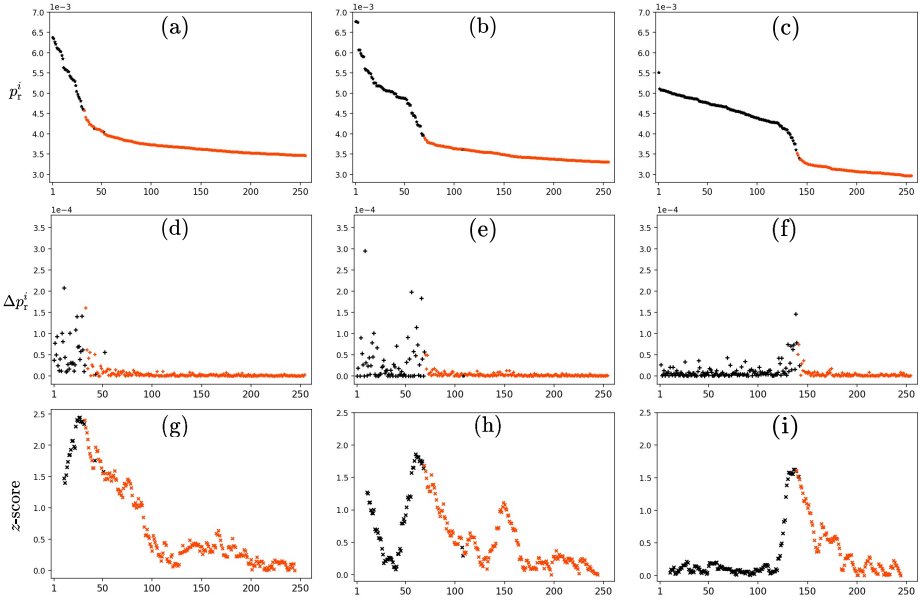


Fig. 3: The observations of data inherent characteristics are shown with three illustrative examples. Each example is a group of three sub-figures in the same column. Given an image x_i, p_r^i is the ranked sequence by descending order of non-zero transition probabilities, *i.e.*, (a)-(c). Δp_r^i is the first-order difference of p_r^i, *i.e.*, (d)-(f). The black and red points represent the images with the same identity of x_i and the images with different identities of x_i respectively. Given a sliding window ω by starting from right to left, we estimate z-score of average Δp_r^i over the samples in the window ω, which is shown in (g)-(i).

3.2 Key Observations

In Fig. 3, we give three examples, each of which is demonstrated by a column, to illustrate our observations. Given an image x_i, its transition probabilities to other images can be obtained from $P(i,:)$. A rank sequence in a descending order of non-zero values of $P(i,:)$ is denoted as p_r^i, which is shown in (a)-(c) in Fig. 3. For x_i, the black and red points represent the images with the same identity of x_i and the images with different identities of x_i respectively. Clearly, with the view from the left to the right, the black points mainly locate at the head region while the red points locate at the tail region. Besides, there exists a mixed region from the head region to the tail region, *e.g.*, from 50 to 100 in (b) of Fig. 3. From the viewpoint of signal processing and clarifying the inherent characteristic of p_r^i, we calculate the first-order difference of p_r^i denoted by Δp_r^i, which is shown in (d)-(f) in Fig. 3. For p_r^i and Δp_r^i, we have the following observations[1].

[1] A detailed study of inherent characteristics of p_r^i is shown in the supplementary materials due to space limitations.

Observation 1. For p_{r}^i, there is no consistent behaviour on the slopes at the head region over all $i \in \{1, \cdots, S\}$. For example, the head region of p_{r}^i for (a) is steeper, compared to that for (b) and (c) in Fig. 3.

Observation 2. For p_{r}^i, there is a stable trend on the slopes at the tail region over all $i \in \{1, \cdots, S\}$, *i.e.*, the slope of p_{r}^i is convergent. In other words, the values of Δp_{r}^i is nearly zero at the tail region.

