Large Scale Face Data Purification based on Correlation Function and Multi-Phase Grouping

Xiangxiang Zhang[1], Zhijun Fang[1,a], Zhenghao Xi[1] and Xiaoshuang Liu[1]

[1]*School of Electronic and Electrical Engineering, Shanghai University of Engineering Science*

Abstract. Recent advances in deep learning technologies enable high performance artificial intelligence, which is an equivalence of human capability or higher for various application. However, deep learning is highly resorted to the large scale training data, which typically contains large number of outlier samples that are difficult to remove. In this paper, we proposed a face image purifying algorithm, which combines the correlation function of deep features with multi-phase grouping technique. A correlation function was proposed to determine the principal class by measuring the similarities between all different samples. The principal class was further used as a prior for the multi-phase grouping algorithm to purify the face data by multiple thresholds. The experimental results demonstrate that the proposed algorithm has significant improvement than the primitive cluster algorithm, such as K-Means.

1 Introduction

Artificial intelligence and machine vision are the frontier of modern industrial and information science, and face recognition technology is one of the earliest implementations of machine vision. Recently, with the development of artificial intelligence in the dynamic face recognition and facial portrait differentiation technology, the face recognition technology has found a new research direction again [2-3]. Furthermore, combining the traditional face recognition technology with the large data technology, training the face features with a large number of samples, and improving the face recognition at the performance has been an issue problem in the field of machine vision research.

However, there are lots of non-target images in the collected face training samples. We define the non-target images in the same kind of face images as the noisy samples, which have the similar features with the target images. The sample noise has seriously affects the training results. At present, how to remove the sample noise in a large number of training samples has become a new focus of face recognition technology [4].

Many recent researchers proposed some approaches for this problem: Fitzgibbon and Zisserman [5] proposed a sample purification method based on Joint Manifold Distance (JMD), this method treats a group of face data as a subspace, and performs face clustering by calculating JMD. Wu B et al. [6] established a clustering model using the probabilistic constraint conditions of the hidden Markov random field, they classified all face sample images into K disjoint clusters, and the K should be given in advance. Because they cannot obtain the accurate K value in advance, the results are very unstable. In [7], Zhang et al. clustered the shape of the face to seven classes. Then, they continue to subdivide the smaller classes according to the facial features extracted by the ASM method. This method relies on the facial contour, but the facial expression is varied, and the influence of light, occlusion and other factors, the deep feature of human face is not easy to extract. Therefore, the method has some limitations.

In addition, some researches combine intelligent algorithm with clustering method. [8] first extracted the deep characterization features of face images using convolution neural network (CNN) algorithm, then they combined the feature with the clustering method (e.g. K-MEANS and hierarchical clustering method) to achieve clustered the face images. However, the problem of this method is similar with [6], the K-MEANS needs to estimate the number of clusters in advance, and the clustering results are more sensitive to the initial values. The hierarchical clustering with different types of feature similarity can classify the data into different clusters, but it cannot be re-classified, the result of cluster purification is not satisfactory. Even so, compared with the method only using clustering algorithm, the methods combined intelligent algorithm and clustering algorithm improves the accuracy and efficiency of sample purification to a certain extent.

In this paper, we proposed an algorithm that combines the correlation function of deep features with multi-phase grouping technique to solve the problems of sample purification. Firstly, we use the VGG Face [1] network to extract the high-level feature information of face images, and we define the correlation function used to compute the principal class. Then, we choose the appropriate multiple thresholds to divide the data into

[a] Corresponding author: zjfang@sues.edu.cn

groups. Finally, we remove the noisy data and get the purification of the sample images by filtering noisy samples in each group.

The remainder of the paper is organized as follows. Section 2 describes feature extraction of face images and similarity measure. The proposed face image purifying algorithm is introduced in detail in section 3, which includes definition of the correlation function、analysis of principal class 、multi-phase grouping and setting thresholds. Section 4 shows the experimental results.

