

武汉理工大学

(申请工学硕士学位论文)

基于 Android 的食品配料表 识别系统研究

培养单位：信息工程学院

学科专业：通信与信息系统

研究生：乔爽

指导教师：肖攸安 教授

2014 年 5 月

分类号_____

密 级_____

UDC _____

学校代码 10497

武汉理工大学

学 位 论 文

题 目 _____ 基于 Android 的食品配料表识别系统研究

英 文 _____ Food Ingredients Identification System Research

题 目 _____ Based on The Android

研究生姓名 _____ 乔 爽

指导教师 姓名 _____ 肖攸安 _____ 职称 _____ 教授 _____ 学位 _____ 博士

单位名称 _____ 信息工程学院 _____ 邮编 _____ 430070

申请学位级别 _____ 硕 士 _____ 学科专业名称 _____ 通信与信息系统

论文提交日期 _____ 2014 年 4 月 _____ 论文答辩日期 _____ 2014 年 5 月

学位授予单位 _____ 武汉理工大学 _____ 学位授予日期 _____

答辩委员会主席 _____ 评阅人 _____

2014 年 5 月

独 创 性 声 明

本人声明,所呈交的论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知,除了文中特别加以标注和致谢的地方外,论文中不包含其他人已经发表或撰写过的研究成果,也不包含为获得武汉理工大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签 名: _____ 日 期: _____

学位论文使用授权书

本人完全了解武汉理工大学有关保留、使用学位论文的规定,即学校有权保留并向国家有关部门或机构送交论文的复印件和电子版,允许论文被查阅和借阅。本人承诺所提交的学位论文(含电子学位论文)为答辩后经修改的最终定稿学位论文,并授权武汉理工大学可以将本学位论文的全部内容编入有关数据库进行检索,可以采用影印、缩印或其他复制手段保存或汇编本学位论文。同时授权经武汉理工大学认可的国家有关机构或论文数据库使用或收录本学位论文,并向社会公众提供信息服务。

(保密的论文在解密后应遵守此规定)

研究生(签名): _____ 导师(签名): _____ 日期: _____

摘要

近年来，食品安全问题越来越引起人们的重视，对于配料表详细信息的诉求也越来越高。然而，在查看配料表时经常会出现如“不溶性聚乙烯聚吡咯烷酮”等词语引起人们的不解。在过去，人们可能不去深究，但随着移动互联网的到来，人们获取信息的方式更加便捷，拿出手机即可进行搜索查询。但截止目前，还没有一个专门的系统可以进行配料表的查询，人们使用的还是通过手机浏览器链接互联网搜索引擎，输入配料词语进行查询，如果遇到难认字或字数很多，依旧会带来很大的麻烦。因此，一个能够方便快捷获取配料信息的系统是很有必要的。

本文对当前的配料表查询方式和文字识别系统进行了分析，考虑到当前人们智能手机的使用习惯，提出了基于 Android 的食品配料表识别系统。用户通过 Android 智能终端进行配料拍摄，将图片上传到服务器，经过一系列预处理及文字识别后，将获取的配料详细信息显示在 Android 客户端，极大地方便了人们关于配料信息的获取。在理论方面，对预处理（二值化、倾斜校正、字符切分、归一化、细化等）、特征提取、分类器设计等进行了详细论述；在实践方面，完成了整个系统的设计及实现，通过与当前配料查询方式、现有文字识别系统在使用方便度、识别效果、结果全面性等进行对比，得到较好的结果。

经过对本文工作的总结，主要的创新工作归纳如下：

- 1) 在拍摄控制的基础上，结合配料表特性，基于投影法，引入聚类分析获取阈值，实现词语和单字分割技术，有效地分离出配料词语或者单字，相较于传统的单字提取，提供了更丰富的模式及更准确的定位。
- 2) 鉴于配料词库的有限性，提出了基于编辑距离的智能纠错算法，来改善文字识别后的误差，有效解决了文字识别率无法达到 100% 的问题。
- 3) 提出了基于整词的配料表识别策略，相对于传统的基于单字的识别，可以提供更多的模式信息，减小维度，提供更准确的识别结果，对其他的特定细分领域具有很好的借鉴意义。
- 4) 鉴于移动终端应用习惯，提出了反馈式自学习识别策略，用户使用次数越多，得到的结果也将越准确。

关键词：光学字符识别，倾斜校正，字符切分，整词识别，配料表

Abstract

Recently, food safety has increasingly attracted people's attention. There are many ingredients on packages such as "Insoluble Polyvinylpyrrolidone" that people cannot understand. In the past, people might not dig deep about them. But with widely-used internet, people can get information faster and more conveniently nowadays. They can even use a mobile phone to search for what they want. So far there hasn't been any special system for searching ingredients. What people do now is just put an ingredient's name in a search engine and then click "enter" to get the information. In this case, difficult words or words with many letters may bring a lot of trouble. Therefore, a system for obtaining information on ingredients conveniently and quickly is of crucial importance.

After analyzing the inquiry mode and character recognition system of current ingredients, this thesis proposes a food ingredient recognition system based on Android considering people's habit of using smart phones. Users take photos of food ingredients with terminals with Android system and upload the photos to the server. The information about the ingredients will be shown on Android terminals after a series of preprocessing and character recognizing procedures, which benefits people a lot for obtaining ingredients information. In aspect of theory, this thesis discusses in detail the preprocessing, including linearization, tilt correction, character segmentation normalization, refining as well as feature extraction, categorizer design. As for practice, this thesis has finished the whole system's design and application, which compared with existing inquiry mode and characters recognition system in the aspects of convenience, recognition effect and the comprehensiveness of results has made great achievements.

The innovations of this thesis are as follows:

- 1) On the basis of the controller and the features of ingredient sheets, this thesis uses projection methods and cluster analysis to get threshold value making segmentation technology separating words into single characters possible, which effectively extract characters and words of ingredients. This technology offers more kinds of modes and more accurate location than traditional single word extraction.

2) Due to the limitation of the database of ingredients, this thesis presents the intelligent error correction algorithm based on edit distance to improve the random error of characters recognition, which effectively solve the problem that is the rate of character recognition cannot reach 100%.

3) The strategy based on the recognition of whole words rather than traditional single characters can provide more mode information and reduce dimensions, which can lead to more accurate recognition results and is instructional to other specific subdivision fields.

4) The thesis comes up with a strategy to learn from feedback according to the habit of using mobile terminals. The more users employ the system, the more accurate their results will be.