Observation 3. The switch position between black and red points, which is usually located in the mixed region of p_{r}^i, may be easily detected by Δp_{r}^i with the view from the right to the left.

Based on the above observations, we clearly find that there is no general fixed value of k in nearest neighbor search, or fixed threshold of transition probability, to locate the switch point for any $i \in \{1, \cdots, S\}$. We present the following assumptions related to data characteristics of P.

Assumption 1. For each x_i, we assume that the images with the same identity of x_i and the images with different identities of x_i are roughly separated by a switch point.

Assumption 2. We assume that the values in Δp_{r}^i at the right-hand side of the switch point form a stable distribution, and the switch point is an outlier to the stable distribution.

We give a brief discussion on Assumptions 1 and 2. In general, there roughly exists a switch point for each image with the ranked transition probabilities in practice due to the effectiveness of a pre-trained feature model. Note that the switch point should lie in the mixed region. This implies that a few black points might locate at the tail region, which in fact results from the imperfect feature model. The switch point is an outlier to the stable distribution implying that there is a dramatic change of probabilities between the head and tail regions.

3.3 FaceMap with Outlier Detection

To adaptively detect the switch point in the ranked transition probabilities for each image, the key point is to discriminate the distribution change between the head and tail regions. Based on Observations 1-3 and Assumptions 1-2, we find that the switch point lies in the mixed region, and represents the outlier point of the stable distribution of the tail region.

The framework of FaceMap is demonstrated in Fig. 4, where an outlier detection module is proposed based on z-score statistic of the average on Δp_{r}^i over a window. With the detected outlier, new transition probabilities are updated and thus fed into Eq. (3). For better illustrations, we show the algorithmic procedures in Algorithm 1. In FaceMap, we first get the ranked probabilities by descending order for each image. Then, we calculate the first-order sequence based on the ranked probabilities of each image. The module of outlier detection captures the switch point by maximizing z-score sequence. Finally, FaceMap adjusts transition probabilities and minimizes the entropy in Eq. (3) with Infomap [24].

Based on Assumptions 1-2, we construct a sliding window ω to calculate the average of Δp_{r}^i, *i.e.*, $\hat{\mu}_j = (\sum_j^{j+\omega} \Delta p_{\mathrm{r}}^i(j))/\omega$ with $j = k - \omega, \cdots, 1$. Note

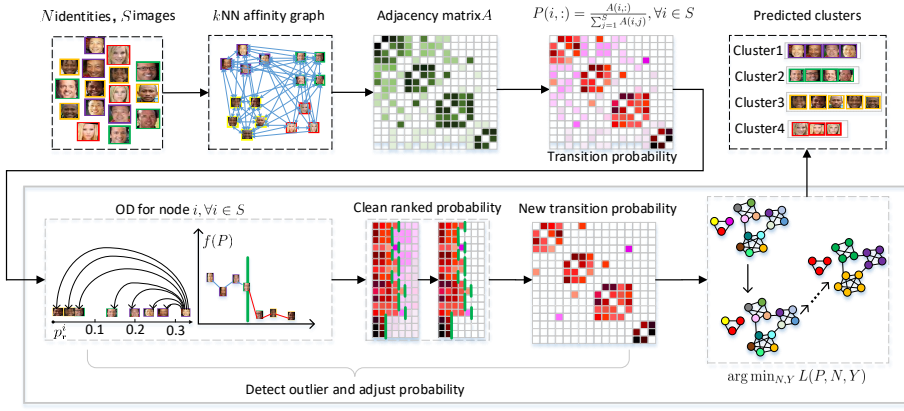


Fig. 4: The framework of FaceMap, where p_r^i is the ranked non-zero transition probabilities of node i and $f(P)$ is a function transformation of P. We design a strategy of outlier detection adaptively adjusting the probabilities. By optimizing the map equation in Eq. (3), the method effectively clusters facial images.

that we start the sliding window from the end of the sequence of Δp_r^i. We introduce the sliding window because the switch point lies in the mixed region. In order to calculate the z-score, we estimate the mean and standard of the stable distribution of the tail region by $\bar{\mu}_j = (\sum_j^k \Delta p_r^i(j))/(k-j-1)$ and $\bar{\sigma}_j = \sqrt{\sum_j^k (\Delta p_r^i(j) - \bar{\mu}_j)^2/(k-j-1)}$ with $j = k-\omega, \cdots, 1$. Note that in Algorithm 1, we set $k = K$.