2 Face Feature Extraction

2.1 Feature Extraction

CNN is an efficient intelligent algorithm, which has been developed rapidly recent years. Because of its deep structure, strong learning ability and hierarchical nonlinear mapping, CNN has been widely used in facial feature extraction [1-3], and it becomes the main method of face recognition [9]. In this paper, we establish the face feature extraction network by the CNN of VGG FACE [1], and the CNN with 13 convolution layers, each convolution layer contains a linear operator, which followed by one or more nonlinear operator, such as ReLU. The last three blocks are fully connected layers (FC). This paper uses the fc7 FC layer to extract the face image features, [1] shows that the face recognition accuracy rate is 97% using this method in the Youtube Faces Dataset. So, we combine it with our proposed face image purification algorithm to refine the face data.

2.2 Similarity measure

We use the CNN method to extract high-level feature $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_j,...,\mathbf{x}_n\}$, where, $\mathbf{x}_j$ is the feature vector of the j^{th} image sample. $j=1,2,3,...,n$. Then the distance between any two samples can be expressed as $dist(\mathbf{x}_i,\mathbf{x}_j)$, and the similarity between them is

$$eachsim(\mathbf{x}_i,\mathbf{x}_j)=1-dist(\mathbf{x}_i,\mathbf{x}_j) \quad (1)$$

where, $dist(\mathbf{x}_i,\mathbf{x}_j)$ is the Euclidean distance after $\mathbf{x}_i,\mathbf{x}_j$ are normalized.

3 Face image purifying algorithm

3.1 Defining the correlation function

Let the face data set as $\boldsymbol{R}=\{c_1,c_2,c_3,...,c_i,...,c_m\}$, where, c_i is the image samples of the i^{th} category, $i=1,2,3,...,m$. The number of the images in c_i is a_i, $i=1,2,3,...,m$. The proportion of the i^{th} category image in the face data set is

$$P(c_i)=a_i/\sum_{k=1}^{m}a_k \quad (2)$$

When $P(c_i)$ is the maximum, c_i is the target sample, and the other category images are taken as noisy samples.

In order to automatically and unsupervised find the face image samples, which distributed densely in the current data set, we define the correlation function, which is the accumulation of similarity of each sample with the other samples, the correlation function for any $\mathbf{x}_i$ is

$$Correlation(\mathbf{x}_i)=\sum_{j=1}^{n-1}eachsim(\mathbf{x}_i,\mathbf{x}_j) \quad (3)$$

where, $i\neq j$, $eachsim(\mathbf{x}_i,\mathbf{x}_j)$ is the similarity between the two feature vector $\mathbf{x}_i$ and $\mathbf{x}_j$.

The data can be divided into several different classes by pre-setting the number of clusters, but this method may cause the data which in a same class to be divided into several categories, or be classified into other cluster so that they become noise. Once the number of clusters is pre-determined, the data samples of the misclassified categories will not be able to be back. This will affect the effect of data purification. The correlation function can reflect correlation of each sample in the global situation, and can estimate principal class among the sample classes. This avoids splitting the same samples into other classes, reduces the misclassification probability of the samples.

3.2 Analysis of principal class

Let $\mathbf{S}=\{\mathbf{f}_1,\mathbf{f}_2,...,\mathbf{f}_i,...,\mathbf{f}_N\}$, $1\le i\le N$ as the set of feature vectors for training samples set, where, $\mathbf{f}_i$ is the feature vector of the i^{th} face image in training samples, N is the number of the samples in the training samples set. We get the similarity $eachsim(\mathbf{f}_i,\mathbf{f}_j)$ of any two different training feature vectors by Eq. (1), where $i\neq j$. The correlation function $Correlation(\mathbf{f}_i)$ of each training feature vector by Eq. (3), and we can get the main component sample by maximizing the $Destin(\mathbf{f}_i)$ as follows

$$\mathbf{f}_{\max}=Max\{Correlation(\mathbf{f}_i)\} \quad (4)$$

where, $\mathbf{f}_{\max}$ is the training feature vector of face images and the target sample vector when the correlation function is maximized.

