# Integrated System for Face Detection, Clustering and Recognition

Anqi Han
Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology
Changsha, Hunan, P.R. China
15580898175,410073
hananqi2016@163.com

Shuohao Li
Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology
Changsha, Hunan, P.R. China
15580868413,410073
shuohaoli@gmail.com

Jun Zhang
Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology
Changsha, Hunan, P.R. China
13308491299,410073
Zhangjun1975@nudt.edu.cn

Haodong Yang
Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology
Changsha, Hunan, P.R. China
18684895316,410073
yanghaodong12@nudt.edu.cn

Rui Wang
National defense acquisition and system engineering management, National University of Defense Technology
Changsha, Hunan, P.R. China
18673119805,410073
406919994@qq.com

## ABSTRACT

Recent years, many approaches have achieved remarkable performances in face detection, clustering and recognition. However, real life can barely see systems designed for their combination, and thousands of videos and images need to be handled in a timely manner. In this paper, an end-to-end face detection, clustering and recognition system has been proposed. The input of our system can be videos or images. For videos, we firstly utilize ffmpeg to transfer these collections into separated frames and for images this step can be skipped. Secondly, we employ joint detection and alignment to detect faces in frames. Then the clustering step is conducted to find face relations based on Interpretive Structure Model. Finally, we establish our own dataset and train our own classifier to realize face recognition based on FaceNet. Most important of all, we propose our original clustering method which avoids duplicate feature computation and repetitive face recognition. And it enhances the efficiency of handling data and shortens the runtime of our experiments greatly.

This system is evaluated using our designed database of 43 persons' faces with varying scales and poses obtained on different complex backgrounds. The performance of the system is quite good and it achieves average accuracy of 87% to 92%.

## CCS Concepts

• **Computing methodologies → Neural networks**


*ICMLT 2018*, May 19--21, 2018, JINAN, China

ACM ISBN 978-1-4503-6432-4/18/05…$15.00
DOI: https://doi.org/10.1145/3231884.3231899

## Keywords

Face detection; face clustering; face recognition; integrated system.

## 1. INTRODUCTION

In the last few years, computer vision and deep learning in face field have grown rapidly. Both the academic world and the industry are pushing forward to speed up the developments and the researches in this area. However, there are no existing systems or relevant devices in real life. And thousands of images and videos need to be manipulated and analyzed in cases that needs to understand how many people exist, who these people are and who they are related to in social media for intelligence analyst. These represent a major issue that such cases must be proceeded in a timely system. And it's worth mentioning that if only merging face detection and recognition stage, the gathered information will be too trivial to deal with which are quite similar and redundant. Therefore, face clustering is the fundamental and critical step to arrange the duplicate data and reduce the complexity of computation.

In this paper, we present FDCR system for face detection (if this is a person), clustering (classify persons among these faces) and recognition (who is this person). Obviously, our work provides the most fundamental and essential component for the intelligence analyst, and we mainly focus on the analysis of news and images collected in important conferences. Our system is composed of four modules--video framing, face detection, face clustering and face recognition.

For video framing, one of the most popular tools ffmpeg is being used to extract frames in our collected news based on our requirements. For face detection, we present a face detector that is capable of both accuracy and speed. Our detecting work is motivated by two major contributions. One is the accurate face alignment which is helpful to distinguish faces and non-faces. The other is the recent advances in cascade face alignment, as the cascade structure has been proven effective in detection field. So we propose to combine the two to benefit each other. And we

make some improvements of this method to detect multiple faces in images and add an algorithm to decide if it's a group photo.

Then the objective of our clustering step is to learn a mapping from face images to a compact Euclidean space based on FaceNet[1]. In Euclidean spaces, distances directly correspond to a measure of face similarity, that is, faces of the same person have small distances and faces of distinct people have large distances. When this stage receives cropped faces from last stage, we exploit the Euclidean embeddings to compute an adjacency matrix where each element represents a distance between two faces. And then we acquire reachable matrix based on interpretive structure model to distinguish faces' tag and divide them into different groups. This makes subsequent step more cost-effective instead of repetitive consumption for tackling the same face features.

Once this embedding has been produced in FaceNet, face recognition becomes straightforward. Our method directly trains the output to be a compact 128-D embedding. We adhere to triplet-based loss function which aims to separate the positive pair from the negative by a distance margin. And theses triplets consist of two matching face thumbnails and a non-matching face thumbnail which are tight crops of the face area. The basic recognition principle is unchanged, however, public dataset cannot satisfy our research direction. So we establish our own face dataset and train our classifier based on FaceNet to tackle our recognition task.