Without loss of generality, we mark the middle point of a sliding window as the candidate of outlier. We calculate the z-score of the sliding window as $z = |\hat{\mu}_j - \bar{\mu}_j|/\bar{\sigma}_j$, and maximize z-score over the sequence. We give the theoretical insight of maximizing the z-score. Based on Chebyshev-Bienayme's Inequality [28], we are ready to have the tail probability of the stable distribution as $\mathbf{P}\left[\frac{|\hat{\mu}_j - \bar{\mu}_j|}{\bar{\sigma}_j} > v\right] \leq \frac{1}{v^2}$, for all $v > 0$. It means that the larger value of z-score of a point, the less likely it comes from the same distribution and thus the more possibility to be a switch point.

4 Experiments

In this section, we conduct experiments on three large-scale face datasets[2]. The comparisons based on traditional metrics are shown via two categories, unsupervised and supervised settings. We further show the performance on new metrics among all methods. We also conduct ablation studies to demonstrate the consistent performance with respect to hyper-parameters of k and ω in FaceMap, and the superiority of outlier detection in FaceMap for face clustering.

[2] More experiments are shown in the supplementary materials due to space limitations.

Algorithm 1 FaceMap in pseudocode

input: a set of face images with features $\{x_i\}_{i=1}^{S}$, K, window size ω
construct an affinity graph A based on kNN with $k = K$
for $i = 1$ to S **do**
  $P(i,:) = A(i,:)/\sum_{j=1}^{S} A(i,j)$ // calculate the original transition probability
  $p_{\mathrm{r}}^{i} = \text{rank-d}(P(i,:))[:K]$ // rank K probabilities by descending order
  for $j = 1$ to $K-1$ **do**
    $\Delta p_{\mathrm{r}}^{i}(j) = p_{\mathrm{r}}^{i}(j) - p_{\mathrm{r}}^{i}(j+1)$ // first-order values of the ranked probabilities
  end for
  initialize $z = \text{zeros}(K)$
  for $j = K-\omega-1$ to 1 **do**
    $\hat{\mu}_j = \sum_{j}^{j+\omega} \Delta p_{\mathrm{r}}^{i}(j)/\omega$ // mean of $\Delta p_{\mathrm{r}}^{i}$ in the sliding window
    $\bar{\mu}_j = \sum_{j}^{K-1} \Delta p_{\mathrm{r}}^{i}(j)/(K-j-1)$ // mean of $\Delta p_{\mathrm{r}}^{i}$ in the tail region
    $\bar{\sigma}_j = \sqrt{\frac{\sum_{j}^{K-1}(\Delta p_{\mathrm{r}}^{i}(j)-\bar{\mu}_j)^2}{K-j-1}}$ // standard deviation of $\Delta p_{\mathrm{r}}^{i}$ in the tail region
    $q = j + \lceil \omega/2 \rceil$ // mark the middle point as a candidate of outlier
    $z(q) = \frac{|\hat{\mu}_j - \bar{\mu}_j|}{\bar{\sigma}_j}$
  end for
  $q^* = \arg\max_q z(q)$ // maximize z-score to detect the outlier
  for $j = 1$ to S **do**
    $P(i,j) = P(i,j) >= p_{\mathrm{r}}^{i}(q^*) ? P(i,j) : 0$ // adjust transition probabilities
  end for
end for
$N_{\mathrm{P}}, Y_{\mathrm{P}} = \arg\min_{N,Y} L(P,N,Y)$
output: Y_{P}

Table 1: Statistics of datasets. AVG stands for average image numbers per identity. STD stands for standard deviation of image numbers per identity.