3.3 Multi-phase grouping

In the dataset $\boldsymbol{R}$, we classify the main component samples by $\mathbf{f}_{\max}$, and let the classified main component samples set be $\boldsymbol{A}$, the rest of the samples will belong to set $\boldsymbol{B}$, $\boldsymbol{R}=\boldsymbol{A}+\boldsymbol{B}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ both are the proper subsets of the data set.

Let the threshold as T_1, we compare the similarity between $\mathbf{x}_i$ and $\mathbf{f}_{\max}$ by Eq. (1), if the similarity is greater than T_1, then the sample will belong to $\boldsymbol{A}$, if not, it will belong to $\boldsymbol{B}$, which can be expressed as

$$eachsim(\mathbf{x}_i, \mathbf{f}_{\max}) = \begin{cases} \geq T_1, \mathbf{x}_i \in \boldsymbol{A} \\ < T_1, \mathbf{x}_i \in \boldsymbol{B} \end{cases} \quad (5)$$

We initially purify samples of the principal class by T_1, and divide the $\boldsymbol{R}$ into $\boldsymbol{A} = \{\mathbf{f}_{A_1}, \mathbf{f}_{A_2}, \mathbf{f}_{A_3}, ..., \mathbf{f}_{A_i}, ..., \mathbf{f}_{A_G}\}$, and $\boldsymbol{B} = \{\mathbf{f}_{B_1}, \mathbf{f}_{B_2}, \mathbf{f}_{B_3}, ..., \mathbf{f}_{B_j}, ..., \mathbf{f}_{B_D}\}$, and $\boldsymbol{A}$ is the principal class and $\boldsymbol{B}$ is the group to be filtered, where $\mathbf{f}_{A_i}$ $1 \leq i \leq G$ is the feature vectors which are divided into the principal group of $\boldsymbol{A}$, and f_{B_j} $1 \leq j \leq D$ is the feature vectors which are divided into the group to be filtered of $\boldsymbol{B}$, $G + D = N$. However, the face change or an unclear image will lead to diversification of face images, which makes it be more difficult to comprehensively characterize multiple principal samples. In this paper, we use the centroid of the feature vectors, and the strict thresholds to solve the problem of multi-features fusion when deciding whether each sample is a main component target sample. Let $\mathbf{O}_1$ be the centroid of $\boldsymbol{A}$, then, $\mathbf{O}_1$ can be expressed as

$$\mathbf{O}_1 = \frac{1}{G} \sum_{i=1}^{G} \mathbf{f}_{A_i} \quad (6)$$

To minimize the loss of the number of samples of principal class as much as possible, we check for the principal sample images that may be left in the set $\boldsymbol{B}$, and we purify the $\boldsymbol{B}$ group again. Let the threshold as $T_2 T_3$, then we compare $\mathbf{O}_1$ with each of feature vectors in the set $\boldsymbol{B}$, and filter the target samples, which can be expressed as follows

$$eachsim(\mathbf{O}_1, \mathbf{f}_{B_j}) = \begin{cases} \geq T_2, \mathbf{f}_{B_j} \in A \\ < T_2, \mathbf{f}_{B_j} \in B \end{cases} \quad (7)$$

Let the set of the second purifying principal class be $\mathbf{A}' = \{\mathbf{f}_{A'_1}, \mathbf{f}_{A'_2}, \mathbf{f}_{A'_3}, ..., \mathbf{f}_{A'_i} ..., \mathbf{f}_{A'_E}\}$, $1 \leq i \leq E, G \leq E \leq N$. We purify the group of principal class and update the centroid to $\mathbf{O}_2$. Then, let the threshold as T_3, the principal class will be purified as follows

$$eachsim(\mathbf{O}_2, \mathbf{f}_{A'_i}) = \begin{cases} \geq T_3, \mathbf{f}_{A'_i} \in A' \\ < T_3, \mathbf{f}_{A'_i} \in B \end{cases} \quad (8)$$

We divide the samples into different groups by the different thresholds to be purified, and classify the samples of principal class from each group, so that the purity of the samples is improved continuously. The multi-phase grouping does not only extend the interclass distance, but also minimizes distance within the class. The flowchart of the proposed algorithm is shown in Figure 1.