The contributions of our work can be summarized as follows:

(1)We present an original end-to-end FDCR system which could handle videos or images. It could not only analyze if the target exists in social media, but also it can test if the image is a group photo.

(2)We propose our own clustering method distinguished from previous approaches. Our algorithm is based on Interpretive Structure Model adhering to the solid math foundation. It can remove redundant data and leave out the unnecessary facial feature computation greatly.

(3)We firstly establish our face list composed of 43 persons, and collect each person's images through crawler. There are about 400 or more images for each person. And we clean this dataset manually in case that there are some other people mixed in each group. This dataset contains leaders in China and other foreign countries, and we use it to train our own classifier to recognize these leaders in big events.

An overview of the remaining paper is as follows: in section 2 we review the literature in face research. Section 3 gives us a detailed description of each part of our FDCR system. And section 4 presents some quantitative results of our system and corresponding performances. Finally we gives a conclusion and further potential development of this work in section 5.

## 2. RELATED WORK

Recently, there is a wide range of methods using deep learning algorithm in the field of face research.

As for face detection, previous approaches often focused on multi-view[2,3], their face detectors were trained separately under various viewpoints or head poses. However, the viewpoint estimation problem was difficult and quantization also brought about inaccuracy, and they would lead to more difficult training and slower detectors. So new approaches that do not use boosted cascade spring up like mushrooms. Shen et al.[4] exploited advanced image retrieval techniques to avoid the expensive sliding window search. Zhu et al.[5] proposed a combination of deformable part models to capture large face variations on conditions of different viewpoints and expressions which could estimate facial points and head poses simutaneously. However, those accurate detectors are all quite slow as a result of their high complexity. Our detection approach takes advantage of the recent advances in cascade face alignment[6,7,8,9]. In such works, we present a face detector qualified for both accuracy and speed which adheres to the principle of "boosted cascade structure + simple features". We use differences of pixels as feature which benefits the efficiency. Features learnt this way are called shape indexed features which could present more invariance to the geometric variations in the face shapes and they are crucial for high alignment accuracy and speed. In this case, our detector achieves the best accuracy on the challenging datasets [5,10] significantly and outperforms all existing academia solutions including [4,5].

The following part of our system is our original face clustering module which helps classification and arrangements of massive data in disordered state. Ho et al.[11] proposed spectral clustering which computed the affinity matrix based on Lambertian object and compared the local gradients of the images. Wang et al.[12] primarily developed a kNN graph construction method to construct the nearest neighbor lists. Zhu et al.[13] developed a dissimilarity measure to perform hierarchical clustering based on the rank-order distance function. The above ways used complex algorithm based on solid math foundation but were only applicable for small datasets. Clustering results depend not only on the choice of clustering algorithm, but also on the quality of the underlying face representation and metric. In this paper, a clustering method has been proposed using the core thought of Interpretative Structural Modeling(ISM) method. After we get the Euclidean embedding per image learning from a deep convolutional neural network(CNN), we attain more precise face representation and then we compute an adjacency matrix each element of which represents distance between two faces. And finally we obtain reachable matrix based on ISM to find identities between these faces and divide them into different groups.

As for recognition, there is a vast corpus of works. Zhenyao et al.[14] employed a deep neural network(DNN) to warp faces into a frontal view and then learned CNN that classified each face to a known identity. Taigman et al.[15] proposed a multi-stage approach that aligned faces to a general 3D shape model. Sun et al. [16,17] used an ensemble of 25 of these networks, each operating on a different face patch. Their recognition loss was similar to the triplet loss we employed in [18]. Then were both optimized by minimizing the $L_2$-distance between faces of the same identity and enforcing a margin between the distance of faces of different identities. The main difference of them was that only pairs of images were compared, whereas the triplet loss encouraged a relative distance constraint. Our approach is a purely data-driven method based on Facenet which learns face representation directly from the pixels of the faces. In this paper, we mainly exploit two deep network architectures. The first architecture is based on the Zeiler&Fergus[19] model which is made up of multiple interleaved layers of convolutions, non-linear activations, local response normalizations, and max pooling layers. And the second network architecture is based on the Inception model of Szegedy et al. which was used as the winning approach for ImageNet 2014 [3]. These networks use mixed layers that run several different convolutional and pooling layers parallelly and concatenate their responses.

