K-L Divergence Based Image Classification and the Application

[1]*Fuhua Chen*[1], *Xuemao Zhang*[2] *and Guangtai Ding*[3]

[1]*Department of Physical Sciences & Math, West Liberty University*
West Liberty, WV 26074, USA
[2]*Department of Mathematics, East Stroudsburg University*
East Stroudsburg, PA 18301, USA
[3]*School of Computer Engineering and Science, Shanghai University*
Shanghai, 200444, China

[1]*fuhua.chen@westliberty.edu*, [2]*xzhang2@esu.edu*, [3]*gtding@shu.edu.cn*

Abstract

Image classification is widely used in many fields. Traditional metric learning based classification methods always maximize between-class distances and minimize within-class distances based on features calculated from each individual. Different from traditional methods, this paper takes each class as a distribution and try to maximize the distances among different distributions using information geometry. In order to minimize the distance among individuals within a class, the paper assumes that each class follows a joint Gaussian distribution and takes an exploratory study on the relation between a within-class distance and the determinant of the covariance matrix of the distribution. It is found that under some assumptions, the average within-class distance among the same class is proportional to the standard deviation (for a random variable) or the product of standard deviations of each feature (for a random vector). As a result, the standard deviation (for a random variable) or the determinant of the covariance matrix (for a random vector) is used to substitute the within-class distance in the metric learning. The proposed method thereinafter saves a lot of computational cost. The method is then applied to person re-identification, which is a very important application in a 5G time, such as smart city. To our surprise, the proposed method is very competitive compared with many state-of-the-art methods while saving the computational cost in the learning stage. Experimental results demonstrate the effectiveness of the proposed method.

Keywords: *Classification, Gaussian distribution, metric learning, K-L divergence, Mahalanobis-distance, Person re-identification.*

[1] *Corresponding author: fuhua.chen@westliberty.edu, Peer-editor: Xuemao Zhang.*

1. Introduction

Image Classification is a fundamental task that attempts to comprehend an entire image as a whole. The goal is to classify the image by assigning it to a specific label, called classification. Typically, Image Classification refers to images in which only one object appears or needs to be considered or analyzed. Examples of research in image classification include human face recognition, handwriting digits recognition, handwriting words recognition, vehicle recognition, object detection, tumor recognition, and so on. Image classification has been applied in many fields, such as medical science, remote sensing, autopilot, surveillance, and so on. Taking handwriting digits as an example, there are many different patterns for same digits. Fig.1 below shows different forms of digits being presented. Sometimes, the difference among the presented forms of the same digit can be huge. It can be easy for a person to recognize a handwriting digit. However, it may not be easy for a computer to recognize it automatically. Image classification is to develop an algorithm for a computer or a robot to automatically recognize an object.

Figure 1: Handwriting digits (http://yann.lecun.com/exdb/mnist/)

Another example of image classification is human face recognition. Fig.2 is a list of human faces. It is easy to figure out a person that you are familiar with from a short list of faces. However, if there is a huge number of faces and you want to find the particular one, it will take you a lot of time and won't be an easy work. On the contrary, it will be much faster for a computer to find a particular face from a face database even with a huge size. Meanwhile, when the database of human faces is very big, containing a huge number of faces, many faces can be very similar. It is then very challenging for a computer or a robot to recognize the particular face.

Figure 2: Human faces (https://www.wired.com/story/artificial-intelligence-fake-fakes/)

Many traditional classification methods can be used in image classification, such as the k-nearest neighbors method (kNN) [1] and the large margin nearest neighbor (LMNN) classification [2]. But image based classification has its own characteristics and can be solved more effectively using some variations of traditional classification methods. Current image classification methods can be divided into two categories: either concentrated on hand-crafted system, where features are extracted by users, or based on deep learning techniques [3-7], where features are extracted automatically by a neural network. Deep learning techniques usually have a higher accuracy but the cost is also higher. Therefore, hand-crafted systems are still very useful and the related research is still active. Among hand-crafted systems, there are also two big categories: (1) Feature-concentrated methods [8-11]; and (2) Metric-learning based methods [12-17]. Feature-concentrated methods focus on extracting discriminative features to improve classification, while metric-learning based methods focus on learning an effective metric. Given a collection of classes and a given object, by computing the distance between the object and each class, the object is assigned to the class that has the shortest distance from the object.

There are two major drawbacks for existing metric-learning based methods. First, current metric-learning based methods always directly use distances between a pair of instances. But different classes may have different number of instances, which causes an imbalanced computation for each class [16, 18]. Second, even if each class has the same number of instances, there is still imbalanced computation between distances for between-class instances and distances for within-class instances [15]. Moreover, when the number of instances for each class increases a little, the computational complexity will increase a lot. Suppose there are totally N classes and each class has K instances. If one particular class increases one instance, then there are K more within-class distances needed to be calculated from the same class and

$(N-1)K$ more between-class distances needed to be calculated from different classes. Fig.3 shows the imbalanced situation when one more object is added to Class A.

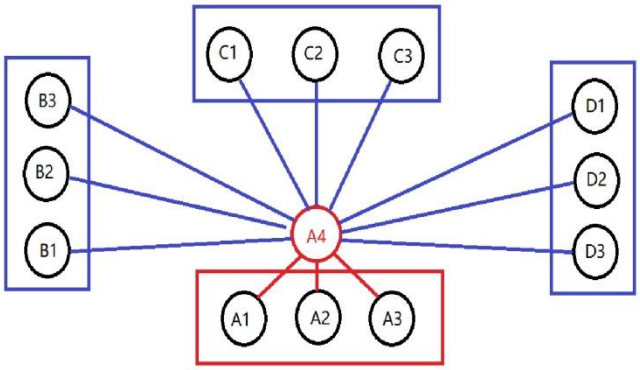

Figure 3: Imbalanced situation. Suppose there are totally four classes ($N = 4$): A, B, C, and D, each class containing three instances ($N = 3$). When Class A increases one instance A4, there are $K = 3$ more pairs to be added within Class A, which are (A1, A4), (A2, A4), and (A3, A4). However, there will be $(N-1)K = (3)3 = 9$ more pairs to be added, which are (B1, A4), (B2, A4), (B3, A4), (C1, A4), (C2, A4), (C3, A4), (D1, A4), (D2, A4), and (D3, A4).