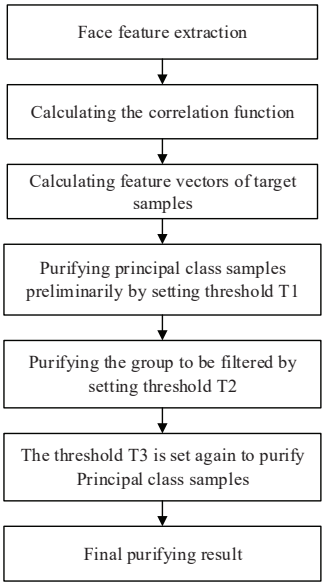

Figure 1. Flowchart of our face purifying algorithm.

3.4 Thresholds setting

Our thresholds can strictly filter the noisy data, and prevent the noisy data from contaminating the target samples. Figure 2 shows the distribution of similarity between each two samples. The horizontal axis represents the similarity values, and the vertical axis represents the number of samples. When the threshold is below 0.6, the similarities of samples are low, this part represents most of the noisy data. When the threshold is above 0.7, the similarities of samples are high, this part represents target samples. In this paper, we get the threshold between 0.7 and 0.8 by the experience.

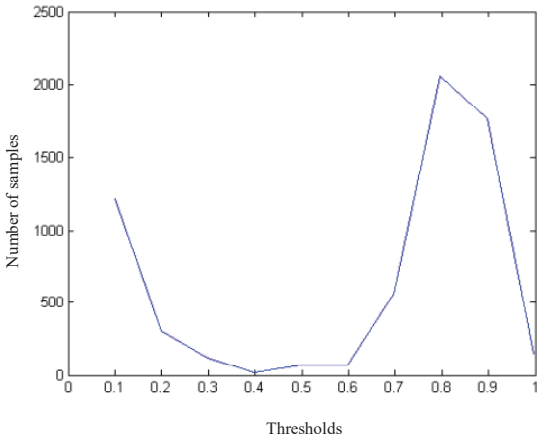

Figure 2. The distribution of similarity between each two samples.

4 Experimental results and analysis

The effect of the proposed algorithm is evaluated using precision and recall of [9]. The precision and recall of class i in the j^{th} cluster are defined as follows

$$P = precision(i, j) = N_{ij} / N_i \quad (9)$$

$$R = recall(i, j) = N_{ij} / N_j \quad (10)$$

where, N_{ij} is the number of class i in the j^{th} cluster, N_j is the number of the samples, N_i is the number of target samples in class i , the F-measure of the class i is

$$F(i) = 2PR / (P + R) \quad (11)$$

The experimental database was selected from the opened face data MS-Celeb-1M [10], which is provided by Microsoft Research Institute and includes images of 1,000,000 celebrities. In our experiment, we randomly select 10,000 face images of the data.

In our experiment, we keep the T_1 unchanged, adjust T_2 and T_3 to obtain the satisfactory purification results. Figure.3 shows the F-measure results using different thresholds.