## 3. MODEL

In this section, we will introduce the architecture and implementation of the system. Different from other systems, our system will automatically manipulate video or image collections with high efficiency and low complexity. It contains four main steps: video framing, face detection, face clustering and face recognition. Figure.1 shows the architecture of our system, and A, B, C, D represents video framing, face detection, face clustering and face recognition respectively. In this system, we use FaceNet to extract features, and we attain euclidean distance based on theses features to represent similarity. Finally, we implement our clustering and training process.

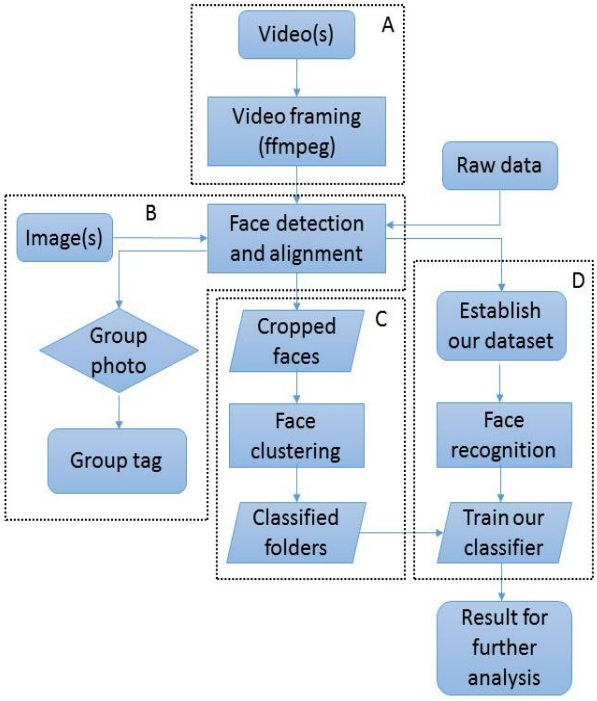


**Figure 1. The architecture of end-to-end FDCR system. It contains four major components: A video framing; B face detection; C face clustering; D face recognition.**

### 3.1 Video Framing

This phase is only for video input, image input is only related to section 3.2, 3.3 and 3.4. To handle video, we mainly use ffmpeg tool. Figure.2 is the main principle of ffmpeg.

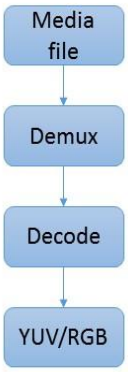


**Figure 2. The principle of ffmpeg.**

We know that multimedia files include both audio and video. Although they are compressed separately, they are bundled together for transmission's convenience. So the first step in our decoding is to separate these bundled audio and video streams, which is the demultiplex called Demux. Secondly, a multimedia file must be compressed in one or more formats known as video and audio coding to reduce the amount of data for storage devices. When compressed frames are extracted from video streams and audio streams, this process is called decoding, and then we get our YUV/RGB data.

And ffmpeg has the following usages:

(1) Extracting a single frame, that is we extract just a single frame from the video into an image file.

(2) Periodic thumbnails, that is we create one thumbnail image every given second.

(3) I-frame thumbnails, that is we create one thumbnail image every I-frame (I is the given number).

After all this consideration, we choose the third method to extract the key frames from a video with a few minutes' duration, and we get their position information to have a better understanding of videos.

### 3.2 Face Detection

In this part, we employ a face detector based on [20] which follows the "boosted cascade structure + simple features" principles.

Our detection method contains two major steps:

(1) Alignment helps detection: a post classifier.

(2) A unified framework for cascade face detection and alignment.

We learn a weak classifier in step (1) shown as Eq.(1) and learn a tree regressors in step (2) shown as Eq.(2). And we can see that both Eq.(1) and Eq.(2) share a similar additive form. This approach learns both classification and regression in the same cascade decision tree benefiting from both accuracy and speed.

$$f = \sum_{t=1}^{T} \sum_{k=1}^{K} C_k^t (x, S^{t-1}) \quad (1)$$

$$R^t(x, S^{t-1}) = \sum_{k=1}^{K} R_k^T (x, S^{t-1}) \quad (2)$$

Each regression tree $R_k^t$ in Eq.(2) is reinforced to a mixed tree $CR_k^t$ that outputs a classification score as well as its shape increment. And the classification and regression parts are evaluated simultaneously during these tests. Details can be seen in [20]. Subsequently, we make some improvements of this work. Noticeably, FaceNet can only detect and crop one face in image, but our system is aimed to detect all faces in images and analyze the potential target. So we change a certain condition in it to crop multiple faces and resize them to $160 \times 160$ pixels for standardization. In addition, we eliminate some side faces and low-resolution faces by setting a threshold which human eyes couldn't tell. Figure.3 illustrates detection examples in video frames. Figure.4 demonstrates detection examples in images. This shows our detection method achieves a good performance besides some low-resolution images. In addition, we add an algorithm for detecting whether an image is a group photo. We can find the relation among these images and dig more hidden clues for searching for the target. To achieve this part, we acquire four vertex coordinates of these cropped faces and compare these invariance information among these. If it meets specific deviation

conditions, we define it's a group photo and it achieves a satisfactory rate of approximately 87%.

