生物特征识别算法中对类内差距的评价*

王永鑫+，蔡莲红，贾珈，马磊

(清华大学 计算机科学与技术系，北京市 100084)

摘 要: 生物特征识别作为一种替代传统身份识别方式的技术，得到了迅速的发展，相应的评测技术也备受关注。由于传感器不同、个体使用的差异，引起类内距离加大，从而使生物特征识别正确率的下降。本文分析了类内距离对识别的影响，给出了对类内距离和类间距离之间关系的评价算法，并对在一个指纹数据库上进行了实际测试，验证了本文提出的评价算法的合理性。

关键词: 生物特征识别,评测

1. 引言

生物特征识别是一种利用人体固有生理特征和行为特征对人的身份进行辨认与确认的技术，与传统的身份验证方式相比，其特征是具有稳定性、防伪性和唯一性。因而，生物特征识别的研究近年来得到了很大的发展。但是，它的非普遍适用性、非活体蒙蔽限制了生物识别应用的范围；特别是由于传感器不同、采集设备的个体使用差异，使得一个人的生物特征，两次采集所得到的结果也可能不同。这样身份确认系统就不一定可以接受一个正确的身份声明。这是由生物特征的本身的类内差距所决定的。因此，生物特征的识别算法与识别系统的性能评价问题成为生物识别研究中非常重要的一个组成部分。

对生物特征识别算法与识别系统进行性能评价，是随着生物特征的识别系统一起发展起来的。现在，对某些生物特征，已经有了比较成熟的生物特征数据库，并开展了相关的识别算法竞赛，如世界指纹确认竞赛（FVC, Fingerprint Verification Competition）[2]等。国内也开展了脸像识别、指纹识别评测等的评测竞赛[4]。在这些评测实践中，形成一系列的评测理论与方法。

对生物特征识别算法进行评测，在[1]中被分为三种方式：技术评测、场景评测与操作评测。在前面提到的生物特征识别竞赛中，较多采用技术评测。这种评测方式主要对识别算法进行评测。用作生物特征识别算法评测指标的有错误匹配率(FMR, false match rate)、错误不匹配率(FNMR, false non-match rate)、接受操作曲线(ROC, receive operating curve)等[2]。它们收集算法在特定数据库上的得分，并考察特定阈值下匹配错误的数量。

匹配错误包括错误匹配——将来自两个不同个体的生物特征认为是匹配的，与错误非匹配——将来自同一个体的两个生物特征认为不匹配。这两种错误，在对实际系统进行评测时也常被称为错误接受(FA, false accept)与错误拒绝(FR, false reject)。

在生物特征识别过程中，这两种错误是难以避免的。这是由于生物特征存在类内差距。

所谓类内差距是指，由于个体的使用差异而引起的同一个体的生物特征在两次采集

* * 本工作由国家自然科学基金（60433030,60418012）资助.

* 王永鑫 Email: wangyongxin@mails.tsinghua.edu.cn

中所得到结果之间的差距。而不同个体的生物特征之间的差距称为类间差距。由于类内差距的存在，生物特征识别算法就必须要容忍一定的类内差距，才能保证不将正确的用户拒绝。对算法来说，可以用样本的类内距离与类间距离来表示这种差异。但是类间距离与类内距离相比，相差并不十分显著，有时还是互相交叉的。这就使生物特征识别算法难以给出完全正确的结果。

已有的对生物特征识别算法的评测方式，多数是根据错误率对算法进行评测。这样做所得到的评测指标的意义很明确，但有时不能为指导算法的改进发挥很大的作用。由于算法可能在某一些特定形式的数据上比较容易出现问题，所以如果可以找到这些数据并分析它们的特点，就可以更好地指导对算法进行进一步的改进。

本文分析了类内距离对识别的影响，给出了对类内距离和类间距离进行评价的算法，并对一个指纹数据库进行了实际测试，验证了本文提出的评价算法的合理性。

2. 生物特征识别系统的基本工作流程

生物特征识别系统在工作过程中，通常会有注册与识别两个基本过程。

在注册过程中，识别系统要经过数据采集、数据预处理、特征提取、模型建立等步骤。其中数据采集指由物理设备得到相关生物特征的原始数据；数据预处理指对原始数据进行预处理，去除在采集过程中由采集设备引入的一些噪声等；特征提取指从原始数据中提取一些可以表示这个数据的特征，这步工作的实际作用是数据的降维，减少数据量；模型建立指在提取到的特征的基础上，建立一个表征这一个原始数据的模型。在实际的算法与系统中，可能会有某些步骤缺失。