In order to overcome these drawbacks, this paper takes each class as a Gaussian distribution. The distance between two classes is denoted by the distance between two distributions, which avoids unbalanced computation when different classes have different number of instances and also avoids unbalanced computation between inter-class distance and intra-class distance. Moreover, since each class is denoted by a single distribution, the computational cost can be greatly reduced. This paper aims at such applications in which there are fixed number of known classes, each of which has enough number of known instances (labelled ones). Given a new probe (unknown instance), the purpose of this work is to figure out which class the probe belongs to or is closest to. The major contribution of this paper lies in proposing a distribution based model in which the within-class distance is measured by the determinant of the shared covariance matrix. As an example of applications, the method is then applied to person re-identification, a very popular application on surveillance. The rest of the paper is organized as below. In Section 2, we derived the proposed model in detail. In section 3, we compared the proposed model with some related work. Section 4 is about the implementation, where many relevant issues are discussed. In Section 5, we introduced the topic about person re-identification and the application of the proposed method on it.

Experimental results are also shown in this part. Finally, there is a short conclusion in Section 6.

2. Model Development

Given a gallery of images, suppose there are totally N classes of images, denoted by $C_1, C_2, \cdots, C_N$, each of which contains a set of images. Given a probe image I, the general way classifying the image I is to extract its features, such as color and texture, and then use these features to make comparison with all known images. The comparison is based on metric. Such a metric can be a Euclidean distance or a non-Euclidean distance such as Mahalanobis distance. If the image I is closest to the representative of a particular class C_i $(i = 1,2, \cdots, N)$, then I will be classified to C_i. The task of an image classification is to find a metric such that within each class, the distances between any two images are small, while the distance between any two images that are located in different classes is large.

Suppose there are totally n different features are extracted for the classification. Let $V = \{(x_1, x_2, \ldots, x_n)\}$ be the feature space that are extracted for the classification. Assume the feature set of each class in each image is a random vector following a joint distribution. Based on the features of the samples from each class in the gallery, we formulate its probability density function (p.d.f.). Let $p_i(\boldsymbol{x}|\boldsymbol{\theta}_i)$ be the p.d.f. of the i-th class, where $\boldsymbol{x}$ is a continuous multivariate random variable of features and $\boldsymbol{\theta}_i$ stands for the set of parameters of the distribution. The Kullback-Leibler divergence (K-L divergence) from the i-th class to the j-th class for continuous multivariate random variable $\boldsymbol{x}$ is given by

$$D_{KL}(p_i||p_j) = \int_{-\infty}^{\infty} p_i(\boldsymbol{x}|\boldsymbol{\theta}_i) \log \frac{p_i(\boldsymbol{x}|\boldsymbol{\theta}_i)}{p_j(\boldsymbol{x}|\boldsymbol{\theta}_j)} d\boldsymbol{x} \tag{1}$$

Based on the geometric meaning of K-L divergence, a best classification should have a maximal total divergence, corresponding to the total between-class distance. In other words, we want to maximize the following energy functional F with respect to the parameter set θ.

$$F(\theta) = \sum_{i,j=1;i\neq j}^{N} D_{KL}\left(p_i||p_j\right) \tag{2}$$

Assume that the feature set of a class follows a joint Gaussian distribution. We use $\mathrm{p_i}(\mathbf{x}|\boldsymbol{\theta}) = \mathrm{p}(\mathbf{x}|\boldsymbol{\mu}_\mathrm{i}, \boldsymbol{\Sigma}_\mathrm{i})$ to denote the joint p.d.f. Namely,

$$p_i(\boldsymbol{x}|\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\mu}_i, \Sigma_i) = \frac{exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right)}{\sqrt{2\pi|\Sigma_i|}} \tag{3}$$

where **x** stands for a feature vector, $\boldsymbol{\mu}_i$ is the mean feature vector of the ith class, and Σ_i is the covariance matrix of the ith class among all features. Σ_i^{-1} and $|\Sigma_i|$ are the inverse matrix and the determinant of Σ_i, respectively. By these assumptions, the K-L divergence from the ith class to the jth class is simplified to

$$\begin{aligned} D_{KL}(p_i||p_j) = & \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \Sigma_j^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\ + & \frac{1}{2}\left(tr(\Sigma_j^{-1}\Sigma_i) - n + \ln\left(\frac{|\Sigma_j|}{|\Sigma_i|}\right)\right) \end{aligned} \tag{4}$$

where $\mathrm{tr}(\cdot)$ stands for the trace of a square matrix.