Key words: Optical Character Recognition, Tilt Correction, Character Segmentation, Recognition of Words, Ingredients

目录

摘要 I

Abstract II

目录 I

第 1 章 绪论	1
1.1 课题来源及研究背景	1
1.2 研究目的与意义	2
1.2.1 文字识别原理	2
1.2.2 汉字识别难点	4
1.2.3 配料表识别难点	5
1.2.4 论文研究意义	5
1.3 国内外发展现状	6
1.3.1 发展历史	6
1.3.2 研究现状	7
1.4 论文研究主要工作	8
第 2 章 系统方案设计	10
2.1 系统综述	10
2.2 系统方案分析	11
2.2.1 服务器设计	12
2.2.2 客户端设计	14
2.3 本章小结	15
第 3 章 预处理	16
3.1 预处理简介	16
3.2 彩色图像的灰度化	16
3.3 二值化	17
3.3.1 全局阈值法	17
3.3.2 局部阈值法	19

3.3.3 二值化算法的改进	21
3.4 倾斜检测与校正	22
3.4.1 倾斜检测	23
3.4.2 倾斜校正	24
3.5 基于投影与聚类的版面分析与字符切分	26
3.5.1 配料表版面分析介绍	26
3.5.2 版面分析常用方法	26
3.5.3 基于投影法的文档图像分析	27
3.5.4 基于聚类分析的文本切割	28
3.6 归一化	29
3.7 汉字细化	30
3.8 本章小结	31
第4章 配料识别	32
4.1 汉字特征提取	32
4.1.1 特征提取概述	32
4.1.2 基于统计的特征提方法	32
4.2 分类器原理	33
4.2.1 距离分类器介绍	35
4.2.2 识别的策略	37
4.3 基于智能纠错的分类识别策略	37
4.3.1 分类器——粗分类	38
4.3.2 分类器——细分类	39
4.3.3 基于编辑距离的智能纠错	40
4.4 基于整词识别的配料表识别策略	42
4.4.1 分类归一化	42
4.4.2 分类器重新设计	42
4.4.3 整词、分字双重识别设计	43
4.5 本章小结	44
第5章 系统实现与测试	45
5.1 系统实现	45
5.1.1 算法流程	45
5.1.2 客户端界面	46

5.1.3 操作步骤	47
5.2 测试及对比分析	48
5.2.1 功能测试对比	48
5.2.2 性能测试对比	50
5.3 本章小结	54
第 6 章 总结与展望	55
6.1 本文工作总结	55
6.2 展望	55
参考文献	57
致谢 61	
攻读学位期间获得的科研成果	62

第1章 绪论

1.1 课题来源及研究背景

基于 Android 的食品配料表识别系统的研究目的是为使用者简化查询，迅速了解配料详细信息，指导食用。项目基于 Android 客户端进行配料表拍摄，简单处理后发给服务器进行配料提取、OCR 识别、配料查询，最终将配料释义反馈给用户。

食品安全问题，涉及到广大人民群众的健康和生命安全，已经成为人们普遍关注的核心问题。当前社会、经济发展十分迅速，食品工业也取得了更大、更新的发展。然而，虽然种类越来越繁多的食品添加剂给人们的生活带来了很大改善，却也使得食品安全问题日益突出^[1]。近年来频频发生的如疯牛病、禽流感等各种危害人们健康和生命的疾病也属于食品安全问题，不仅损害了经济的发展，有的甚至导致了政治风波^[2]。随着人们对食品安全问题的关注，食品配料表也进入到人们的视野^[5-6]。然而，虽然包装上面标明了所含配料，但除了“水”“蛋白质”等常见配料，广大食用者还是不了解所食为何物^[7]。鉴于版面限制与美观问题，又不能写出详细信息，所以，一款方便的能够查询配料信息的系统就呼之欲出了。

截止到目前，还没有一款成形的系统能够满足人们的需求。大部分的应用场景还是人们拿出手机，打开浏览器，输入配料，查询，如果遇到生僻字就更加纠结。而现有的文字识别系统一个是识别率不高，另一个是没有查询功能，相比较手动输入汉字也许带来更多的麻烦。

由于目前还没有专门的配料表智能识别系统，而其核心为文字识别系统，因此以下针对文字识别系统进行讨论。

模式识别^[8]是在 20 世纪 20 年代诞生的一门学科。G.Tauschek 在 1929 年发明了能够阅读 0~9 的数字的阅读机开启了文字识别的大门。Fisher 在上个世纪的 30 年代提出了一种基于统计的分类理论，从而奠定了模式识别的基础。到后来，分别出现了句法模式识别、模糊模式识别、神经网络模型、小样本学习、支持向量机等，均受到了很大的重视。一般的模式识别系统基本构成见图 1-1。

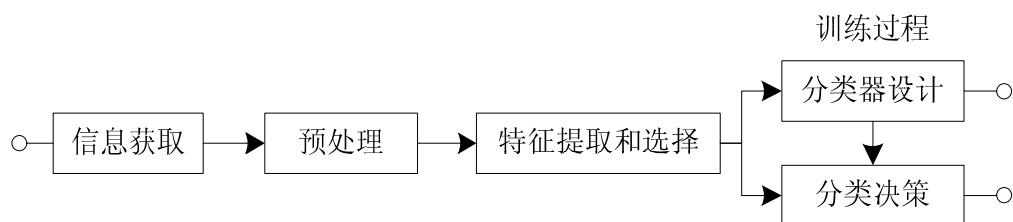


图 1-1 模式识别系统基本构成

作为模式识别的一种，文字识别尤其是汉字识别是一个难点^[9]。怎样能够方便、快捷地将文字输入到计算机一直是人机接口效率中一个很难突破的瓶颈。现在，大部分的输入主要靠人工，虽然也出现了文字识别和语音输入，但是效率和正确率还有待提高。在文字识别领域，标准文本（打印的纸质文本）录入较简单，对于复杂背景（如封面、商标等）、字体多变、字号较小的文字识别较难，还停留在实验室阶段^[10-11]。