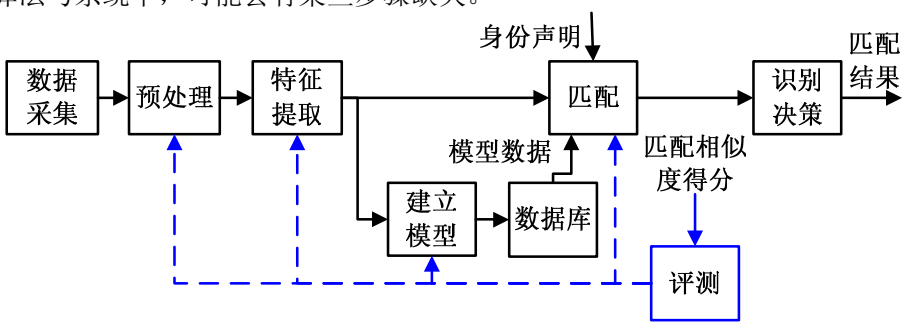

图 1 生物识别系统框图

在识别过程中，识别系统要经过数据采集、数据预处理、特征提取、匹配、识别决策等步骤。相应的步骤与注册过程是一致的。匹配步骤是指，算法根据实际的需求，将从新的数据中得到的特征与数据库的某个或某些模型进行比较，并给出匹配相似度的得分；识别决策过程是指算法根据得到的匹配得分与实际问题的要求，对实际的问题给出一个回答。例如在确认系统中，匹配步骤会把新数据的特征信息与其声明的身份的模型信息进行比较，并给出一个匹配相似度得分；然后在识别决策过程中，算法通常将得分与一个阈值进行比较，并根据比较的结果给出匹配或者是不匹配的结论。

算法给出的相似度得分，是对样本的相似性的一个度量。它与算法所看到的样本之间的距离有负相关关系。本文就以相似度得分为基础，给出一些衡量算法中类内距离与类间距离的评价方法，并通过实验验证了方法的有效性。

3. 评价算法的类内差距

算法出现误匹配的原因是由于算法不得不容忍类内差距，而算法在某些情况下难以

区分类内差距与类间差距。这样，对算法给出的类内的差距与类间差距的估计就成为一个重要的问题。

d'(d prime)[3]就是一个从整体上对类内差距与类间差距进行估计的指标。

在第 2 节中提到过，算法中匹配过程给出的相似度得分与样本之间的距离是呈负相关关系的。因此，可以用相似度得分来代表样本之间的距离。得分越高，距离越近。

将两个同一身份的样本之间的匹配称作真实匹配，匹配所得到的相似度得分称为真实匹配得分(gms, genuine match score)；将不同身份的两个样本之间的匹配称作虚假匹配，匹配所得到的相似度得分称为真实匹配得分(ims, imposter match score)，那么 d'可以记为(s 表示得分)：

$$d' = \frac{\overline{s_{gms}} - \overline{s_{ims}}}{\sqrt{[\sigma^2(s_{gms}) + \sigma^2(s_{ims})]/2}} \qquad \text{公式（1）}$$

其中，$\overline{s_{gms}}$ 是所以真实匹配得分的平均值，$\sigma^2(s_{gms})$ 是所以真实匹配得分的方差。$\overline{s_{ims}}$ 与 $\sigma^2(s_{ims})$ 的含义类似。

可以看到，这一指标反映类内的匹配得分与类间的匹配得分之间的分布差距。由于匹配得分反映了样本之间的距离，所以该指标也反映了该算法意义下类内距离与类间距离的一种统计关系。显然 d'越大，说明算法可将类内差距与类间差距区分地分得越好，也就是算法的效果将越好。但是，它不能具体指出算法在哪些数据上表现得不好。