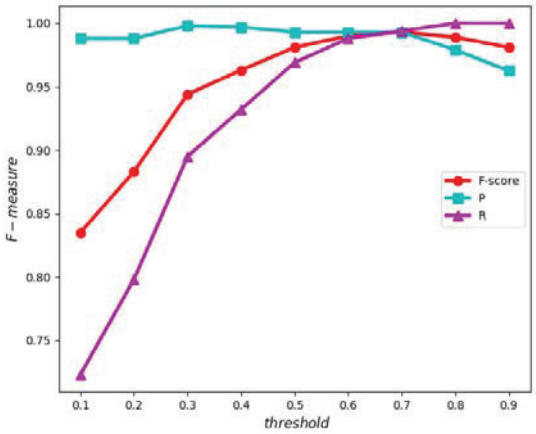

Figure 3. The F-measure results using different thresholds

The F-score reflects the result of the face images purification, recall reflects the purity of the data, and precision reflects the number of target samples loss. However, the result of the purification is the best when the threshold is increased to about 0.7.

We select part of the samples whose similarities are so high to compare with K-MEANS method, and to calculate recall rate of principal samples, the result is shown in Figure.4.

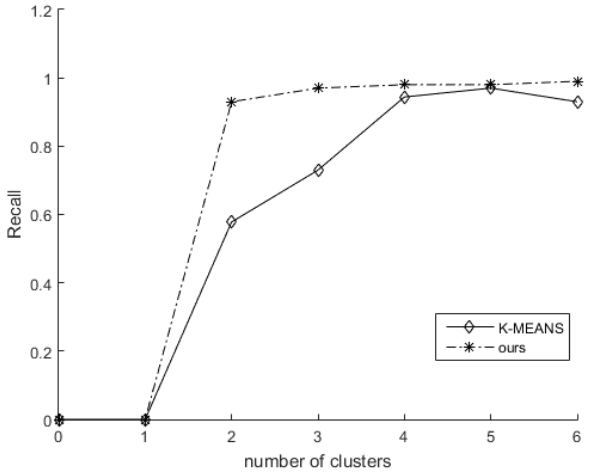

Figure 4. Recall rate of different clustering results

5 Conclusion

Data purification is an important method for big data analysis and application. To reduce the data loss and purify face images as much as possible, we proposed a simple yet effective purification algorithm for face data. We used this novel algorithm to purify the target face samples in big data samples by selecting the appropriate thresholds. The experiment results indicate that the proposed algorithm is helpful for improving the rates of precision and recall. The highlights of this paper are that the proposed algorithm is more accurate and stable than traditional clustering algorithm.

References

1. Parkhi O M, Vedaldi A, Zisserman A. Deep Face Recognition[C]// British Machine Vision Conference. 2015:41.1-41.12.
2. Taigman Y, Yang M, Ranzato M, et al. DeepFace: Closing the Gap to Human-Level Performance in Face Verification[C]// Computer Vision and Pattern Recognition. IEEE, 2014:1701-1708.
3. Schroff F, Kalenichenko D, Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering[C]// Computer Vision and Pattern Recognition. IEEE, 2015:815-823.
4. Frénay B, Verleysen M. Classification in the presence of label noise: a survey[J]. IEEE Transactions on Neural Networks & Learning Systems, 2014, 25(5):845-869.
5. Fitzgibbon A W, Zisserman A. Joint manifold distance: a new approach to appearance based clustering[A].Fitzgibbon A W. IEEE Conference on Computer Vision and Pattern Recognition[C]. Madison: IEEE, 2003.
6. Wu B, Zhang Y, Hu B G, et al. Constrained Clustering and Its Application to Face Clustering in Videos[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2013:3507-3514.
7. Zhang S C, Fang B, Liang Y Z, et al. A face clustering method based on facial shape information[C]// International Conference on Wavelet Analysis and Pattern Recognition. IEEE, 2011.
8. Hu Y, Dong Y. Face clustering using high-level feature based on deep learning[J]. 2015.
9. STEINBACH M, KARYPIS G, KUMAR V. A comparison of document clustering techniques[C] / /Proc of KDD-2000 Workshop on Text Mining. 2000.
10. Guo Y, Zhang L, Hu Y, et al. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition[C]// European Conference on Computer Vision. Springer, Cham, 2016:87-102.