近年来，智能手机已经成为生活必需品，基于手机拍照的识别系统也浮出水面。继有道词典推出拍照取词之后，微信的“扫一扫”更是推出了扫描二维码、扫描图书封面、扫描街景等应用，志在改变用户应用习惯^[12]。年前百度推出的百度翻译，其拍照翻译功能更是瞬间火遍全国，虽然官方公布识别率只有20%（实际更低），但依然没有减少人们的使用热情。然而，除了有道词典的翻译具有真正的实用价值外，其他的还停留在娱乐层面。专门针对与人们生活息息相关的食品配料识别还没有成型的产品或研究，因此，本文重点研究基于Android的配料表智能识别系统，旨在为人们提供方便的配料信息查询。

1.2 研究目的与意义

1.2.1 文字识别原理

文字识别说白了就是对带有文字的白纸或者其他文字载体进行拍摄得到图片，获取点阵的文件^[18]，然后再对图像文件进行一系列的处理，最终找到一个相匹配的结果。具体的环节如图 1-2 所示^[19]。

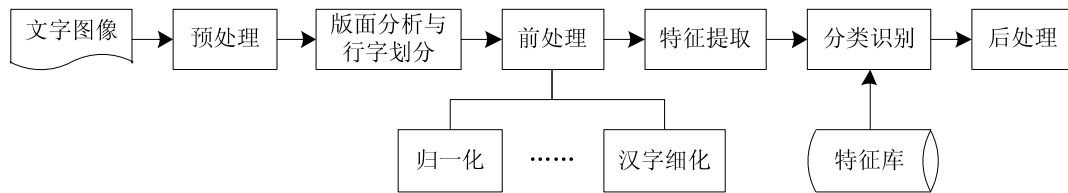


图 1-2 文字识别流程

1) 预处理

预处理也就是对图像真正进行识别前，为了简化模式提取工作，进行的一系列准备工作。比如图像受到了干扰或者倾斜，那么我们就先对干扰进行去除，对角度进行纠正，争取使文字水平，然后再将彩色图像变成二值图像等一系列手段。

2) 版面分析

所谓版面分析，就是通过分析、理解提取图像中的文本、图像、表格等不同区域。食品配料表的版面还是比较复杂的，不像文本图像，白底黑字，都是文字。食品配料表中，会有各种表格以及图文混排现象，要想进行文字识别，就要把相应的区域找出来^[20-21]。

3) 字符分割

字符分割在食品配料表的识别过程中，还是比较重要的，只有把单个字或者单个词找出来，才能进行识别。如果不能够分割出想要识别的字符，那么基本很难进行文字的识别。

4) 前处理

这个是对预处理的进一步处理。也就是针对于分割出来的字符，进行进一步加工，把笔画变细，提取出骨架，将每个字符图像统一大小，以便于下一步的具体操作。

5) 特征提取

这个就是找出字符图像的特征，是什么参数导致了这个字与其他字的不同，一旦能够提取出分辨率高的特征，就会极大地压缩信息量，提高识别速度。所以，这是一个关键性的技术，关系到食品配料表的识别正确率。

6) 文字的识别

到了这一步，就是对各种分类器进行组合，通过一系列的分类判断，找到与所要识别的字符图像相匹配的特征，然后确定识别结果。众所周知，汉字的

字符数很大，如果一个一个进行对比，那么每一个字都要很长的时间来完成。如果消耗时间超过人工输入，那么配料表的识别系统就将没有意义。所以，要设计一个多级分类器，先确定一个小的范围，再从这个范围内进行具体的识别，这样，就会大大提高效率。

7) 后处理

前文中说过，OCR 是永远不可能达到 100%的技术，尤其是食品配料表这种复杂的识别环境，拍照角度、光线干扰、背景的复杂、褶皱等，都会给识别带来很大的麻烦。因此，在识别之后，进一步的处理来提高识别率也是一个不可或缺步骤。

1.2.2 汉字识别难点

针对识别过程，可以得出汉字识别的难点：

1) 汉字量大

在 GB2312-80 中，有 3755 个一级汉字，6763 个二级汉字^[22]。文字识别系统要想达到良好的识别效果，一般要能够支持 3000-4000 个常用字，当然，对于食品配料表的文字识别可以减少到 2000 字以内。但是还是会给识别带来很大的计算量。如果用一个分类器去比较两千次，速度可想而知。而如果采用多级分类器，先进行繁简程度识别或者结构识别，再进行文字识别，必然可以提高速度。但是如果步骤太多也不好，如果最开始就错了，后面进行的就是无用功。

2) 字体多

汉语博大精深，汉字也是一门艺术，呈现出多种字体如宋体、黑体、楷体等，虽然它们拥有相同的拓扑结构，但笔画的位置、走势、长短还是有所差别，各个部分的比例和位置也不尽相同。要想保证足够的正确率，就要适应不同字体特征。

3) 结构复杂、字形相似

对于英文字母，一共只有 26 个，且每个最多 3 笔，而汉字则不同，平均就有 11 笔，多的有将近 40 笔，非常复杂。此外，不同汉字的区分度还不高，比如“日”和“曰”，就只是形状的些许差异，笔顺笔画完全相同，若想区分开，特征就要有很好的细分能力。而这需要很大的计算量，所以在这之前，就要进行一个粗分类。

1.2.3 配料表识别难点

相对于传统汉字识别，配料表识别的相对容易在于所用汉字较少，常用汉字六千多，但经统计，在配料表中出现的汉字只有一千多^[23]，大概四分之一，计算量较传统文字识别小很多，可以提高速度。然而配料表由于印刷在各种包装上面，也出现了其特有难点，如图 1-3。

- 1) 背景复杂：食品包装为了鲜艳，吸引人，一般会应用多色复杂的图案作为背景，导致背景与文字区分度很低，或者有强烈的干扰，造成文字识别困难；
- 2) 表面非平面：很多包装并非正方体，如圆柱体就会带来配料表部分文字扭曲，还有些塑料包装，容易发生褶皱，带来局部较大变形，也是识别中的难题；
- 3) 高光干扰：由于很多包装为塑料或铁质包装，有很强烈的反光效果，如果角度选取不好，会出现大面积高光，造成文字部分无法提取；
- 4) 各种字符混排：在配料表中，不只有中文，还可能出现英文、数字、连词符等，多语言的支持也是配料表识别系统的一个难点^[24]。