When the targets in those images all belong to a same category, such as human beings, we can further assume that the distributions of different classes share the same covariance matrix, denoted by Σ, and the only difference among different classes is their feature means $\boldsymbol{\mu}_i$. Under this discussion, the joint distribution of feature vectors for each class can be denoted by

$$p(\boldsymbol{x}|\boldsymbol{\mu}_i, \Sigma) = \frac{exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right)}{\sqrt{2\pi|\Sigma|}}, \tag{5}$$

where the mean vector $\boldsymbol{\mu}_i$ of each class's features can be estimated from known images (labeled samples for training). Then the K-L divergence (4) from the ith class to the jth class is simplified as follows.

$$D_{KL}(p_i||p_j) = \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \tag{6}$$

The K-L divergence functional (2) with joint Gaussian distribution can then be written as

$$F(\Sigma) = \sum_{i,j=1}^{N} \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{7}$$

Maximizing the between-class distance becomes maximizing the functional $\mathrm{F}(\Sigma)$ with respect to Σ, or equivalently, to solve the following optimization problem:

$$\min_{\Sigma} G(\Sigma) \triangleq \sum_{i,j=1}^{N} -\frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{8}$$

2.1. Distribution based within-class distance characterization with common covariance matrix

Eq. (8) only contains between-class distances, without any measurement of within-class distances. Therefore, Eq. (8) as a model for classification is not enough. Since we are developing a distribution based model, it would be nice to characterize the within-class distance also using distribution itself. As a matter of fact, the within-class distance of a distribution can be characterized by its variance or covariance matrix. To make it more clear, we first use one-dimensional Gaussian distributions as an example.

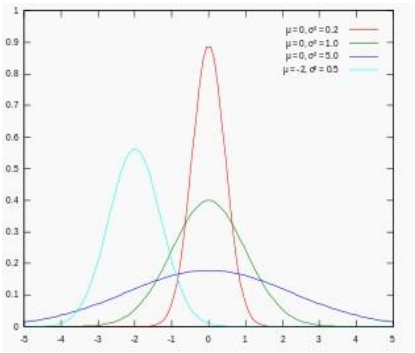

Figure 4: Shapes of one-dimensional Gaussian distributions with different variances.

Fig.4 contains four normal curves. From the perspective of statistics, we know that the variance of the red one is the smallest and the variance of the blue one is the largest. Also, the variance of a distribution actually characterizes the average distance of the instances of the distribution. To illustrate it more clearly, let X_1 and X_2 be two independent random instances of the same random variable X whose variance is σ^2. Consider the quantity $\sqrt{E[(X_1 - X_2)^2]}$, where $E[\cdot]$ stands for expectation. Then $\sqrt{E[(X_1 - X_2)^2]}$ stands for the average distance of the instances from the same variable X. We have the following theorem.

Theorem 1. Let X_1 and X_2 be two independent copies of the random variable X whose variance is σ^2. Then we have

$$\sqrt{E[(X_1 - X_2)^2]} = \sqrt{2}\sigma. \tag{9}$$

Proof. It is equivalent to prove that $\mathrm{E}[(\mathrm{X}_1 - \mathrm{X}_2)^2] = 2\sigma^2$. Let μ be the mean of the distribution of X. Then the mean of X_1 and the mean of X_2 are also μ. So,

$$
\begin{aligned}
& E[(X_1 - X_2)^2] = E\left[\left((X_1 - \mu) - (X_2 - \mu)\right)^2\right] \\
= \;& E[(X_1 - \mu)^2 - 2(X_1 - \mu)(X_2 - \mu) + (X_2 - \mu)^2] \\
= \;& E[(X_1 - \mu)^2] - 2E[(X_1 - \mu)(X_2 - \mu)] + E[(X_2 - \mu)^2] \\
= \;& 2\sigma^2.
\end{aligned} \tag{10}
$$

Definition 1. Let X_1 and X_2 be two independent copies of the random variable X whose variance is σ^2. Define the quantity $\sqrt{E[(X_1 - X_2)^2]}$ to be the average within-class distance of X.
We can generalize the concept of average within-class distance and the corresponding result to random vectors.

Definition 2. Let X_1 and X_2 be two independent copies of the random vector X having n components. Define the average within-class distance of X to be the geometric mean of the average within-class distances of each component. Namely,

$$
d(X_1, X_2) = \sqrt[n]{\prod_{i=1}^{n} \sqrt{E[(X_{1i} - X_{2i})^2]}} \tag{11}
$$

where $E[\cdot]$ stands for expectation and X_{1i} and X_{2i} are component random samples, $i = 1,2,\ldots,n$.
We have the following theorem.

Theorem 2. Let X be a random vector of features. Suppose all of the selected features are independent and the variances of each feature are $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectively. Then the average within-class distance of X is $\sqrt{2}(\sigma_1\sigma_2\ldots\sigma_n)^{1/n}$.
Proof. Let X_1 and X_2 be any two such random vector instances. According to the definition,

$$
\begin{aligned}
d(X_1, X_2) = \;& \sqrt[n]{\prod_{i=1}^{n} \sqrt{E[(X_{1i} - X_{2i})^2]}} \\
= \;& \sqrt[n]{\prod_{i=1}^{n} \sqrt{2}\,\sigma_i} = \sqrt{2}(\sigma_1\sigma_2\ldots\sigma_n)^{1/n}.
\end{aligned} \tag{12}
$$

Note that, when the features are independent, we have

$$
\Sigma = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2). \tag{13}
$$

Then we have the following relation between the average within-class distance and the determinant of the covariance matrix.

$$d_{within} = \sqrt{2}|\Sigma|^{1/2n}, \tag{14}$$

where d_{within} denotes the average within-class distance and $|\Sigma|$ denotes the determinant of Σ. Therefore, the covariance matrix can completely characterize the average distance among samples of a joint Gaussian distribution.

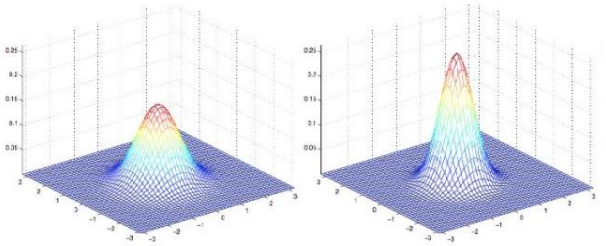

Figure 5: Shapes of two-dimensional Gaussian distributions with different variances.