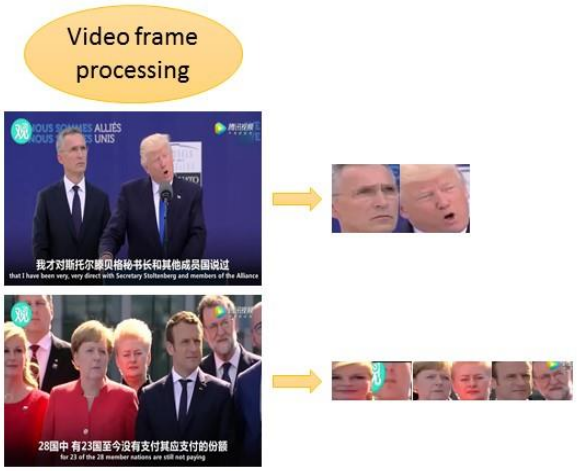


**Figure 3. The cropped faces are from video frames. Each frame in video has been compressed comparable to images, consequently, some faces in frames are maybe in low resolution. And we ignore some side faces which our detection algorithm couldn't find enough features and human eyes cannot tell who they are. All the cropped faces will then be resized to 160×160 pixels.**

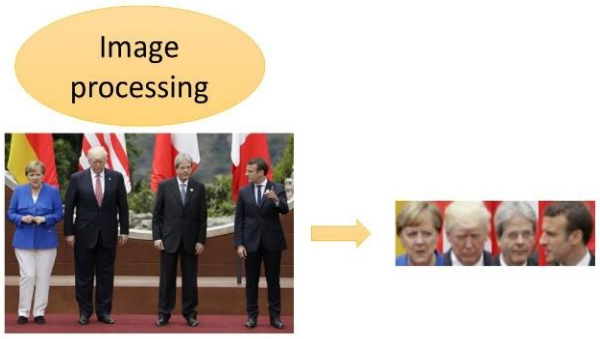


**Figure 4. The cropped faces are from images. Because images have higher resolution than frames in video, so the accuracy of image detection is generally better than video detection. And this image is detected as a group photo.**

## 3.3 Face Clustering

Face clustering is the original part of our system, and it is an important but easily ignored requirement for handling faces. We all know that the faces in video frames are duplicate to a large extent, because the same person may show up in most time of the clip. But we couldn't identify the same face repetitively because of its limited resources and allowed time for users. Existing systems have ignored the importance of arrangements for data, and most of them only take face detection and face recognition into account without considering computation complexity and running time. This comes to our clustering method to solve this problem. Previous approaches are related to k-means which are all given dividing number. However, how many people will appear in the video or image is not certain and it's not informed beforehand. So we propose a face similarity algorithm based on graph illustrated in Figure.5 to measure the similarity of faces.

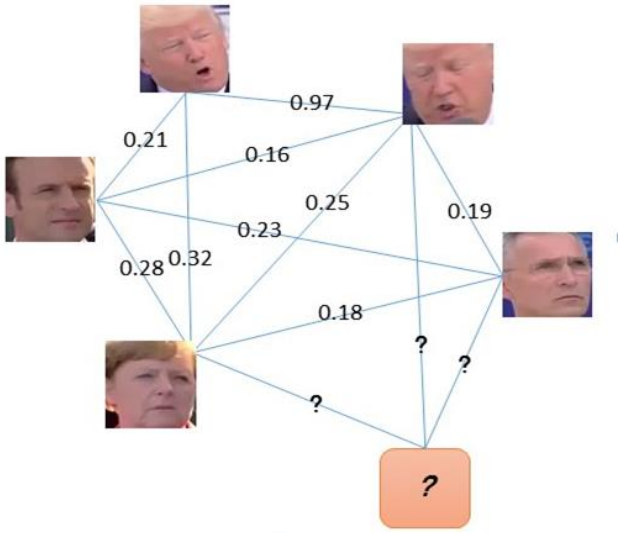


**Figure 5. Face undirected relation graph. The vertex presents the face, and the undirected edge presents the face relation, and the weight presents how close these relations are.**

We assume that there are n faces in group G, then G can be presented as $G = (e_1, e_2, \dots, e_n)$. And we can define $e_i = (i = 1,2, \dots, n)$ as a node of the graph which represents a face, and $e_{ij}(i, j = 1,2, \dots, n)$ as an undirected edge which presents a relation between two faces.