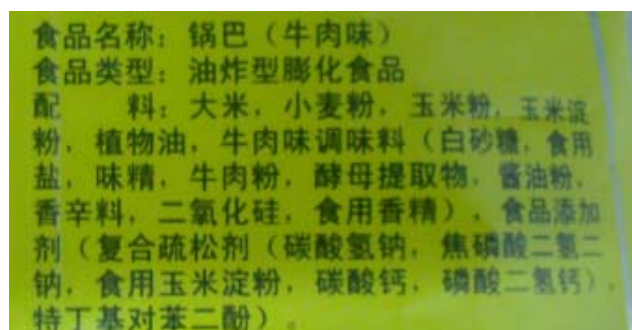


图 1-3 配料表

1.2.4 论文研究意义

本文针对文字识别过程中的主要环节尤其是配料表识别的难点进行了探讨和实践，对图像处理的各个方面也进行了探索性研究。论文通过以下几个方面的研究，使其在应用领域和理论创新均具有较大意义：

- 1) 解决了人们对于配料信息不知所云的问题；
- 2) 改善了传统信息查询方式，改进了依靠手动输入再进行搜索的繁琐操作

流程，提供了一拍即得式配料信息查询方案；

3) 提出了基于聚类分析的整词、单字分割等预处理算法，提高预处理质量；

4) 提出了基于整词的配料表识别策略，相对于传统的基于单字的识别，可以提供更多的模式信息，减小维度，提供更准确的识别结果，对其他的特定细分领域具有很好的借鉴意义。

1.3 国内外发展现状

1.3.1 发展历史

众多学者专家在上个世纪初就已经开始了相关方面的研究。一直发展到了现在，对于印刷体的英文识别、数字识别都取得了相当巨大的成果^[13]。即使是汉字的识别，也有了很大的进步。而此方面的相关研究工作在我国起步还是比国外要晚了一些。一直到了上个世纪的末期，才逐渐尝试着对汉字字符（开始只是印刷体）进行了识别方面的研究。截止到目前，主要分为了三个比较大的阶段，具体见表 1-1：

表 1-1 文字识别发展史

时间	贡献
1996 年	IBM 公司依据模板匹配的方法识别出了 1000 个印刷体汉字
1977 年	东芝综合研究所应用其开发的系统可以识别 2000 个印刷体汉字
80 年代初	日本武藏野电气研究所进一步将汉字识别数提升到 2300 多个

1) 第一阶段，上个世纪 80 年代前后

我国学者从上个世纪的 70 年代末，逐步开始接触印刷体汉字的识别，对其进行深入研究。通过对汉字图像的各种情况进行分析，提出了很多种特征提取方法，并应用这些方法进行相应系统的开发。然而，这些还只是停留在实验阶段，并没有相应的识别系统可以供大家使用。

2) 第二阶段，上个世纪 90 年后

发展到上个世纪的 90 年代，文字识别系统的发展开始步上高速公路，不仅是在实验室里面进行研究，也开始投入市场使用。在这个时期，我国的相关技术也开始发展，然而，也存在着诸如噪声干扰等问题，影响准确率。

3) 第三阶段 90 年代至今

印刷体汉字的研究到了这一阶段，已经比较成熟。其主要工作也从简单地能够识别变为了如何提高性能上面。此外，还研究了如何进行中英文混排的识别，来增强系统的鲁棒性以及稳定性。

由于起步比较晚，直到上个世纪七十年代才开始，所以我国的相关技术发展比别人慢了一拍。但是，也从国外的先进技术那里学来了很多有价值的经验，虽然晚，但是速度快。在此之后，我国也出现了很多的文字识别软件，且识别率也相当不错，有的甚至可以达到 95% 以上。

1.3.2 研究现状

目前，OCR 识别技术已经发展地相对成熟，不仅能够识别较多条件的部分字符以及符号，还能进行版面分析，识别多字体字号的混排文档。汉字字符的识别正确率可以达到 98%，对于一些印刷质量较差的字符，也令人满意地达到了 95%。手写字符识别率相对较低，如果比较工整，可以达到 70%。我国的汉字字符识别研究经过十几年的发展，虽然存在各种困难如起步晚、字符集庞大等，但还是逐渐在图书馆、新闻、邮政等各个方面进行了广泛应用。

当前存在很多专业型字符识别产品，他们面向特定的行业，如邮政、海关等。这种产品需要的字符集较小，一般可依靠专用输入设备进行快速、高效识别。

随着越来越多的移动产品如微信、手机百度等 apps 进入人们的生活，文字识别逐渐从传统的专用设备、计算机识别向智能终端发展，识别场景也更加变化多样。如有道词典识别单词、微信识别图书封面等，都从标准的文本识别向多样化的图片信息转变，这也造成了传统的识别系统不能胜任当今的应用场景。

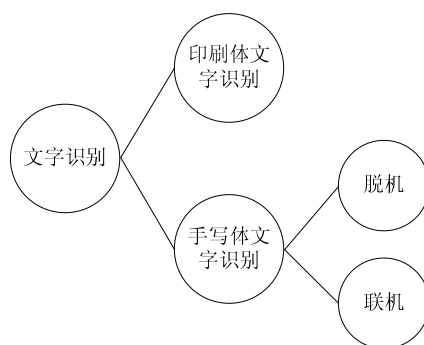


图 1-4 文字识别分类

就识别对象而言，文字识别技术的分类见图 1-4。手写体由于样本的不规范性，需要高维特征来减少不确定性的影响，而印刷体较为规范，特征提取相对稳定，对系统的实时性和准确性要求较高^[15-16]。

文字识别的理论和实践研究，一直是众多学者长期从事的课题，也有很多宝贵的实验数据和应用成果供我们参考，见表 1-2。

表 1-2 文字识别发展现状

学者	研究成果	优点	缺点
Gu Wang	将字符作为二维随机点阵	创新思路	速度慢，效果差
Lijima	多重相似度方法	解决不同字体及形近字问题	运算量大，速度慢
Akiyama	增加特征矢量维数	提高识别率，且可识别印刷体与手写体	速度慢，未建模
戴汝为	将汉字分解为四个层次	汉字识别率提高	
吴佑寿	汉字结构模型	系统地提出了汉字识别模型	

OCR 识别的目标，当然是达到完全正确，但这只能是一个梦想，也许是永远也不可能实现的难题。究其缘由，是因为在整个的识别过程中，会出现很多我们无法控制的干扰因素，这些都会给识别带来麻烦。虽然人工智能领域的一些技术对于识别会有很大的帮助，但是这还需要一个长时间的发展^[17]。