Datasets	# of images	# of identities	AVG	STD
MS1M [11]	5.8M	85.0K	68.5	40.6
VGGFace2 [5]	3.3M	9.1K	362.6	101.3
CASIA [38]	0.5M	10.5K	46.4	59.3

4.1 Datasets and Metrics

The three publicly available large-scale datasets of face images are MS-Celeb-1M [11], VGGFace2 [5] and CASIA [38]. Following [7], we use the refined version of MS-Celeb-1M, denoted as MS1M. The statistics of three datasets are shown in Table 1. From the table, we find that there is a large discrepancy of distribution of image numbers per identity among three datasets. It is challenging to solve face clustering problem across different datasets with good performance.

For evaluation metrics, we first adopt the traditional F_{P} and F_{B} for all methods. Then, for evaluating identity-level quality of clustering, we calculate the values based on new metrics proposed in this paper, *i.e.*, $R_{\#\mathrm{I}}$, $R_{\#\mathrm{S}}$ and F_{I}, which have been discussed in details in Section 2.2.

Table 2: Performance comparison with unsupervised methods. Our FaceMap yields gains of 10.38% on F_{P} and 5.75% on F_{B} on average, compared to the best results of unsupervised methods. We adopt $k = 256$ and $\omega = 20$ for FaceMap.

Datasets	**Part1(0.58M)**		**Part3(1.74M)**		**Part5(2.89M)**		**Part7(4.05M)**		**Part9(5.21M)**		**CASIA**		**VGGFace2**	
Methods	F_{P}	F_{B}	F_{P}	F_{B}	F_{P}	F_{B}	F_{P}	F_{B}	F_{P}	F_{B}	F_{P}	F_{B}	F_{P}	F_{B}
K-means [15]	79.21	81.23	73.04	75.20	69.83	72.34	67.90	70.57	66.47	69.42	36.72	78.53	20.47	81.46
HAC [31]	70.63	70.46	54.40	69.53	11.08	68.62	1.40	67.69	0.37	66.96	61.87	53.65	NA	NA
ARO [22]	85.58	87.75	82.26	84.39	81.08	82.78	79.76	81.60	79.76	80.67	75.88	87.08	78.71	84.35
DBSCAN [9]	67.93	67.17	63.41	66.53	52.50	66.26	45.24	44.87	44.94	44.74	57.25	49.43	66.88	65.49
FaceMap	94.24	92.55	91.31	89.67	89.32	88.20	87.74	87.11	86.37	86.29	92.55	91.24	94.15	93.78
	+8.66	**+4.80**	**+9.05**	**+5.28**	**+8.24**	**+5.42**	**+7.98**	**+5.51**	**+6.61**	**+5.62**	**+16.67**	**+4.16**	**+15.44**	**+9.43**

4.2 Implementation Details

Following the setting in [29, 36, 37], MS1M is divided into 10 parts with nearly equal number of identities. A face clustering test benchmark in multiple scales is built by combining different numbers of subsets. We use number of subsets and face numbers at multiple scales to denote these test sets, *e.g.*, Part3(2.89M) denotes the union of 3 subsets of MS1M with 2.89M face instances. Similarly, We randomly select 10 percent of identities of each dataset as training set and the others as test set for VGGFace2 and CASIA. In the experiments, we take the training set to train the GCN-based models in supervised methods and evaluate on the test set for all the methods to make a fair comparison. We use extracted feature provided in [29] as the face representation of MS1M. We use the pre-trained Arcface model, which is trained on Glint-360K [7, 3], to extract face representation features of VGGFace2 and CASIA. To build the affinity graph, we use Faiss [13] to search nearest neighbors for all methods. In experiments, we set k as 256 for FaceMap and 80 for GCN-based methods. We set $\omega = 20$.