Then we compute adjacency matrix denoted as dist shown in Eq.(3). The undirected face graph is shown as Figure.5. The cropped face plays a role of the vertex, whereas the undirected edge presents a similarity between faces. We can see that President Trump is 97% similar to his own side face. And the smaller the value, the more dissimilar between these faces.

$$\text{dist} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{21} & \cdots & e_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nn} \end{bmatrix} \quad (3)$$

where $e_{ij}$ presents distance between $e_i$ and $e_j$ and it can be computed by Euclidean embeddings described in section 3.4.2.

We firstly employ the embedding per image from FaceNet to obtain and then we set a threshold θ through comparison to decide if $e_i$ is similar to $e_j$. If $e_{ij}$ is greater than or equal to θ, we set $e_{ij} = 1$ which means $e_i$ is relational to $e_j$. On the other hand, if $e_{ij}$ is less than θ, we set $e_{ij} = 0$ which means $e_i$ and $e_j$ have no relation, that is,

$$e_{ij} = \begin{cases} 1 & if\ e_i\ is\ relational\ to\ e_j \\ 0 & if\ e_i\ is\ relational\ to\ e_j \end{cases} \quad (4)$$

After this, we use the core of ISM and establish the reachable matrix M , then element $m_{ij}$ in n×n reachable matrix M meets the condition,

$$m_{ij} = \begin{cases} 1 & if\ e_i\ is\ relational\ to\ e_j \\ 0 & if\ e_i\ is\ relational\ to\ e_j \end{cases} \quad (5)$$

And M possesses the following properties:

(1) Generally for arbitrary integer r, if $e_i$ is reachable to $e_j$ , and the length between $e_i$ and $e_j$ is r, then $m_{ij}$ is 1.

(2) For a cyclic system, when k increases, M (that is $dist^k$) changed periodically until it becomes invariant.

(3) For an acyclic system, M = 0 ( $dist^k = 0$ ) when k increases to a certain value.

So we measure the similarity of faces based on these properties, and we can get the reachable matrix M by multiplication of adjacency matrix dist until M becomes invariant or is equal to 0.

We give the definition of matrix multiplication (⊗) like this :

$$\text{dist} \otimes \text{dist} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{21} & \cdots & e_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nn} \end{bmatrix}^2 = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{21} & \cdots & b_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \quad (6)$$

where $b_{ij} = bool(\sum_{k=1}^{n} e_{ik} e_{kj})$ , that is, after this matrix multiplication, we acquire element $b_{ij}$ by taking the bool value of the sum. And if $b_{ij} = 1$, it means $e_i$ is one-step reachable to $e_j$ . Otherwise, if $b_{ij} = 0$, there are two occasions: (a) $e_i$ is not related to $e_j$ ; (b) $e_i$ is K-step relational to $e_j$ where K > 1 and K is positive integer.

And then we multiply matrix dist again and again until it becomes invariant when k=K, the equation is as follows:

$$\text{M} = \text{dist} \otimes \text{dist} \otimes \cdots \otimes \text{dist} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{21} & \cdots & e_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nn} \end{bmatrix}^K = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{21} & \cdots & m_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{nn} \end{bmatrix} \quad (7)$$

Therefore, we could find the final relation in the reachable matrix. If $m_{ij} = 1$, we consider that $e_i$ is K-step relational to $e_j$ and we believe they belong to the same identity. Otherwise, we consider that $e_i$ is not relational to $e_j$ and they are defined as different identities and divided into different groups. The flow diagram of clustering process is illuminated as Figure.6.

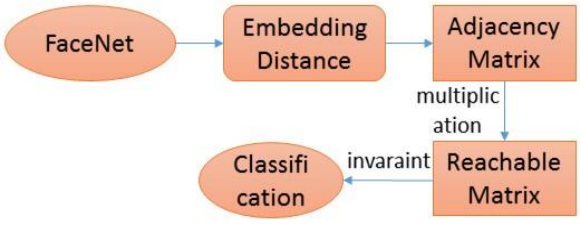


**Figure 6. This flow diagram shows the process of our clustering step.**

## 3.4 Face Recognition

Face recognition is the final part of our integrated system. It realizes the aim of face verification and provide critical information for our system to make important decisions in intelligence analyst. In this part, we set up our own dataset and train our own classifier on this dataset.