1.4 论文研究主要工作

根据前文中食品配料表识别系统的研究目的、研究要求以及国内外对食品配料表识别系统的研究现状，本文的主要研究内容如下：

本章，介绍了本文研究课题食品配料表识别系统的背景及意义。详细阐述了选题来源、配料表识别的研究背景，进而通过文字识别的发展历程、研究现状及配料表的识别难点得出论文的研究意义。

第 2 章，通过配料表识别系统应用场景分析，得出要解决的核心问题，给出初步方案及系统结构，并对服务器设计和客户端设计进行了详细描述。

第 3 章，详细介绍了预处理的各个方面。包括灰度化、二值化、倾斜校正、版面分析与字符切割、归一化、细化，并对二值化进行了改进，提出了基于聚

类分析的字符分割算法，且改进了细化畸变，使得预处理的整体效果有显著改善。

第 4 章，介绍了食品配料表识别的详细过程。在经过预处理后，对配料信息进行特征提取，经过分类识别，得出配料。提出了基于编辑距离的智能纠错算法，对识别结果进一步进行纠错，有效解决识别率无法达到 100% 的问题。创造性地提出基于整词的配料表识别策略。

第 5 章，讲解了系统的实现过程，给出了移动客户端的界面以及操作流程。与现有的配料查询方法、文字识别系统进行了横向对比。

第 6 章，对论文进行了总结和展望。

第2章 系统方案设计

2.1 系统综述

根据绪论中的分析，本文需提出一套能够应用智能手机终端拍照、进行配料识别、查询配料详细信息的系统，完成配料拍摄、后台识别、信息显示三个主要功能。

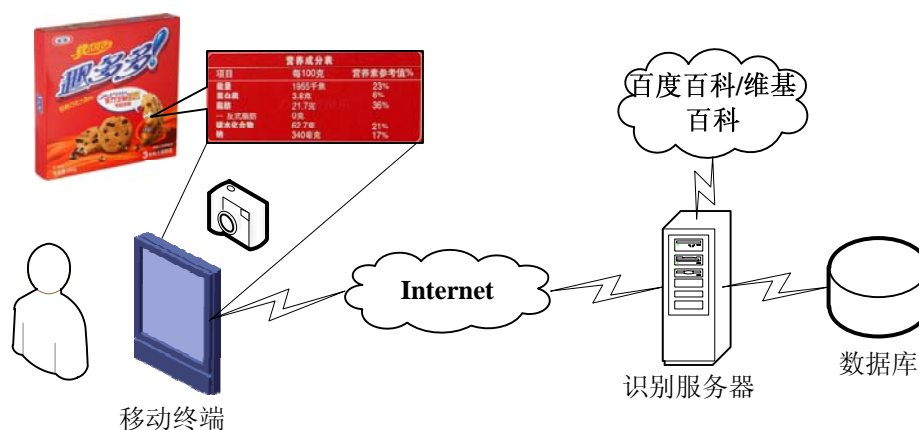


图 2-1 食品配料表识别系统

整个操作流程如图 2-1，用户拍摄食品配料表后上传给服务器，服务器进行预处理、模式提取、识别后，通过百度百科或维基百科提取配料详细信息，反馈给用户移动端。

根据操作流程，可看出整个系统所要解决的核心问题及相关技术：

1) 配料定位：在模式识别中，要解决两个问题，一个是“**What**”，一个是“**Where**”，其中更难解决的是“**Where**”的问题，目前还没有哪个系统或者算法可以较好解决此问题，因此配料的定位需要在移动终端进行限制，由使用者配合采集配料表图片。

2) 版面分割：在拍摄控制的基础上，结合配料表特性，基于投影法，引入聚类分析实现词语和单字分割技术，争取能够有效分离出配料词语或者单字，相较于传统的单字提取，提供更丰富的模式及更准确的定位；

- 3) 错误纠正: 传统的文字识别很难达到 100% 的识别率, 但是由于配料表是有限个词汇, 因此可以根据词库进行智能纠错, 以改善识别率;
- 4) 策略探索: 提出了基于整词的配料识别策略, 相较于传统的基于单字的识别, 可以提供更多的模式信息, 减小维度, 提供更准确的识别结果;
- 5) 效果改进: 鉴于移动终端应用习惯, 提出了反馈式自学习识别策略, 希望用户使用次数越多, 得到的结果也将越准确。

根据整个处理流程, 核心问题分布在不同的阶段, 见图 2-2:

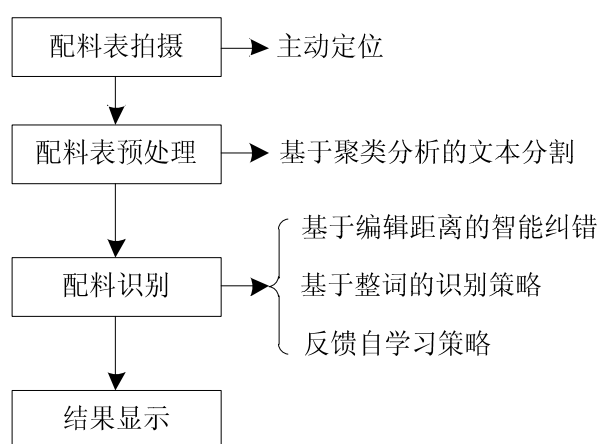


图 2-2 核心问题阶段分布

2.2 系统方案分析

配料表识别系统的一般应用场景为用户在超市或者食用产品时, 对配料信息产生疑问或者想要具体了解, 希望不依赖人工打字的繁琐操作, 快速获取配料详细信息, 对现实行为进行指导。因此, 基于操作流程, 进行系统方案的分析。

服务器基于 OpenCV 实现图像处理及文字识别^[25]; 客户端基于 Android 实现拍照及预处理^[26]。客户端通过网络通信将预处理后的照片传到服务器, 由服务器进行进一步处理、识别, 通过搜索引擎获取配料详细信息并反馈给客户端^[27-28]。

