2022 年第十二届 MathorCup 高校数学建模挑战赛题目

A 题 大规模指纹图像检索的模型与实现

在生物特征识别领域，指纹作为最具独特性与持久性的生物特征之一，被广泛应用于身份识别。

指纹识别过程分为特征提取和比对两个环节。其中特征提取环节会提取用于指纹识别的指纹特征，一般国际上最为常见的指纹特征为“细节点”特征，其可视化展示形式如图 1 中的浅蓝色小圆圈及对外伸出的浅蓝色短线段，短线段用于指示细节点处纹线方向。细节点一般采用三元存储格式：(x,y,θ)，分别表示 x 轴像素坐标、y 轴像素坐标及细节点方向。一般而言：（1）指纹图像坐标体系：左上角为坐标原点，且 x 轴方向向右，y 轴方向向下；（2）细节点表达约定：细节点 x,y 的位置采用指纹图像坐标系表达，其方向规定：零度方向为 x 轴正方向（向右），90 度方向为 y 轴负方向（向上），180 度方向为 x 轴负方向（向左），270 度方向为 y 轴正方向（向下），最大角度为 359 度。角度的最小区分单位为 1 度。

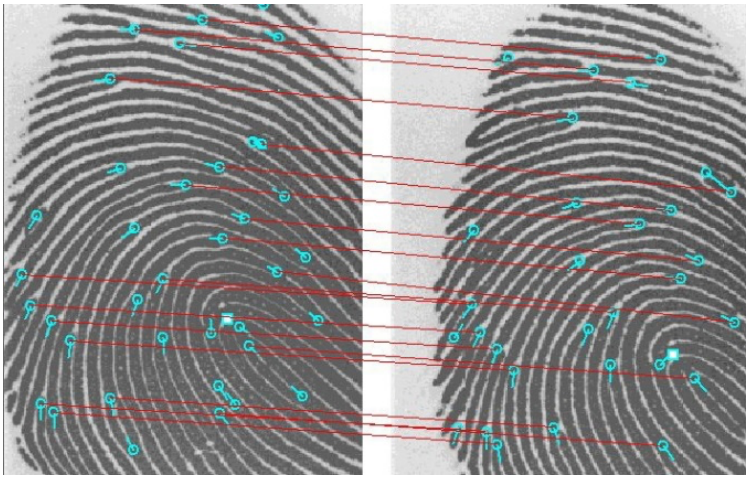

图 1 指纹识别原理

在指纹匹配环节，需要对两幅指纹图像的“同一性”进行定量评价，通常采用相似度指标。常见的两枚指纹之间的相似度评价主要依据每枚指纹图像中各个细节点之间的匹配关系。如图 1 所示，相互具有匹配关系的细节点之间用一根跨越两幅图像的红线将其互相连接，用于可视化展示。

在指纹图像匹配环节，常需要考虑如下的情况：

考虑到在采集指纹图像时，手指按压图像采集设备的角度、轻重及位置各不相同，因此两幅指纹图像需要做图像的旋转、平移后才能相互对准。由于手指皮肤较为柔软，通过按压方式采集到的指纹图像会发生一定程度的不规则弹性形变，在图 1 中会发现两幅指纹图像中，某些相互匹配的细节点在对准时，不能完全“重叠”，有一定幅度的位置及角度的偏差。这一现象也可以从“跨越两幅图像的红线并不是都平行”现象中观察到。

考虑到手指可能存在临时性蜕皮、褶皱等因素，且空气中的湿度及皮肤表面的干燥程度或粘附在皮肤上的异物等都会导致采集到的指纹图像存在纹线模糊、遮挡等现象，最终导致某些原本应该提取到的特征没有提取到，或者提取了一定数量的伪特征。在图 1 中可以观察到并不是所有的细节点都有对应的红线进行关联。

指纹识别问题中有一种一对多比对模式（one-to-many matching），是录入的查询指纹与指纹数据库中的所有已登记指纹逐一进行匹配，直至找到相似度最佳的已登记指纹或搜索完整个指纹数据库后给出无对应已登记指纹的结论。辨识模式主要应用于刑侦指纹自动识别系统、大型指纹考勤系统和门禁系统等。随着社会上指纹识别应用的普及，用于识别人员身份的指纹数据库规模也迅速上升，居民身份证指纹库容甚至达到了亿人级别。

这也会导致“逐一”遍历全库的查找方式因每次遍历时间过长而无法具有实际应用价值，必须引入图像检索技术，缩短数据库的每次遍历时间。

指纹图像的检索原理可以形象的看成：用若干大筛子快速、精准地筛除掉数据库中绝大多数与查询指纹图像明确不具有“同一”关系的图像。检索过程完成后，留下的少量图像和查询指纹具有高度相似性，需要进一步利用指纹识别算法做“逐一”识别。

一般而言，图像检索过程筛选掉的指纹图像少，则在检索过程完成后保留下来“同一”关系指纹图像的可能性大，但是整个识别过程的耗时较长；反之，检索算法筛选掉的指纹图像多，则在筛选后保留下来“同一”关系指纹图像的可能性相对变小，但是整个识别过程的耗时较短。若检索过程中筛选掉了“同一”关系的指纹，则后续“逐一”比对算法无论怎么改进都无法再找到这枚“同一”关系的指纹。因此如何设计高效精准的搜索算法是大规模指纹图像检索的关键问题。

问题 1：分析指纹图像的细节点特征(参见附件三所述三个数据文件)，给出可用于指纹快速检索的检索方法，请阐述该检索方法的原理。重点说明：（1）检索过程中避免筛除掉“同一”指纹的机制；（2）完整、清晰的图像检索模型框架及实现方法；（3）给出该检索方法的时间复杂度、空间复杂度分析。估算检索方法自身占用的内存空间规模及由于采用该检索方法，每一枚指纹图像需要承担的内存空间大小。