### *3.4.1 Face dataset*

Firstly, we examine our system with the face images captured by our own crawler. The dataset we create consists of 43 leaders in China and other surrounding countries. And for each leader, we collect over 400 images in different illumination and poses. But most images are in recent years, because only current state is related to our analysis.

Noticeably, we cannot be certain that images we crawl are clean enough because there is maybe another person in one person's group. So we clean the dataset manually and input this dataset to detection module to extract faces in frames or images and establish our face dataset. After we get all these cropped faces, we also need to clean each group by the person identity because there are maybe other people in group photos.

### *3.4.2 Classifier*

FaceNet uses a deep CNN consisting of two different core architectures: The Zeiler&Fergus[19] style networks and the recent Inception [3] type networks. The details of these networks are described in [1]. Given the model details, we employ the triplet loss used in [21] to achieve end-to-end learning which is the most important point in recognition part. The selection of triplet loss is important for the training of networks, and the loss used in [21] is more suitable for our system through both experiments and logic reasoning. And the encouragement is that the loss motivates all faces of each identity to be projected into a single point in the embedding Euclidian space. To optimize this end, the loss is finally minimized as Eq.(8), the detailed principle and implementation are in [1].

$$\forall(\text{f}(x_i^a), \text{f}(x_i^p), \text{f}(x_i^n)) \in \text{T} \quad (8)$$

$$\text{L} = \sum_i^N [||\text{f}(x_i^a) - \text{f}(x_i^p)||_2^2 - ||\text{f}(x_i^a) - \text{f}(x_i^n)||_2^2 + \alpha]_+ \quad (9)$$

Our approach has made some modification in this recognition step, so we need to train our own classifier to categorize these face collections. At first, we prepare to alter the net structure, but it's quite impossible to reuse the pretrained model by modifying parameters in early layers. To suit our purpose, we follow the instructions on the wiki[22] and generate our own model directly without retraining the final layer of the previous model. And then we modify the classifiy method of FaceNet to recognize each group acquired from clustering step to get respective identity. Figure.7 shows the process of face recognition.

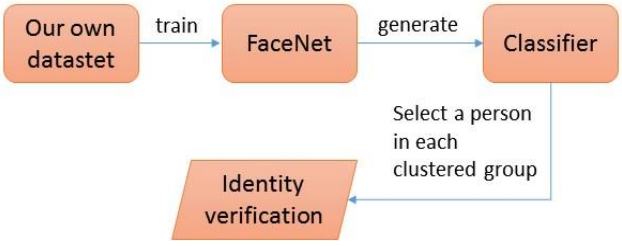


**Figure 7. This flow diagram shows the process of our recognition step.**

# 4. EXPERIMENT

## 4.1 Prepare Our Own Dataset

We collected about 12,000 images containing 43 persons in natural scenes from web. Then we input these images to face

detection module to obtain our face dataset. However, there are maybe multiple faces in one image, so we need to remove people which don't belong to the corresponding person's identity. Details of the raw and the cleaned dataset are illustrated in Table.1.

**Table 1. The face numbers of our datasets in three cases: raw data, cropped data, input data**

| Persons | Raw images | Cropped images | Input images |
|---|---|---|---|
| 1 | 344 | 813 | 49 |
| 2 | 234 | 537 | 39 |
| 3 | 480 | 1272 | 49 |
| 4 | 152 | 407 | 19 |
| 5 | 197 | 539 | 49 |
| 6 | 216 | 573 | 49 |
| 7 | 277 | 649 | 50 |
| 8 | 228 | 667 | 31 |
| 9 | 203 | 581 | 16 |
| 10 | 349 | 761 | 50 |
| 11 | 269 | 913 | 50 |
| 12 | 341 | 632 | 50 |
| 13 | 228 | 611 | 36 |
| 14 | 342 | 994 | 47 |
| 15 | 302 | 1037 | 827 |
| 16 | 290 | 752 | 49 |
| 17 | 341 | 934 | 49 |
| 18 | 361 | 1147 | 50 |
| 19 | 1509 | 3222 | 47 |
| 20 | 293 | 713 | 15 |
| 21 | 221 | 403 | 47 |
| 22 | 204 | 534 | 50 |
| 23 | 295 | 663 | 50 |
| 24 | 246 | 575 | 18 |
| 25 | 197 | 538 | 50 |
| 26 | 218 | 603 | 37 |
| 27 | 119 | 384 | 40 |
| 28 | 278 | 1135 | 49 |
| 29 | 263 | 1036 | 50 |
| 30 | 224 | 954 | 49 |
| 31 | 298 | 1183 | 51 |
| 32 | 216 | 836 | 51 |
| 33 | 382 | 847 | 50 |
| 34 | 161 | 265 | 37 |
| 35 | 31 | 84 | 16 |
| 36 | 132 | 292 | 50 |
| 37 | 201 | 416 | 49 |
| 38 | 250 | 567 | 50 |
| 39 | 310 | 713 | 48 |
| 40 | 91 | 199 | 33 |
| 41 | 206 | 562 | 49 |
| 42 | 281 | 289 | 90 |
| 43 | 216 | 836 | 51 |