4.3 Results

1) Performance comparison with unsupervised methods. We report the clustering results in Table 2. FaceMap outperforms all the unsupervised methods by a large margin on all the test datasets. We note that the performance of most of the unsupervised methods is sensitive to the hyper-parameters, *e.g.*, the pre-set cluster number in K-means. In addition, some methods are not scalable for large-scale datasets, *e.g.*, HAC failed to cluster VGGFace2 in 360 hours. For ARO, we report superior results compared to the performance in [29] by carefully tuning the thresholds. The performance of DBSCAN is not consistent across the datasets as it assumes a balanced density in each cluster. Note that FaceMap has superior performance across different test datasets.

2) Performance comparison with supervised methods. FaceMap achieves a better result on all the test datasets, as shown in Table 3. We directly report the results of MS1M in the related reference. The results of CASIA and VGGFace2 are produced using the source codes provided by the authors. We observe that the performance of supervised methods on CASIA and VGGFace2 is not as good as the ones in MS1M. This gap is caused by the difference of labeled data and the generalization of the hyper-parameters.

Table 3: Performance comparison with supervised methods. Our FaceMap yields gains of 4.80% on F_{P} and 5.04% on F_{B} on average, compared to the best results of supervised methods. We adopt $k = 256$ and $\omega = 20$ for FaceMap.

Datasets	**Part1(0.58M)**		**Part3(1.74M)**		**Part5(2.89M)**		**Part7(4.05M)**		**Part9(5.21M)**		**CASIA**		**VGGFace2**	
Method	F_{P}	F_{B}	F_{P}	F_{B}	F_{P}	F_{B}	F_{P}	F_{B}	F_{P}	F_{B}	F_{P}	F_{B}	F_{P}	F_{B}
L-GCN [33] CVPR 2019	78.68	84.37	75.83	81.61	74.29	80.11	73.70	79.33	72.99	78.60	68.52	79.22	60.74	66.42
GCN-D [37] CVPR 2019	85.66	85.52	83.76	83.99	81.62	82.00	80.33	80.72	79.21	79.71	78.71	83.14	65.80	69.27
GCN-V+E [36] CVPR 2020	87.93	86.09	84.04	82.84	82.10	81.24	80.45	80.09	79.30	79.25	83.40	81.35	50.35	52.47
Clusformer [20] CVPR 2021	88.20	87.17	84.60	84.05	82.79	82.30	81.03	80.51	79.91	79.95	NA	NA	NA	NA
CPC [18] ICCV 2021	90.67	89.54	86.91	86.25	85.06	84.55	83.51	83.49	82.41	82.40	NA	NA	NA	NA
STAR-FC [29] CVPR 2021	91.97	90.21	88.28	86.26	86.17	84.13	84.70	82.63	83.46	81.47	75.34	71.81	84.12	83.49
FaceMap	94.24	92.55	91.31	89.67	89.32	88.20	87.74	87.11	86.37	86.29	92.55	91.24	94.15	93.78
	+2.27	**+2.34**	**+3.03**	**+3.41**	**+3.15**	**+3.65**	**+3.04**	**+3.62**	**+2.91**	**+3.89**	**+9.15**	**+8.10**	**+10.03**	**+10.29**

Table 4: Identity-level comparison via new metrics. FaceMap results high quality clusters at identity level compared to other methods. All the results are reported in %. ↓ indicates the smaller results are better.