Fig.5 shows two two-dimensional normal surfaces (Gaussian surfaces) where the left is lower but open wider and the right is higher but open narrower, which means that the average within-class distance of the left distribution is larger than the average within-class distance of the right one. From the expression of the p.d.f. of Gaussian distribution (3), we see that the shape of the distribution is completely determined by the covariance matrix (and the means determine the location (center) of the distribution only). As a matter of fact, whether or not the distribution is open wider or narrower, which corresponds to whether the within-class distance is larger or smaller, is determined by the determinant of the covariance matrix. Fig.5 indicates that the determinant of the covariance matrix for the left Gaussian distribution is larger while the determinant of the covariance matrix for the right one is smaller. Therefore, we can use the determinant $|\Sigma|$ to approximate the within-class "distance". Note that as a covariance matrix of a joint Gaussian distribution, the covariance matrix Σ satisfies $|\Sigma| > 0$.

Meanwhile, considering that the dimension n of features can be very large, the determinant term will be sensitive to the dimension of features and will affect the numerical implementation largely, we use $\ln|\Sigma|$ to control the within-class distance. So, our normal-distribution based image classification model that measures both between-class and within-class distance is given as below.

$$\min_{\Sigma} G(\Sigma) \triangleq \sum_{i,j=1}^{N} -\frac{1}{2}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T \Sigma^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) + \lambda \ln|\Sigma| \tag{15}$$

2.2. Regularization

To make a machine-learning work very well, it is well known that the sample used for learning should represent the population very well. Therefore, the common covariance matrix in Eq. (15) should be as close to each sample covariance matrix as possible. That means the following sum

$$\sum_{i=1}^{N} ||\Sigma - \Sigma_i||_F^2 \tag{16}$$

should be small, in which $||\cdot||_{\mathrm{F}}$ is the Frobenius norm. More simply, we use Σ_0 to denote the arithmetic mean of all the N covariance matrices Σ_{i}, namely,

$$\Sigma_0 = \frac{\Sigma_1 + \Sigma_2 \cdots + \Sigma_N}{N},$$

and force $||\Sigma - \Sigma_0||_{\mathrm{F}}^2$ to be small. The final Gaussian distribution based person re-identification model is to given below.

$$\min_{\Sigma} G(\Sigma) \triangleq \sum_{i,j=1}^{N} -\frac{1}{2}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T \Sigma^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) + \lambda \ln|\Sigma| + \frac{\gamma}{2}||\Sigma - \Sigma_0||_F^2 \tag{17}$$

where λ and γ are two positive parameters used to balance the three terms.

3. Related work

Traditional Euclidean distance between vectors take each component as equally important. When the components are correlated, the traditional Euclidean distance cannot well characterize the difference between two vectors. For this reason, Mahalanobis distance [19] is developed for measuring the distance (or difference) between two random vectors when there probably some correlations among different components. Let x and y be two random vectors of the same distribution and Σ be the covariance matrix between the two vectors. The Manalanobis distance between x and y is defined by

$$d(x, y) = (x - y)^T \Sigma^{-1}(x - y) \tag{18}$$

where Σ is the covariance matrix between X and Y, and Σ^{-1} denotes the inverse matrix of Σ.

Since the Manalanobis distance can counteract the influence of correlations among different components of vectors, it has been applied to many metric learning based classifications [12-17]. The proposed model is very similar to those metric learning models that use Mahalanobis distance. However,

there are significant differences between the proposed model and existing Mahalanobis distance based metric learning models.

1. Almost all of the existing work using Mahalanobis distance measure the distance between two instances, while our model measures the difference between two distributions (That means the difference between two classes).
2. Since the Mahalanobis-distance based models measure distances among instances, it needs to calculate distances between every pair of instances (images) inside a class for the within-class distance, while the proposed model, which is based on classes (distributions), not instances, only calculates an average within-class distance inside a class using the determinant of the common covariance matrix. Therefore, the computational cost is significantly reduced.
3. The matrix Σ in the proposed model takes the role of the common covariance matrix among the features of all classes. Therefore, the matrix in the developed model has a richer statistical meaning. With the new statistical meaning, some constraints can be added to the model using known classes (the third term in the proposed model), which can guarantee the numerical solution not to deviate too much from the optimal solution.
4. One drawback of Mahalanobis-distance based models is that it enlarges the effect of particular samples such as outliers. Our model is based on distributions, not instances, and therefore can reduce the negative influence of particular samples in the learning stage. If the outliers are excluded during the training stage, it is possible to improve the performance of the model implementation.

4. Implementation

In this section, we discussed data preparation, initialization, parameter selection, numerical solution, and classification. We directly iterate on the covariance matrix and proved that when the parameters are selected properly, the iterated matrices can be guaranteed to be always positive definite.

4.1. Data preparation

Suppose all available data have already been preprocessed and the selected features are already calculated. Before the numerical solution is obtained, the mean feature vectors for each class need to be calculated. During this stage, control charts for normal distributions can be used to find and drop outliers so that the mean feature vectors can better stand for the

corresponding distributions and therefore lead the learning process more robust. All the covariance matrices Σ_i are also required to be calculated before the iterations are performed.