Each of the numbers in column Person has its corresponding identity in our dataset list. We can notice that preprocessed data is far smaller than raw data and cropped data. This is because some of images we collect is much fewer than others resulting from their limited resources on Internet. So we balance the face numbers in each group to prevent overfitting. In this experiment test, we use these images and their flipped versions to realize data augment for training.

## 4.2 Experimental Results and Discussions

The implementation of our deep learning architecture is based on tensorflow which can be tested on CPU or GPU. We make experiments on CPU to evaluate its practical performance because some places lack GPU devices in real life and we want to make our system applicable to general occasions.

To make an evaluation of the real time performance of our end-to-end system, we test the time spent by different steps after inputting numbers of videos and images. For face detection, we test 22 videos and we count the total faces in frames. We get only 35 errors out of total 2410 faces in frames and most of the error faces are side faces, blurring or low-resolution faces. The detection accuracy of 98.55% is quite perfect. For face clustering, we test about 30 videos. The number of clustered groups are 316, and there are about 40 plus faces misplaced, so we get the quite satisfactory clustering rate of approximately 87.3%. For recognition part, we test about 300 persons, and there are about 40 people are false recognized, so the final accuracy of face recognition is about 87.9%. Table.2 illustrates a video example of the time consumption in each part and their accuracies. We can see that clustering time is larger because it needs to reload the model and amounts of math computations will occupy many resources. But it benefits much to face recognition module when computing facial feature embeddings. Since we got such excellent rate of test evaluation for video, the performance for images is definitely better than video because its high resolution and quality than compressed frames. Most important of all, the running time of our system is faster than systems consisting of only face detection and face recognition. We test about three videos and two images. Table.3 shows the comparison between ours and other systems without clustering step where number 1, 2, 3 presents video and number 4, 5 presents image. It can be seen that there is a great improvement between ours and others with large decrements, and the longer the time of video, the larger the difference value of processing time between ours and other systems. However, there are not obvious changes for images because not many people need to be detected and clustered in one

image, which suggest that our system is better for those occasions when massive data need to be analyzed.

**Table 2. The system performance in three major parts.**

| Step | Duration | Accuracy |
|---|---|---|
| Video | 190s | - |
| Video framing | 2.89s | - |
| Face detection | 7.34s | 0.986 |
| Face clustering | 27.69s | 0.873 |
| Face recognition | 21.23s | 0.879 |

**Table 3. The system performance with and without clustering step.**

| Test number | Duration time | Processing time with clustering | Processing time without clustering |
|---|---|---|---|
| 1 | 107s | 42.3s | 62.4s |
| 2 | 154s | 56.4s | 82.1s |
| 3 | 201s | 58.9s | 93.7s |
| 4 | - | 34.0s | 33.8s |
| 5 | - | 37.5s | 39.6s |

The results can suggest that our end-to-end FDCR system is excellent on our dataset. And our dataset contains all scenarios with all illumination, variations and occlusions which is robust to most natural scenes.

In conclusion, our system is very applicable to general occasions in real life and it will provide exact evidence and clue for further analysis in intelligence analyst.

# 5. CONCLUSION

In this paper, we design an end-to-end FDCR system based on FaceNet. The input of our system is various which can be video or image or their collections, so it's very practical in industrial field. And above all, we propose a clustering method to arrange the redundant data which shortens the runtime of our experiments greatly. Then we establish our own dataset and clean it based on aligned detection, and train our own classifier on cleaned dataset. The results of experiments show the good performance of our system and prove the effectiveness of this method.