如图 2-3 所示, 整个配料表识别系统的结构可划分为移动客户端和服务端两

部分。

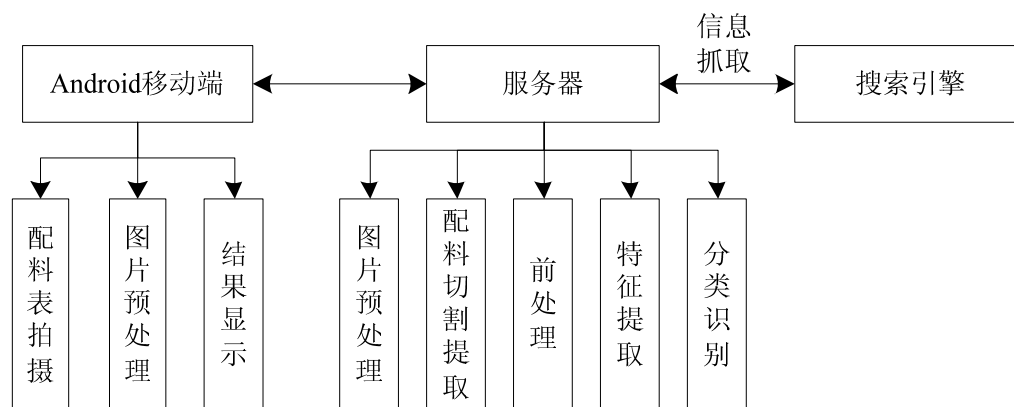


图 2-3 系统框图

其中，Android 移动端负责配料表的拍摄，图片的预处理如二值化，将图片上传给服务器并接收服务器反馈的配料详细信息显示在屏幕上。服务器接收到移动端传来的图片后，进一步进行预处理，进行倾斜纠偏、文字切割、特征提取、分类识别等，再将识别结果通过搜索引擎如百度、Google 等进行信息查询，将结果抓取下来反馈给客户端。

此外，针对配料表识别的特殊应用场景，要求满足以下功能：

- 1) 通过拍摄食品配料表，能够获取配料的详细信息；
- 2) 能够有效应对光线对识别结果的干扰；
- 3) 能够应对非平面问题，如圆柱体包装的识别、褶皱塑料包装的识别；
- 4) 能够识别多种混排字符，包括汉字识别、英文识别、数字识别、特殊符号识别以及由以上四种字符搭配成的配料。

2.2.1 服务器设计

鉴于目前手机硬件配置与电脑还是无法同日而语，为了达到良好的处理速度，提供更好的用户体验，采用服务器端进行文字识别。另外，在服务器进行识别还有一个优点就是可以获取用户反馈，搜集样本，不断提升识别性能。

为了处理高并发问题及提高事件处理速度，服务器与客户端采用 socket 通信，且服务器采用 Epoll 模型^[29]。

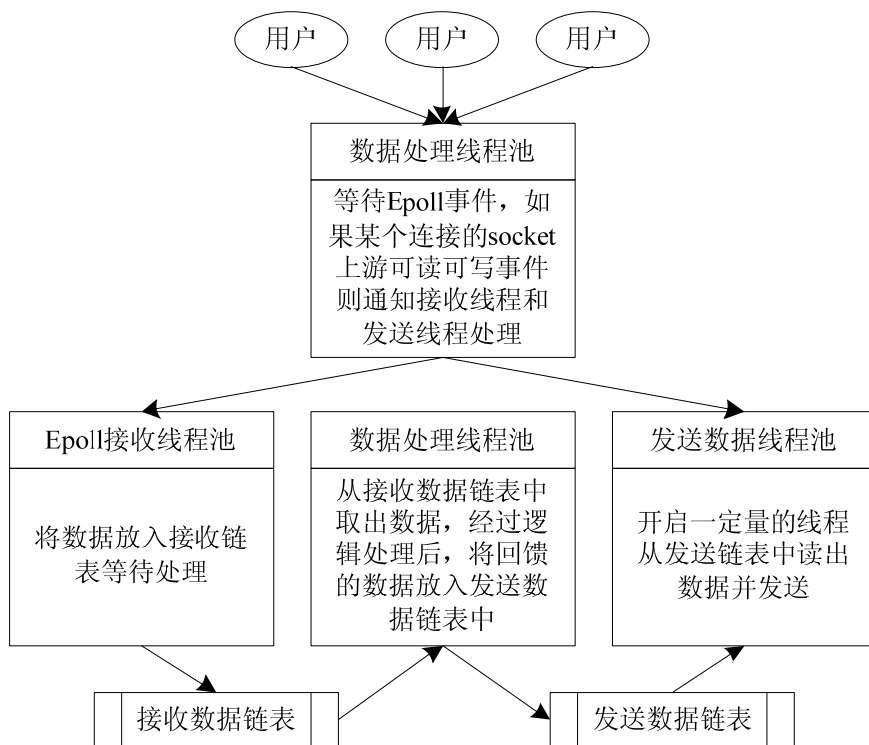


图 2-4 Epoll 模型

图 2-4 中，数据处理线程池完成主要识别工作，根据功能需求，分为配料词库解析模块、图片预处理模块、特征提取模块、机器学习模块、识别模块，经过图 2-5 进行逻辑处理。

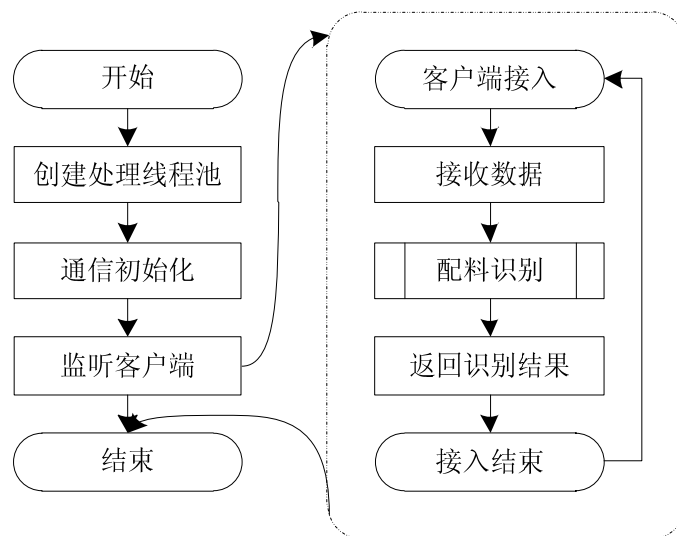


图 2-5 服务功能逻辑图

2.2.2 客户端设计

自从 Android 于 2007 年 11 月 5 日被 Google 及其开放手机联盟推出以来，凭借其开放、免费、功能强大等先天优势，吸引了全球的电信行业、手机制造商等加入到 Android 阵营，同时也吸引了众多软件开发厂商、科研机构和开发者投身其中。在与 iOS 的竞争中，Android 也逐渐处于领先，在全球主要使用智能手机的国家，Android 市场占有率全部在 50%以上，在德国、拉美、西班牙甚至超过了 80%，在国内也达到了 78.6%。所以，本文选用 Android 智能手机作为客户端。