4.2. Numerical solution

Model (17) contains both Σ and Σ^{-1}, which will cause trouble in numerical implementation. We therefore convert Model (17) into the following equivalent optimization problem.

$$\begin{aligned}\min_{\Sigma^{-1}} G(\Sigma^{-1}) \triangleq & \sum_{i,j=1}^{N} -\frac{1}{2}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T \Sigma^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) \\ & - \lambda \ln|\Sigma^{-1}| + \frac{\gamma}{2}||\Sigma^{-1} - \Sigma_0^{-1}||_F^2\end{aligned} \tag{19}$$

Using M to replace Σ^{-1} leads to the following optimization problem.

$$\min_{M} G(M) \triangleq \sum_{i,j=1}^{N} -\frac{1}{2}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T M\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) - \lambda \ln|M| + \frac{\gamma}{2}||M - \Sigma_0^{-1}||_F^2 \tag{20}$$

For most applications, the number of images for each class is usually smaller than the number of features, leading to Σ_0 to be not invertible. That means Σ_0^{-1} in (20) may not exist. To overcome this problem, we use $\Sigma_0 + \epsilon \mathrm{I}$ to replace Σ_0 with small positive ϵ so that $\Sigma_0 + \epsilon \mathrm{I}$ is invertible. By differentiation, the Euler-Lagrange equation of M is

$$\begin{aligned}\sum_{i,j=1}^{N} -\left(\mu_i - \mu_j\right)\left(\mu_i - \mu_j\right)^T - & \lambda\left(2M^{-1} - diag(M^{-1})\right) \\ & +\gamma[2(M - \Sigma_0^{-1}) - diag(M - \Sigma_0^{-1})]\end{aligned} \tag{21}$$

Let

$$B = \sum_{i,j=1}^{N} \left(\mu_i - \mu_j\right)\left(\mu_i - \mu_j\right)^T. \tag{22}$$

The iterations for M can then be written as

$$\begin{aligned}M^{(t+1)} = & M^{(t)} + S \cdot \left\{B + \lambda\left[2\left(M^{(t)}\right)^{-1} - diag\left(\left(M^{(t)}\right)^{-1}\right)\right]\right\} \\ & -\gamma S \cdot \left[2\left(M^{(t)} - \Sigma_0^{-1}\right) - diag\left(M^{(t)} - \Sigma_0^{-1}\right)\right]\end{aligned} \tag{23}$$

where t stands for the tth iteration and S is the step size matrix of positive entries. When S is a scalar matrix, all entries in the matrix M evolve at the same step size. If the matrix S is taken with different entries, the evolution of different entries in M take different step sizes. In this case, $S \cdot ()$ stands for matrix dot product.

4.3. Symmetric positive-definite property

One possibility that may hinder the algorithm to work properly is the positive-definite promise of the covariance matrix Σ. For example, even if the initial value of the matrix $\Sigma^{(0)}$ is symmetric and positive-definite, the successive $\Sigma^{(t)}$ $(t = 1,2,3,\ldots)$ may not be positive definite. The following theorem provides a condition that guarantees the iterations always preserve the positive-definite property for our general Model (20) or Model (17).

Theorem 3. For Model (20) or Model (17), suppose $M^{(t)}$ is symmetric and positive-definite. If the step size matrix S is taken in such a way that all non-diagonal items are constant while the diagonal is twice of non-diagonal items, i.e.,

$$S = \frac{1}{2}s1_{n\times n} + \frac{1}{2}sI_n = \begin{bmatrix} s & \frac{1}{2}s & \cdots & \frac{1}{2}s \\ \frac{1}{2}s & s & \cdots & \frac{1}{2}s \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{2}s & \frac{1}{2}s & \cdots & s \end{bmatrix} \tag{24}$$

then the successive matrix $M^{(t+1)}$ in the iterations (23) is also symmetric and positive-definite providing one of the following conditions is satisfied:

1. $$s\gamma \le 1 \tag{25}$$
2. $$s\gamma > 1 \ \ and \ \ \sqrt{\frac{s\lambda}{s\gamma - 1}} > \max_i\{x_i\} \tag{26}$$

where x_i $(i = 1,2,\ldots,N)$ are the eigenvalues of M.

Proof. It is easy to see that $M^{(t+1)}$ is symmetric since every term in the iterations is symmetric.

Note that when the step size is taken as the given matrix, the iterations (23) can be written as

$$M^{(t+1)} = S \cdot B + s\gamma\Sigma_0^{-1} + (1 - s\gamma)M^{(t)} + s\lambda\left(M^{(t)}\right)^{-1} \tag{27}$$

where $s > 0$ is a scalar. It is easy to see that $S \cdot B + s\gamma\Sigma_0^{-1}$ is positive semi-definite for $s > 0$. Therefore, it is enough to show that $(1 - s\gamma)M^{(t)} + s\lambda\left(M^{(t)}\right)^{-1}$ is positive definite.

Let $M^{(t)} = P\Lambda P^{-1}$ be the orthogonal decomposition for $M^{(t)}$. Then $\left(M^{(t)}\right)^{-1} = P\Lambda^{-1}P^{-1}$. Then we have

$$\begin{aligned} & (1-s\gamma)M^{(t)} + s\lambda\left(M^{(t)}\right)^{-1} \\ = \; & (1-s\gamma)P\Lambda P^{-1} + s\lambda P\Lambda^{-1}P^{-1} \\ = \; & P(s\lambda\Lambda^{-1} - (s\gamma-1)\Lambda)P^{-1} \\ = \; & P[diag(s\lambda x_i^{-1} - (s\gamma-1)x_i)]P^{-1} \end{aligned} \tag{28}$$

where x_i are eigenvalues of $M^{(t)}$ ($i = 1,2,\ldots,N$) which must be positive. So, the eigenvalues of $M^{(t+1)}$ are

$$s\lambda x_i^{-1} - (s\gamma-1)x_i, \;\; i = 1,2,\ldots,N. \tag{29}$$

- Case 1: $s\gamma \leq 1$. In this case, $s\lambda x_i^{-1} - (s\gamma-1)x_i \geq s\lambda x_i^{-1} > 0$. That means all eigenvalues of $M^{(t+1)}$ are positive.
- Case 2: If $s\gamma > 1$ and $\sqrt{\frac{s\lambda}{s\gamma-1}} > \max_i\{x_i\}$, then we have

$$\begin{aligned} & \frac{s\lambda}{s\gamma-1} > x_i^2 \;\; for \;\; all \;\; i \\ & So, \;\; s\lambda > (s\gamma-1)x_i^2 \\ & So, \;\; s\lambda\frac{1}{x_i} - (s\gamma-1)x_i > 0 \end{aligned} \tag{30}$$

That means all eigenvalues of $M^{(t+1)}$ are also positive and so the matrix $M^{(t+1)}$ is also positive definite.