于是，我们可以使用另一种评测方式。即：

$$d_{ij} = \frac{\overline{s_i} - \overline{s_{ij}}}{\sqrt{[\sigma^2(s_i) + \sigma^2(s_{ij})]/2}} \qquad (i \neq j) \qquad \text{公式（2）}$$

其中，s_i为身份标识同为i的不同的样本之间的匹配结果，$\overline{s_i}$ 为其平均值，$\sigma^2(s_i)$ 为其方差；而s_{ij}为身份标识分别为i, j的不同身份的样本之间的匹配结果，$\overline{s_{ij}}$ 为其平均值，$\sigma^2(s_{ij})$ 为其方差。

$\overline{s_i}$ 称为类内平均相似度得分，d_{ij}称为类间相对距离。又令：

$$d_i = \frac{1}{N-1}\sum_{j \neq i} d_{ij} \qquad \text{公式（3）}$$

称为类间相对平均距离。其中 N 为数据库中的拥有不同身份的个体总数。

类内平均相似度得分越高，也就是类内距离越小的时候，算法的在该类上得到正确结果的可能性就越高。同样，类间相对平均距离越大的时候，算法就越容易将该类的样本从整个数据库中分离出来，算法的结果也将更好。

通过这样一种方式，就可以计算两个不同的小类，也就是在生物特征识别中两个不同身份之间的差距。由于该方法得到的结果是对每一对身份之间的进行匹配的得到的，因而可以更仔细地看到在该算法下哪些数据发生了混淆。通过对这些产生混淆的数据特点的分析，就可以进一步更有针对性地对算法进行改进。

这个方法在进行评测时，对某些算法还时常会得到负的d_{ij}。这说明，有时对这两个不同的类，算法所看到的类间差距要比类内差距更小。这样一来，无论算法的匹配阈值设为多少，在涉及这两类的匹配都会产生很多错误。这也将成为算法设计在下一步工作时需要重点解决的问题。

4. 评测实例

以下是某指纹识别算法在BVC2004[4]的数据库 2B中的测试结果。

该数训据库中共包括了 20 个不同手指的指纹，每个手指有 10 个指纹图片，共有 200 个指纹图片。

算法的等错误率点 EER 为 0.313304，d'值为 0.900528。

下面考察算法在某一类上平均错误率。该错误率以有该类参与的所有匹配中，错误拒绝与错误接受率的平均来表示： $AvgErrRate_i = (FNMR_i + FMR_i)/2$ 。在计算错误时，匹配阈值取算法取得等错误率点时的阈值。

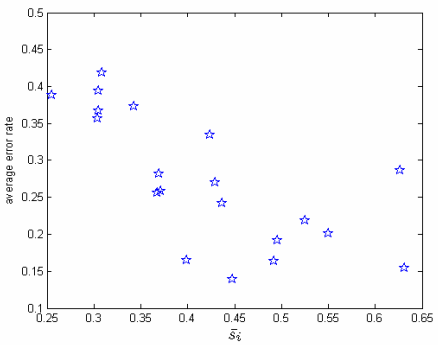

图 2 平均错误率与类内平均相似度得分的关系

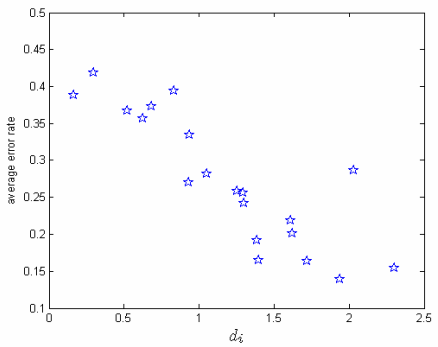

图 3 平均错误率与类间相对平均距离的关系

从图 2 中可以明显看到，当类内平均相似度得分比较小的时候，匹配错误率变得到很高。当算法的类内平均相似度得分比较大，也就是类内距离比较小的时候，算法在该类上的平均错误率就较小。

从图 3 中可以明显看到，平均错误率随着相对平均距离的增大而减小。

因而在算法的进一步改进的过程中，要重点考虑类内平均相似度得分较小，以及类间相对平均距离较小的数据类，使算法得到的类内距离更小而类间相对距离更大。这样就可能使算法的性能得到进一步的提高。