在客户端，主要完成配料表拍摄和结果显示，具体操作流程如图 2-6：

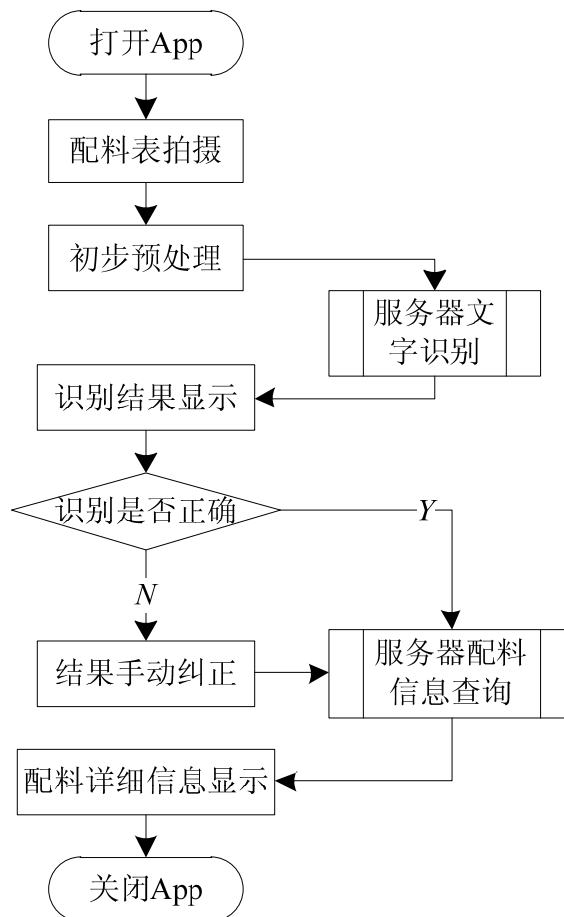


图 2-6 客户端操作流程

2.3 本章小结

本章首先进行了系统综述，通过操作流程讲解引出要解决的核心问题，结合特定环境，引出文章的创新点。其次，对食品配料表识别系统结构进行了介绍，分析了整个识别的处理流程，给出了服务器处理流程和客户端处理流程。

第3章 预处理

3.1 预处理简介

如果对一幅图像进行处理，其色彩信息量极高，但是更多的是视觉效果，对食品配料表的识别并没有太大的帮助，却会导致计算量极大。此外，大部分的图像处理算法都是基于灰度图像的，因此在食品配料表的识别过程中，最好先将其转换成灰度图像，再进行处理，这也是预处理的一部分。而在拍照的过程中，如果角度不好、强光干扰、食品包装油褶皱，这些都会给图像的质量带来很大的损失，所以要对图像中的噪声进行过滤、对角度进行纠正，确保水平或者与模板相同。此外，还要区分出背景和文字，获取相关信息，而这将用到二值化技术。以上的各种技术都可以称为食品配料识别的预处理^[30]。

对于配料表图像中的噪声也无非加性噪声、椒盐噪声等几类，有很多方法可去噪^[31]。如图 3-1 所示的几个重要预处理步骤，下文中将一一介绍。

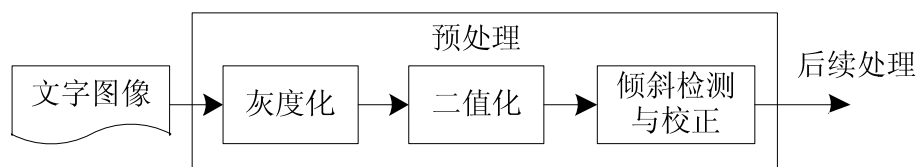


图 3-1 预处理过程

3.2 彩色图像的灰度化

目前的彩色图像，每一个像素点都用一个三维的矢量（R，G，B，即三原色）来表示。图像的灰度就是明暗程度，对于每一个像素点，0 为黑色，255 为白色，在黑白之间的 256 个等级，表示不同的明暗。我们可以根据这个特点，将三维矢量投影成一维矢量，也就是将彩色图像变为灰度图像，称为灰度化。下面介绍两种常用的灰度化方法：

- 1) 平均值法，见公式（3-1）：

$$gray = (R+G+B)/3 \quad (3-1)$$

2) 加权平均法，见公式（3-2）：

$$gray = k \times R + l \times G + m \times B \quad (3-2)$$

其中 $k+l+m=1$ ， $k=0.587$ ， $l=0.229$ ， $m=0.114$ 。

3.3 二值化

对于灰度图像，还要进一步处理，把目标点从背景点中分离出来，通过找到一个合适的阈值，将两者分开。在预处理的过程中，最关键的问题就是二值化。假设 t 为图像 $f(x,y)$ 在某区域内的灰度值范围为 G 内选取的阈值 ($t \in G$)，那么二值化的过程则可用公式（3-3）表示：

$$f_i(x,y) = \begin{cases} b_1, & f(x,y) \leq t \\ b_2, & f(x,y) > t \end{cases} \quad (3-3)$$

公式（3-3）中的 (b_1, b_2) 是一个二值化对，一般取 0 和 1，当然也可以取 0 和 255。对于阈值 t 的选取，如果太大，那么也许会把文字部分的像素点看成背景来处理；如果太小，又可能将背景的像素点当成文字，就造成了一大片的黑块。所以对于阈值的选取，是一个很关键的问题。二值化的最理想结果，就是处理后，刚好把配料表文字和后面的背景完全分开，既保留了所有文字的特征信息，又过滤了干扰。二值化一般分为两类：也就是经常提到的全局阈值法以及局部阈值法。

3.3.1 全局阈值法

此种算法是通过对整体灰度空间的分布特征进行分析，对分析结果处理后，提取一个单一的阈值 t ，进而进行处理的方法。比较典型的全局阈值法有 Doyle 提出的 p -tile 法、Ostu 提出的最大类间方差法、Prewitt 提出的直方图双峰法等。

全局阈值法计算比较简单，但是对于那些光照不均匀，背景较复杂，有很大噪声干扰的图像，处理效果不够理想^[32-33]。

1) p -分位数法

p -分位数法是一个高效简单的算法，但具有一定的局限性，即要求知道图像中目标像素和背景像素比例这一先验概率。我们假设 $k=p_1/p_2$ 是目标和背景的像素比例，就可以依此对阈值 t 进行确定，当 $f(x,y) \geq t$ 时，我们将所要确定的像素点定义为目标点； $f(x,y) < t$ 时，则相反地将所要确定的像素点确定为背景点。