Therefore, in either case, $M^{(t+1)}$ is always positive definite.

The theorem is very important for the implementation. The reason is that, if we could not guarantee the resulted matrix to be still positive definite after iterations, the algorithm will be meaningless. The theorem gives us a hint that during the iterations, when the step size is taken in such a way as in Eq. (24), the parameter γ should be taken in such a way that $s\gamma \leq 1$. When the parameters are chosen in such a way, the resulted matrix after iterations is guaranteed to be still positive definite and can still be used to form a distance for distributions as in the first two terms in the model.

4.4. Initialization and parameter selection

Since the matrix Σ is symmetric positive-definite, the initial value of M in iterations (23) must be also symmetric positive-definite. One suggestion for the initial value of (23) is $(\Sigma_0 + \epsilon I_n)^{-1}$ where ϵ is taken in such a way that $\Sigma_0 + \epsilon I_n$ is invertible and $\epsilon > 0$ is small enough. Also based on Theorem 3, s is recommended to chosen in such a way that $s\gamma \leq 1$.

4.5. Algorithm

The overall algorithm for learning the matrix M is listed below.

Algorithm: Learning matrix M

- Input the number of classes, N, and each feature vector x_{ij};
- Input $s > 0$ for step size matrix based on (25) or (26);
- Input error controller $\delta > 0$, number of classes N, and maximum iterations Max > 0;
- Assign Σ_0: $\Sigma_0 \leftarrow \left(\sum_{i=1}^{N} \Sigma_i\right)/N$;
- Pick $\epsilon > 0$ such that $\Sigma_0 + \epsilon I$ is invertible;
- Initialize $M^{(0)}$: $M^{(0)} \leftarrow (\Sigma_0 + \epsilon I)^{-1}$ and $t \leftarrow 0$;
- Compute $M^{(1)}$ based on (23);
- While $||M^{(t+1)} - M^{(t)}||_F > \delta$ and $t < Max$ do
  - Update $M^{(t)}$ based on (23);
  - $t \leftarrow t + 1$.
- Output M.

4.6. Computational cost analysis

Suppose there are N classes and each class has K images. Then the feature-based methods such as relative distance methods need to compare

$$NC_K^2 + C_N^2 K^2 = \frac{(KN)^2 - KN}{2} \tag{31}$$

pairs during the training stage, while our distribution-based method only need to compare $C_N^2 = \frac{N^2 - N}{2}$ pairs, where C_N^2 is the number of combinations selecting 2 objects from N objects and C_K^2 is the number of combinations selecting 2 objects from K objects. Therefore, the proposed method is much faster in the training stage than traditional methods.

5. Application on Person Re-identification

Person re-identification (re-ID) aims at matching people across multiple non-overlapping camera networks. It manages to re-identify a target in one camera when he/she disappears from another. Person re-identification is one of the most important tasks in surveillance systems. Person re-identification is extremely challenging because of the variation of lighting, angle of imaging, low resolution and so on. Person re-identification is very useful for discovering the high-level semantic including criminal tracking, behavioral identification, mass disturbances prediction and etc. In this section, the proposed method of image classification is applied to person re-identification to verify the effectiveness of the method.

5.1. Introduction to person re-identification

Current person re-identification methods can be divided into two categories: either concentrated on hand-crafted system or based on deep learning techniques [5-7, 20-21]. Most recently published deep learning methods are video-based person re-identification [22-24]. Overall, deep learning techniques have higher matching rates. But person re-identification research using hand-crafted systems is still very active for two reasons: First, the computational cost as well as the cost of hardware for deep learning based techniques is very high; Secondly, some techniques used in hand-crafted systems can be imbedded/integrated into deep learning based techniques, such as [25-27]. The role of metric learning in deep learning based methods has been replaced by the loss function designed to guide the feature representation learning [28-31]. Hand-crafted systems can also be divided into two categories: (1) Feature-concentrated methods, such as [8-11]; and (2) Metric-learning based methods, such as [10], [14], [16-18], [32]. Just list some here.

5.2. Experimental Results

Four popular data sets are selected for the experiments: 3DPeS [34], CAVIAR4REID [35], Market-1501 [36], and DukeMTMC-reID [8] that are compared with most popular methods in person re-identification. In the experiment, two different ways of feature representations are used to make the comparison: ELF [44]) and LOMO [8]. For the ELF representation, we equally partitioned each image into 6 horizontal stripes, and RGB, HSV, YCbCr, Lab, YIQ and 16 Gabor texture features were extracted for each stripe. Except for the Gabor texture features, all other features are color features. For each feature channel, a 16D histogram was extracted and then normalized by L1-norm. All histograms were concatenated together to form a single 2688 dimensional vector. And use PCA to compress them to the low dimension with 95 percent energy to be preserved. For the LOMO representation, we extract the LOMO descriptors for each image, to represent the human appearance. The LOMO extractor has shown impressive robustness against viewpoint changes and illumination variations by concatenating the maximal pattern of joint HSV histogram and SILTP descriptor.