Datasets	**CASIA**				**VGGFace2**				**Part1(0.58M)**			
Methods	$F_{\mathrm{I}}(0.5)$	$F_{\mathrm{I}}(0.9)$	$R_{\#\mathrm{S}}(\downarrow)$	$R_{\#\mathrm{I}}$	$F_{\mathrm{I}}(0.5)$	$F_{\mathrm{I}}(0.9)$	$R_{\#\mathrm{S}}(\downarrow)$	$R_{\#\mathrm{I}}$	$F_{\mathrm{I}}(0.5)$	$F_{\mathrm{I}}(0.9)$	$R_{\#\mathrm{S}}(\downarrow)$	$R_{\#\mathrm{I}}$
HAC [31]	4.45	0.28	1764.75	1990.40	NA	NA	NA	NA	8.83	2.13	896.29	1432.17
ARO [22]	32.36	22.11	348.01	505.03	4.88	1.29	321.66	3953.92	26.18	14.51	372.17	596.36
DBSCAN [9]	3.60	0.19	1917.78	2151.39	1.68	0.02	9013.06	10019.62	7.64	1.63	981.63	1571.83
STAR-FC [29]	7.89	1.93	1281.97	1387.47	2.20	1.55	7313.88	7413.10	23.72	17.62	466.91	574.72
FaceMap	**92.45**	**66.30**	**3.15**	**110.88**	**63.99**	**51.88**	**26.55**	**209.85**	**89.05**	**67.53**	**1.94**	**110.77**

3) Performance comparison of identity-level quality. We compare our methods from the perspective of identity-level quality in Table 4. Here, we only report the result of the state-of-the-art STAR-FC as the representative supervised methods. We can observe that $R_{\#\mathrm{I}}$ is closer to 100% and the $R_{\#\mathrm{S}}$ is much smaller compared to other methods. This result demonstrates that FaceMap has a great advantage to estimate the true number of identity. In addition, our method yields a large boost on F_{I} with θ=0.5 and 0.9, which demonstrates that a large number of identities with high quality clusters are generated by FaceMap.

4.4 Ablation Studies

1) Study on the sensitivity of K and ω. FaceMap has two hyper-parameters, *i.e.*, K and ω. We note that K affects the construction of the affinity graph, and ω can influence the detection of the change position in outlier detection. Table 5 shows the performance of FaceMap when we vary K and ω on the MS1M Part1(0.58M). We observe that FaceMap exhibits stable performance on all the metrics except $R_{\#\mathrm{S}}$, and larger K and smaller ω result in smaller $R_{\#\mathrm{S}}$. This result shows that FaceMap is robust.

2) Study on the effectiveness of outlier detection. We denote FaceMap without the OD module by FaceMap-OD. We introduce hand-crafted rules to reduce the noisy edges in the affinity graph by fixing a similarity threshold δ based on kNN, represented by FaceMap-OD(k,δ). In addition, we delete the OD module and adopt a learned affinity graph from GCN-based method (STAR-FC) as input for the map equation in Eq. (3), which is represented by FaceMap-GCN. We can observe that FaceMap consistently outperforms all other methods with

Table 5: Performance of FaceMap on MS1M Part1(0.58M) with different K and ω. All the results are reported in %. ↓ indicates the smaller results are better.

K	128					256					512					Mean	STD
ω	10	15	20	25	30	10	15	20	25	30	10	15	20	25	30		
F_{P}	94.02	94.21	94.36	94.39	94.32	94.21	94.23	94.24	94.16	94.01	94.02	93.99	93.85	93.84	93.57	94.10	0.22
F_{B}	92.31	92.50	92.61	92.63	92.58	92.54	92.56	92.55	92.51	92.37	92.42	92.41	92.31	92.28	92.06	92.44	0.15
$F_{\mathrm{I}}(0.9)$	64.33	65.41	66.49	67.03	67.40	66.41	67.23	67.53	67.58	67.97	66.50	67.25	67.48	67.60	67.74	66.93	0.95
$R_{\#\mathrm{S}}(\downarrow)$	1.99	2.25	2.89	3.10	3.01	1.35	1.54	1.94	2.16	1.98	1.00	1.19	1.40	1.66	1.33	1.92	0.64
$R_{\#\mathrm{I}}$	120.32	118.06	115.22	113.90	111.79	115.19	112.99	110.77	109.46	106.95	113.18	110.71	107.85	106.89	104.41	111.85	4.23