In the future, we can also build our face relation graph based on this system and make an analysis of this relation chain, thus reasoning some truth and tracking the critical clue. In addition, this can also be used as abnormality detection for monitoring some significant situations. Furthermore, our work only contains face manipulation, we could make a holistic system to deal with event prediction which needs not only face techniques but also time series analysis and relation reasoning. It requires more knowledge and abilities to make more consummate of our practical system and we will make our best to improve this later.

# 6. ACKNOWLEDGMENTS

We would like to thank Maarten Bloemen for his discussions and great insights on face recognition and Florian Schroff, Dmitry Kalenichenko and James Philbin for providing network architectures like [14] and discussing network design choices. And our work is primarily supported by National Natural Science Foundation of China(NSFC) with grant number 61671459.

# 7. REFERENCES

[1] Schroff, Florian, Dmitry Kalenichenko, and James Philbin."Facenet: A unified embedding for face recognitionand clustering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[2] Huang, C., Ai, H., Li, Y., Lao, S.: High-Performance Rotation Invariant Multiview Face Detection. IEEE Transactions on PAMI 29, 671-686 (2007).

[3] Y. Tian, W. Liu, R. Xiao, F. Wen, and X. Tang. A face annotation framework with partial clustering and interactive labeling. In CVPR, pages 1-8, 2007.

[4] Shen, X., Lin, Z., Brandt, J., Wu, Y.: Detecting and Aligning Faces by Image Retrieval. In: Computer Vision and Pattern Recognition (2013).

[5] Zhu, X., Ramanan, D.: Face detection, pose estimation and landmark localization in the wild. In: Computer Vision and Pattern Recognition (2012).

[6] Xiong, X., DelaTorre, F.: Supervised descent method and its applications to face alignment. In: Computer Vision and Pattern Recognition (2013).

[7] Cao, X., Wei, Y., Wen, F., Sun, J.: Face Alignment by Explicit Shape Regression. In: Computer Vision and Pattern Recognition (2012).

[8] Y. Sun, X.W., Tang, X.: Deep convolutional network cascade for facial point detection. In: Computer Vision and Pattern Recognition (2013).

[9] Ren, S., Cao, X., Wei, Y., Sun, J.: Face Alignment at 3000 FPS via Regressing Local Binary Features. In: Computer Vision and Pattern Recognition (2014).

[10] Jain, V., Learned-Miller, E.: Fddb: A benchmark for face detection in unconstrained settings. Tech. Rep. UM-CS-2010-009, University of Massachusetts, Amherst (2010).

[11] Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In CVPR, volume 1, pages I-11, 2003.

[12] J. Wang, J. Wang, G. Zeng, Z. Tu, R. Gan, and S. Li. Scalable k-nn graph construction for visual descriptors. In CVPR, pages 1106-1113, 2012.

[13] C. Zhui, F. Wen, and J. Sun. A rank-order distance based clustering algorithm for face tagging. In CVPR, pages 481-488, 2011.

[14] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical view faces in the wild with deep neural networks. CoRR, abs/1404.3543, 2014. 2.

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In IEEE Conf. on CVPR, 2014. 1, 2, 5, 7, 8, 9.

[16] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. CoRR, abs/1406.4773, 2014. 1, 2, 3.

[17] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. CoRR, abs/1412.1265, 2014. 1, 2, 5, 8.

[18] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In NIPS. MIT Press, 2006. 2, 3.

[19] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013. 2, 3, 4, 6.

[20] Chen D, Ren S, Wei Y, et al. Joint cascade face detection and alignment. In ECCV, 2014.7.

[21] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. CoRR, abs/1406.4773, 2014. 1, 2, 3.

[22] https://github.com/davidsandberg/facenet/wiki/Train-a-classifier-on-own-images.

[23] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In CVPR, pages 2707-2714.IEEE, 2010.

[24] B. S. Swann. FBI video analytics priority initiative. In 17th Annual Conference & Exhibition on the Practical Application of Biometrics, 2014.

[25] C. Zhui, F. Wen, and J. Sun. A rank-order distance based clustering algorithm for face tagging. In CVPR, pages 481-488, 2011.

[26] Li L, Zhang J, Fei J, et al. An incremental face recognition system based on deep learning[C]// Fifteenth Iapr International Conference on Machine Vision Applications. 2017:238-241.

[27] Otto C, Klare B, Jain A K. An efficient approach for clustering face images[C]// International Conference on Biometrics. IEEE, 2015:243-250.