For each data set, similar parameters are chosen to apply in the model. For the data set 3DPeS, the step size and parameters are $s = 10^{-7}$, $\gamma = 10^{5}$, and $\lambda = 10^{4}$; For the data set CAVIAR4REID, Market-1501, and

CUHK01, the step size and parameters are $s = 10^{-10}, \gamma = 10^{9}, \lambda = 10^{4}$; For the data set DukeMTMC-reID, the step size and parameters are $s = 10^{-8}, \gamma = 10^{7}, \lambda = 10^{2}$. The experiment is performed in Intel(R) Core(TM) i7-6700K CPU @4.00GHz. RAM: 32.0GB. platform: MATLAB R2016a. The performance of all the methods is evaluated using tables that is similar to cumulative matching characteristic (CMC), which is a standard measurement for Re-ID. The CMC tables present each probability of finding the correct matching over the top r classes in the gallery image ranking, with r varying from 1 to 30. Our results are located in the last two rows of each table. Best performance are highlighted with bold numbers.
We first use ELF feature representation to compare with LMNN[37], PRDC [38], and KISSME [39] at Rank 1, Rank 10, Rank 20 and Rank 30, respectively, on 3DPeS and CAVIAR4REID. The results are shown in Table-1 and Table-2, respectively. Our method outperforms all the three methods.

Table 1: Comparison of top ranked matching rates (%) on 3DPeS

Method	Rank 1	Rank 10	Rank 20	Rank 30
LMNN (ELF)	8.4	26.7	50.9	57.3
PRDC (ELF)	17.3	50	61.7	73.6
KISSME (ELF)	7.5	29	46.3	58.7
OURs (ELF)	**35**	**68.3**	**78.3**	**83.3**

Table 2: Comparison of top ranked matching rates (%) on CAVIAR4REID

Method	Rank 1	Rank 10	Rank 20	Rank 30
LMNN (ELF)	14.2	56.6	75.6	82.7
PRDC (ELF)	23.8	74.6	86.7	95.4
KISSME (ELF)	16.6	58.5	77.5	86.6
OURs (ELF)	**55**	**90**	**97.5**	**100**

On DukeMTMC-reID, we compared our method with LMNN [37], PRDC [38], KISSME [39], XQDA [8], MMFA [42], and PAUL [43] at Rank 1, Rank 10, Rank 20, and Rank 30, respectively. Except for PAUL method, our method is better than other methods overall. However, PAUL uses a neural network to train for selecting optimal features, while our model only used manually selected features. So the cost of our model is much smaller.

Table 3: Comparison of top ranked matching rates (%) on DukeMTMC-reID

Method	Rank 1	Rank 10	Rank 20	Rank 30
XQDA (LOMO)	30.7	56.6	76.6	83.3
LMNN (ELF)	28.6	54.8	70	80.3
PRDC (ELF)	22	46.3	66.6	78.8
KISSME (ELF)	25.1	36.5	73.3	80.5
MMFA	45.3	-	-	-
PAUL	**72.0**	**86.0**	-	-
OURs (ELF)	45	83.3	**91.6**	**95**
OURs (LOMO)	60	85	85	88.3

On Market-1501, we compared our method with most recent methods: XQDA [8], CAMEL [40], MAR [41], MMFA [42], and PAUL [43] for the matching rate. Our results are a little better than other methods. Considering that some of them used deep learning with heavy cost, our method is therefore very competitive on Market-1501.

Table 4: Comparison of top ranked matching rates (%) on Market-1501

Method	CAMEL	MAR	PAUL	MMFA	XQDA	OURs
Matching Rate	54.5	67.7	68.5	56.7	68.5	**71.6**

6. Conclusion

In this paper, we developed a novel image classification method based on distributions. The developed model directly works on distributions, not on specific images in the training stage. It largely reduces the computational cost. An interesting thing about the model development is that, it starts from the K-L divergence, which is not a metric actually, but ends at a distance based model. Although the final model is still about metric learning, the original idea in this paper is only based on K-L divergence. The future work contains two aspects. (1) Distinguish outliers in the pre-processing stage and then exclude all outliers before metric learning. In this way, the matching rate can be improved. (2) Consider to use other information

metrics, instead of the K-L divergence, and extend the hypothesis to any other kind of distributions so that the model is more applicable.

Acknowledgments

We thank Dr. Hongchao Zhang (Louisiana State University), Dr. Weihong Guo (Case Western Reserve University), and Dr. Ouyang Yuyuan (Clemson University) for reviewing the paper. Their hard work is highly appreciated.

References

[1] Cover,T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions in Information Theory*, Vol. 13, No. 1, 21-27.

[2] Weinberger, K. and Saul, L. (2009). Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research*, Vol. 10, 207-244.

[3] Affonso, C. et. al. (2017). Deep learning for biological image classification. *Expert Systems With Applications*, Vol. 85, 114-122.

[4] Wang, J. and Perez, L. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. arXiv:1712.04621v1 [cs.CV], 13 Dec. 2017.

[5] Rao, Y.M., Lu, J.W. and Zhou, J. (2019). Learning Discriminative aggregation network for video-based face recognition and person re-identification. *International Journal of Computer Vision*, Vol. 127, 701-718. [doi: 10.1007/s11263-018-1135-x]

[6] Yao, H. et al. (2019). Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, DOI: 10.1109/tip.2019.2891888.

[7] Zhou, S. et al. (2019). Discriminative feature learning with consistent attention regularization for person re-identification. *ICCV*, 8040-8049.

[8] Liao, S. et al. (2015). Person re-identification by local maximal occurrence representation and metric learning. *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2197-2206.

[9] Varior, R.R., Wang, G., Lu, J. and Liu, T. (2016). Learning invariant color features for person re-identification. *IEEE Trans. on Image Processing*, Vol. 25, No. 7,3395-3410.

[10] WANG, H.Y. et al. (2018). Person re-identification by Semi-supervised dictionary rectification learning with retraining module. *Journal of Electronic Imaging*, Vol. 27, No. 4, 043043-1 - 043043-9. [doi:10.1117/12.2309840]

[11] Chen, G.Y., Lu, J.W., Yang, M. and Zhou, J. (2019). Spatial-Temporal Attention-Aware Learning for Video-Based Person Re-Identification. *IEEE Trans. on Image Processing*, Vol. 28, No. 9, 4192-4205. [doi: 10.1109/TIP.2019.2908062]

[12] Xiang, X., Nie, F. and Zhang, C. (2008). Learning a Mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, Vol. 41, 3600–3612.