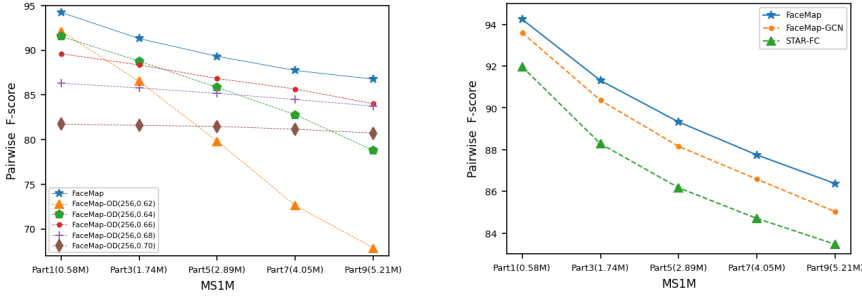


Fig. 5: Ablation study on the effectiveness of outlier detection.

different thresholds, which demonstrates the effectiveness and robustness of the outlier detection module across multiple scales of datasets. Besides, we observe that FaceMap-GCN performs better than STAR-FC, and FaceMap exhibits a consistent better performance compared to FaceMap-GCN. This result further validates the superiority of the proposed outlier detection module.

5 Conclusions

In this paper, we propose an unsupervised method, named as FaceMap, for large-scale face clustering. We formulate face clustering as community detection with the map equation. By minimizing the map equation on the entropy of the structure in affinity graph from facial images, we obtain the predicted clusters of facial images. In order to alleviate the noisy transition probability, we develop a strategy of outlier detection to adaptively adjust transition probabilities. We also illustrate the limitations of the traditional metrics for face clustering and design three metrics for comprehensive evaluations. The proposed FaceMap achieves new state-of-the-arts in light of traditional metrics on three large-scale datasets, where our method significantly outperforms the prior methods. We demonstrate the superiority of FaceMap in identity-level quality via new metrics. From a viewpoint of practical applications with super large-scale datasets, we may take the advantage of the distributed characteristic in calculation of outlier detection module for FaceMap, and reduce computation complexity with a high-performance distributed system. Further works on this direction lie in a broader applications of the map equation to more computer vision tasks.

References

1. Amelio, A., Pizzuti, C.: Is normalized mutual information a fair measure for comparing community detection methods? In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015. pp. 1584–1585 (2015)
2. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval **12**(4), 461–486 (2009)
3. An, X., Zhu, X., Gao, Y., Xiao, Y., Zhao, Y., Feng, Z., Wu, L., Qin, B., Zhang, M., Zhang, D., et al.: Partial fc: Training 10 million identities on a single machine. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1445–1449 (2021)
4. Bijl, E.: A comparison of clustering algorithms for face clustering. Ph.D. thesis (2018)
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition. pp. 67–74. IEEE (2018)
6. Cao, X., Wei, X., Han, Y., Lin, D.: Robust face clustering via tensor decomposition. IEEE Transactions on Cybernetics **45**(11), 2546–2557 (2014)
7. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
8. Edler, D., Guedes, T., Zizka, A., Rosvall, M., Antonelli, A.: Infomap bioregions: Interactive mapping of biogeographical regions from species distributions. Systematic biology **66**(2), 197–204 (2017)
9. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd. vol. 96, pp. 226–231 (1996)
10. Gan, G., Ma, C., Wu, J.: Data clustering: theory, algorithms, and applications. SIAM (2020)
11. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 87–102. Springer (2016)
12. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM computing surveys (CSUR) **31**(3), 264–323 (1999)
13. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019)
14. Lambiotte, R., Rosvall, M., Scholtes, I.: From networks to optimal higher-order models of complex systems. Nature physics **15**(4), 313–320 (2019)
15. Likas, A., Vlassis, N., Verbeek, J.J.: The global k-means clustering algorithm. Pattern Recognition **36**(2), 451–461 (2003)
16. Lin, W.A., Chen, J.C., Castillo, C.D., Chellappa, R.: Deep density clustering of unconstrained faces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8128–8137 (2018)
17. Lin, W.A., Chen, J.C., Chellappa, R.: A proximity-aware hierarchical clustering of faces. In: IEEE International Conference on Automatic Face & Gesture Recognition. pp. 294–301. IEEE (2017)
18. Liu, J., Qiu, D., Yan, P., Wei, X.: Learn to cluster faces via pairwise classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3845–3853 (2021)