5. 总结

生物特征识别算法的评测，是研究生物特征识别算法的一个重要的工具。它的主要作用就是根据算法的运行结果，给出一些有关算法的性能的评价。同时，它应该起到指导算法的进一步改进的作用。

本文中主要论述了生物特征识别算法中识别错误的产生的原因。并且，针对这些产生的原因，给出了一些评测算法的方法与指标。这些指标可能对数据采集、特征预处理以及匹配算法的改进提供思路。也期望为算法的评测，带来新的思路。

参考文献

[1] P.Jonathon Phillips, Alvin Martin, C.L.Wilson, Mark Przybock．“An Introduction to Evaluating Biometric Systems”．IEEE Computer．2000．P56-63

[2] D.Maio, D.Maltoni, R.Cappelli, J.L.Wayman, A.K.Jain．“FVC2000: Fingerprint Verification Competition Report”．http://bias.csr.unibo.it/fvc2000/

[3] Ruud M.Bolle, Nalini K.Ratha, Sharath Pankanti．“Performance Evaluation in 1:1 Biometric Engines”．Sinobiometrics 2004．Springer-Verlag Berlin Heidelberg．2004．Guangzhou, China．P27-46

[4] http://www.sinobiometrics.com/sinobiometrics%2704.htm#2

Evaluation on the Variance within Same Identity in Biometrics

WANG Yong-Xin+, CAI Lian-Hong, JIA Jia, MA Lei

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)
+ Corresponding author: Phn: +86-01-5153-8002, E-mail: wangyongxin@mails.tsinghua.edu.cn

Key words: Biometrics; Evaluation

Abstract: Biometrics-based authentication has developed rapidly these years as a possible replacement of traditional authentication methods, and the evaluation of such systems has become more and more import. The main purpose of evaluation is to tell how well a system or algorithm is working. Many different performance statistics have been found these years. They can represent the performance of biometrics algorithms or systems.

But the traditional evaluation methods can only tell the overall performance of an algorithm on a given database or limited subjects. They cannot tell the designer on which kind of data the algorithm or system is likely to fail, which means much to people working on the algorithm. In this paper, I tried to show what kind data is likely to cause a specific algorithm to fail.

In the process of biometrics authentication, algorithm gets data from capture devices. And then the raw data is preprocessed to reduce noise. The noise might be brought in within the capture device, along with the data captured. And it might also come from different capture devices, or subjects' different habits using the capture devices. After the preprocessing step, features are extracted from data. Features from different data with the same identity are then combined into a template, and the template is stored into a database. The above steps are called enrollment. In a match, features of a new subject are matched with one of the templates in database using a specific matching algorithm. The algorithm then gives a match score base on similarity.

In capture step, noise can be brought in. Noise here made data with the same identity have different appearance. And sometimes there may be some data from a different identity which seem to be more similar to a given piece of data than other data from the same identity. That is the cause why the matching algorithm would fail.

In other steps, efforts are made to reduce the impact of noise. They made data from the same identity be close. But there's still variance within the same identity, and sometimes the variance is still large enough to get the algorithm to give a wrong result.

In this paper I tried to give a way to evaluate the variances within data with the same identity. As the similarity match score given by the matching algorithm have negative correlation to the distance between different samples, I used the match scores to estimate the distance between different samples. The lower the match score, the greater the distance between the two samples. Based on the estimation of distance, I found that when an identity's data have a smaller variance, the algorithm will make fewer mistakes on that identity's data. When the variance between some identity is great, the algorithm is likely to make more mistakes with that identity's data. That shows that the algorithm cannot deal with the noise came from that identity's data effectively, and gives researchers a suggestion that how the algorithm can be improved.

I also compared the distances with the same identity with distances between data from that identity and some other identity. That is important because distances can be only meaningful when compared relatively. I assumed that the distribution of scores between each

pair of two identities is a normal distribution, and compared distances between each pair of different identities with distances within data of one of the identity in the pair. The result is called relative distance between identities. The relative distance can be calculated between each pair of different identity. Then we can get an average of all the relative distances concerning one identity, and call it average relative distance, which can be used to show how well one identity is separated with the others. I found in experiments that the greater the average relative distance is, the less likely the algorithm to make errors concerning the data of that identity.