[13] Gallego, G., Cuevas, C., Mohedano, R. and Garca, N. (2013). On the Mahalanobis Distance Classification Criterion for Multidimensional Normal Distributions. *IEEE Transactions on Signal Processing*, Vol. 61, No. 17, September 1, 2013.

[14] Zheng, W-S., Gong , S. and Xiang, T. (2016). Towards Open-World Person Re-Identification by One-Shot Group-based Verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 3, 591-606.

[15] Ding, Z. and Fu, Y. (2017). Robust Transfer Metric Learning for Image Classification. *IEEE Transactions on Image Processing*, Vol. 26, No. 2, Feb. 2017,660-670.
[16] Wang, Y. et al. (2019). Dynamic curriculum learning for imbalanced data classification. *Proceedings of the IEEE International Conference on Computer Vision*, 5017 - 5026.
[17] Jia, J. et al. (2020). View-specific Subspace Learning and Re-ranking for Semi-supervised Person Re-identification. *Pattern Recognition*, Vol. 108, 2020: 107568.
[18] Feng, L. et al. (2018). Learning a Distance Metric by Balancing KL-Divergence for Imbalanced Datasets. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 49, No. 12, 2384 - 2395.
[19] Mahalanobis, P.C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, Vol. 2, No. 1, 49-55. Retrieved 2016-09-27.
[20] Zheng, L. et al. (2017). Person re-identification in the wild. *CVPR*, 1367-1376.
[21] Ni, T.G. et al. (2018). Discriminative deep transfer metric learning for cross-scenario person re-identification. *Journal of Electronic Imaging*, Vol. 27, No. 4, 043026-1 - 043026-10. [doi: 10.1117/1.JEI.27.4.043026].
[22] Zhang, W. et al. (2019). Multi-scale Spatial-temporal Attention Model for Person Re-identification in Videos. *IEEE Transactions on Image Processing*, Vol. 29, 3365 - 3373. DOI: 10.1109/TIP.2019.2959653.
[23] Bai, X. et al. (2020). Deep-person: Learning discriminative deep features for person re-identification. *Pattern Recognition*, Vol. 98, 2020: 107036.
[24] Hadjkacem, B. et al. (2020). A novel Gait-Appearance-based Multi-Scale Video Covariance Approach for pedestrian (re)-identification. *Engineering Applications of Artificial Intelligence*, Vol. 91, 2020: 103566.
[25] Zheng, Z. et al. (2016). A discriminatively learned CNN embedding for person re-identification. arXiv:1611.05666.
[26] Wojke, N. and Bewley, A. (2018). Deep cosine metric learning for person re-identification. *WACV*, 748-756.
[27] Ye, M. et al. (2019). Unsupervised embedding learning via invariant and spreading instance feature. *CVPR*, 6210-6219.
[28] Sun, Y. et al. (2017). Svdnet for pedestrian retrieval. *ICCV*, 3800-3808.
[29] Deng, W. et al. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. *CVPR*, 994 - 1003.
[30] Tian, M. et al. (2018). Eliminating background-bias for robust person re-identification. *CVPR*, 5794-5803.
[31] Zheng, M. et al. (2019). Re-identification with consistent attentive siamese networks. *CVPR*, 5735-5744.
[32] McLaughlin, N. et al. (2017). Person reidentification using deep convnets with multitask learning. *IEEE Trans. Circuits Syst. Video Technol*, Vol. 27, No. 3, 525 - 539.
[33] Wang, J. et al. (2017). DeepList: learning deep features with adaptive listwise constraint for person reidentification. *IEEE Trans. Circuits Syst. Video Technol*, Vol. 27, No. 3, 513 - 524.
[34] Baltieri, D. et al. (2011). 3DPeS: 3D people data set for surveillance and forensics. *Joint ACM Workshop on Human Gesture and Behavior Understanding*, New York: ACM, 59-64.
[35] Cheng, D. et al. (2011). Custom Pictorial Structures for Re-Identification. *Proceedings of British Machine Vision Conference*, 1-11.
[36] Zheng, L. et al. (2016). Scalable Person Re-identification: A Benchmark. *Proceedings of the 14th IEEE International Conference on Computer Vision*, 1116-1124.

[37] Dikmen, M. et al. (2010). Pedestrian recognition with a learned metric. *Asian Conference in Computer Vision*, 501-512.

[38] Zheng, W-S., Gong, S. and Xiang, T. (2013). Re-identification by Relative Distance Comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 3, 653-668.

[39] Kostinger, M. et al. (2012). Large scale metric learning from equivalence constraints. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, DOI: 10.1109/CVPR.2012.6247939.

[40] Yu, H., Wu, A. and Zheng,W. (2017). Cross-view asymmetric metric learning for unsupervised person re-identification. *Proceedings of the IEEE International Conference on Computer Vision*, 994-1002, doi:10.1109/iccv.2017.113.

[41] Yu, H.-X., Zheng, W.-S., Wu, A., Guo, X., Gong, S. and Lai, J.-H. (2019). Unsupervised Person Re-identification by Soft Multilabel Learning. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2148-2157, doi:10.1109/cvpr.2019.00225.

[42] Lin, S., Li, H., Li, C.T. and Kot,A.C. (2018). Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *Proceedings of the British Machine Vision Conference*.

[43] Yang, Q., Yu, H.-X., Wu, A. and Zheng, W.-S. (2019). Patch-based discriminative feature learning for unsupervised person re-identification. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3633-3642, doi:10.1109/ cvpr.2019.00375.

[44] Gray, D. and Tao, H. (2008). Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. *European Conference on Computer Vision*, 262-275.