19. Moreno, J.G., Dias, G.: Adapted b-cubed metrics to unbalanced datasets. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 911–914 (2015)
20. Nguyen, X.B., Bui, D.T., Duong, C.N., Bui, T.D., Luu, K.: Clusformer: A transformer based clustering approach to unsupervised large-scale face and visual landmark recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10847–10856 (2021)
21. Nicolini, C., Bifone, A.: Modular structure of brain functional networks: breaking the resolution limit by surprise. Scientific reports **6**(1), 1–13 (2016)
22. Otto, C., Wang, D., Jain, A.K.: Clustering millions of faces by identity. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 289–303 (2017)
23. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 17–35. Springer (2016)
24. Rosvall, M., Axelsson, D., Bergstrom, C.T.: The map equation. The European Physical Journal Special Topics **178**(1), 13–23 (2009)
25. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**(4), 1118–1123 (2008)
26. Rosvall, M., Bergstrom, C.T.: Mapping change in large networks. PloS one **5**(1), e8694 (2010)
27. Scherhag, U., Rathgeb, C., Merkle, J., Breithaupt, R., Busch, C.: Face recognition systems under morphing attacks: A survey. IEEE Access **7**, 23012–23026 (2019)
28. Seneta, E.: A tricentenary history of the law of large numbers. Bernoulli **19**(4), 1088–1121 (2013)
29. Shen, S., Li, W., Zhu, Z., Huang, G., Du, D., Lu, J., Zhou, J.: Structure-aware face clustering on a large-scale graph with 107 nodes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9085–9094 (2021)
30. Shi, Y., Otto, C., Jain, A.K.: Face clustering: representation and pairwise constraints. IEEE Transactions on Information Forensics and Security **13**(7), 1626–1640 (2018)
31. Sibson, R.: Slink: an optimally efficient algorithm for the single-link cluster method. The computer journal **16**(1), 30–34 (1973)
32. Tian, Y., Liu, W., Xiao, R., Wen, F., Tang, X.: A face annotation framework with partial clustering and interactive labeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007)
33. Wang, Z., Zheng, L., Li, Y., Wang, S.: Linkage based face clustering via graph convolution network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1117–1125 (2019)
34. Xing, Y., He, T., Xiao, T., Wang, Y., Xiong, Y., Xia, W., Wipf, D., Zhang, Z., Soatto, S.: Learning hierarchical graph neural networks for image clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3467–3477 (2021)
35. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Transactions on neural networks **16**(3), 645–678 (2005)
36. Yang, L., Chen, D., Zhan, X., Zhao, R., Loy, C.C., Lin, D.: Learning to cluster faces via confidence and connectivity estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13369–13378 (2020)
37. Yang, L., Zhan, X., Chen, D., Yan, J., Loy, C.C., Lin, D.: Learning to cluster faces on an affinity graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2298–2306 (2019)

38. Yi, D., Lei, Z., Liao, S., Li, S.: Casiawebface: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
39. Zhang, Y., Deng, W., Wang, M., Hu, J., Li, X., Zhao, D., Wen, D.: Global-local gcn: Large-scale label noise cleansing for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7731–7740 (2020)
40. Zhu, C., Wen, F., Sun, J.: A rank-order distance based clustering algorithm for face tagging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 481–488. IEEE (2011)