问题 2：针对提出的检索方法：将数据文件中提供的具有“同一”关系的全部指纹对子(TZ_同指.txt)作为查询图像在指纹数据集中进行检索方法验

证。具体要求参见附件一。

问题 3：针对 TZ_同指 200_乱序后_Data.txt 数据文件采用和问题 2 相同的方式进行检索，给出检索结果，并将结果数据压缩为.zip 格式后上传到竞赛系统的计算结果中。具体要求参见附件二。

问题 4：（1）在完成本赛题时你们可能尝试了不同类型的数学模型及技术路线，并利用问题 2 的数据验证、优选出最佳的检索方法。请介绍并评价你们所考虑过的模型及技术路线的优缺点；（2）本赛题最高的筛选量为 97%，针对 97%以上的筛选量，你们在检索精度、检索时间及内存占用等方面有什么更好的改进策略或者会尝试什么新的检索方法？

附件一 算法验证要求：

问题 2 所需数据集的构成为：TZ_异指.txt 和 TZ_同指.txt 两个文件，合计 10500 枚指纹。

给出当检索方法分别过滤掉数据集中 80%，90%，95%，97%图像后，在剩下的小规模子集中，仍含有“同一”指纹匹配对子的数量占总指纹“同一”匹配对子数量（500 对，TZ_同指.txt 的总共含有 500 对）的比例。

例如，采用 TZ_同指.txt 全部 500 个匹配对子（合计 1000 行）的数据逐一在 TZ_异指.txt 文件与 TZ_同指.txt 组成的数据集中进行检索，某检索算法按照 95%的过滤水平做检索，发现 TZ_同指.txt 文件中有 400 个匹配对子的 ID 出现在各自的检索结果中（例如用 ID 为 10200_0 的指纹检索，在检索结果中除了发现它自己（10200_0）之外，能找到 ID：10200_1），而 100 个匹配对子的 ID 没有出现在各自的检索结果中（例如用 ID 为 10200_0

的指纹检索，在检索结果中除了发现它自己（10200_0）之外，没有找到 ID：10200_1），则问题 2 的需要计算的比例结果为：在 95%的过滤水平上，穿透率为 400/500=0.8

附件二 算法计算要求：

问题 3 所需数据集的构成为：TZ_异指.txt 和 TZ_同指 200_乱序后_Data.txt 两个文件，合计 10200 枚指纹。

针对 TZ_同指 200_乱序后_Data.txt 数据文件采用和问题 2 相同的方式在数据库中进行检索，分别建立名称为“result_90”、“result_95”、“result_97”的目录，在上述三个目录中分别保存过滤掉 90%，95%，97%的指纹图像（TZ_异指.txt）后，200 枚图像（TZ_同指 200_乱序后_Data.txt）的检索结果，要求每行数据对应一个文本文件。文本文件的格式如下（ID 代表了每一指纹的识别标识）：

检索筛选后留下的指纹图像个数,ID1,ID2,ID3,…

例如，在 97%的过滤水平下，利用 TZ_同指 200_乱序后_Data.txt 文件的 ID 为 A1 的第一行数据在数据库中进行检索，取得的检索结果保存在 result_97 目录下的 A1.txt 文件中。检索结果中留下 4 枚高度相似的指纹，其 ID 编号分别是：10200_0,10201_0,A1,A4。

A1.txt 文件内保存的内容如下：

4,10200_0,10201_0,A1,A4,

附件三：数据文件格式说明

1. 共计三个文本文件：TZ_同指.txt，TZ_异指.txt，TZ_同指 200_乱序后_Data.txt

2. TZ_同指.txt 文件共计 500 个手指，每个手指重复采集 2 次，这两次重复采集的图像，相互之间形成同一匹配关系。

3. TZ_异指.txt 文件共计 10000 个手指，每个手指只采集一次，相互之间没有同一匹配关系。

4. TZ_同指.txt，TZ_异指.txt 以及 TZ_同指 200_乱序后_Data.txt 的文件格式相同。每一行代表一枚指纹，包括指纹 ID，细节点总数，每一个细节点三元信息，细节点的每一元之间、细节点之间均用逗号隔开（特别注意:因为每枚指纹提取到的细节点个数不同，所以每一行的长度不同）。具体格式如下：

例如：10200_0,3,20,30,40,40,60,34,100,110,89,

10200_1,2,23,33,43,43,63,34,

A1,2,25,44,54,66,43,29,

解释：10200_0 的指纹图像提取到了 3 个细节点，其三元组数据为：(20，30，40),(40，60，34),(100，110，89)

解释：10200_1 的指纹图像提取到了 2 个细节点，其三元组数据为：(23，33，43),(43，63，34)

解释：A1 的指纹图像提取到了 2 个细节点，其三元组数据为：(25，44，54),(66，43，29)

特别说明：（1）真实数据中，每一枚指纹图像提取到的细节点个数平均约为 20-40 个，具体参见数据文件，明显多于 10200_0、10200_1 这两个解释例子；（2）TZ_同指 200_乱序后_Data.txt 文件启用了新的指纹 ID 进行标注，如 A1。

5. 对于 TZ_同指.txt 文件而言：两个指纹 ID，若在“_”符号前的字符相同则是同一人的同一手指，“_”符号后的数字代表了第几次采集。这也意味这 2 次采集的图像具有“同一”关系。例如 10199_0 和 10199_1 是同一手指的指纹，分别是第一次采集和第二次采集。这两枚指纹之间具有“同一”匹配关系。