2) 直方图双峰法（mode 法）

这是一个很好理解的算法。在图 3-2 中，出现了类似于驼峰的两个波峰一个波谷，相对应的是图像的前景点和背景点，它们的区分度较高时，就会形成这种现象，而阈值就可以用波谷点的灰度值代替。

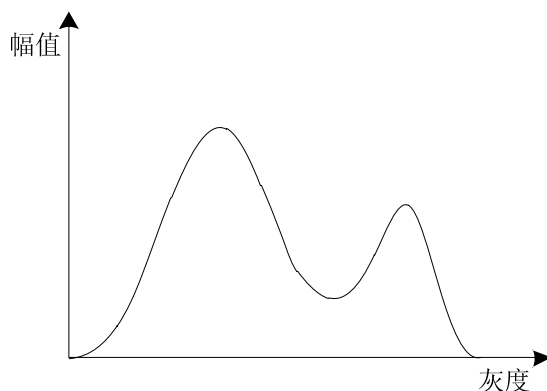


图 3-2 灰度直方图呈双峰形

3) 全局迭代法

这是一个迭代逼近法，设定一个初始值，在某种规则下，寻找一个最优值，将初始值替换，进行迭代。如定义 $g_{\max} = \max(f(x,y))$ 为最大灰度值， $g_{\min} = \min(f(x,y))$ 为最小灰度值，我们假设 $t = (g_{\max} + g_{\min})/2$ 为阈值，那么根据此二值化后会得到两个区域。对这两个区域分别进行刚才的操作，不断循环下去，一直到阈值不再变化，那么这个不变的阈值就是所谓的全局阈值。

4) Ostu 法（最大类间方差法）

该方法参考了模式识别的判别理论，如果样本分属于不同的类别，那么类别间灰度值的方差将会达到最大，而如果将灰度值分为一类，那么它们的方差将达到最小。在此思想上，Ostu 法应用最小二乘法原理，实现阈值自动选取。

公式（3-4）中的两类 C_1 和 C_2 由阈值 t 来区分：

$$C_1=\{0,1,2,\cdots,t\}, C_2=\{t+1,t+2,\cdots,255\} \quad (3-4)$$

公式（3-5）是最小二乘法推导式中三个判别准则，Ostu 法利用其中的一个来进行最大化求解，进而找到目标阈值 t 。

$$\lambda = \frac{\sigma_B^2(t)}{\sigma_W^2(t)}, \kappa = \frac{\sigma_T^2}{\sigma_W^2(t)}, \eta = \frac{\sigma_B^2(t)}{\sigma_T^2} \quad (3-5)$$

公式（3-5）中 $\sigma_W^2(t)$ 、 $\sigma_B^2(t)$ 和 σ_T^2 分别代表了为类内方差、类间方差和总方差，我们发现只有总方差是与 t 无关的函数，而其他都是关于 t 的函数。所以我们选择 η 进行计算，这样比较简便，见公式（3-6）：

$$t^* = \text{Argmax}(\sigma_B^2) \quad (3-6)$$

当图像中的目标和背景区分度很明显时，使用 Ostu 算法简单、快速。若区分度不明显，就会出现把配料表的背景作为配料词语进行划分，出现黑色块，无法进行辨别。

3.3.2 局部阈值法

如果食品配料表的背景中，有其他图案进行干扰，或者出现部分反色，若只用全局阈值法，要么将背景作为目标，导致黑色块或者多余信息，要么将目标识别为背景，造成信息丢失，这都是我们不允许的。而局部阈值法就是对目标点的邻域进行考察，获取阈值。

当然，局部阈值法也存在它的局限性：不像全局阈值法对整体进行统计，局部阈值法要计算每个像素的邻域，并寻找阈值，这就无形中极大地增加了工作量，必然会导致处理速度变慢。如果在这个情况下，图像再比较大的话，那么也许一个字的处理就要好久，不符合现实要求。此外，由于对细节处理过多，对于一些细小的噪声，也许会被识别为目标点，很难分离出文字部分。

1) Chow 和 Kaneko 的方法

将灰度分布的局部信息与直方图双峰法相结合，可以得到更好的处理效果。其基本步骤为：

- a) 将样本图像进行多等份分割；
- b) 对每个等份依据双峰法得到局部阈值；
- c) 对所有等份中的阈值进行插值化算法，得到阈值。

该方法是 1972 年 Chow 和 Kaneko 提出的。它的问题在于等份的选取，如果太多了，直方图双峰法效果差；如果太少了，那么效果又与全局阈值法差别不大。

2) Niblack 算法

Niblack 的核心是不仅仅依靠统计直方图，而是对每个目标点的邻域进行统计计算，例如公式（3-7）中，像素点 (x,y) ， $s(x,y)$ 为其邻域灰度值的均值， $m(x,y)$ 为其邻域内灰度值的标准差，通过其计算阈值。难点是邻域大小的选择。

$$t(x,y) = m(x,y) + s(x,y) + k \quad (3-7)$$

3) Bernsen 算法

我们定义 $f(i,j)$ 是图像在像素点 (i,j) 处的灰度值，则 $(2\omega+1) \times (2\omega+1)$ 为以 (i,j) 为中心的邻域大小， $T(i,j)$ 为其阈值，见公式（3-8）：

$$T(i,j) = 0.5 \times \left(\max_{\substack{-\omega \leq m \leq \omega \\ -\omega \leq n \leq \omega}} f(i+m, j+n) + \min_{\substack{-\omega \leq m \leq \omega \\ -\omega \leq n \leq \omega}} f(i+m, j+n) \right) \quad (3-8)$$

公式（3-9）为图像中各像素点 (i,j) 用 $b(i,j)$ 逐点进行二值化：

$$b(i,j) = \begin{cases} 0, & f(i,j) < T(i,j) \\ 1, & f(i,j) \geq T(i,j) \end{cases} \quad (3-9)$$

Bernsen 算法适合处理在拍照过程中，光束的干扰，但是无可避免地具有局部阈值法速度慢的特性，并且容易出现断裂、伪影等一系列对结果产生不好影响的现象。

3.3.3 二值化算法的改进

现在已经有几十种阈值计算方法被提出。其中，Otsu 算法的性能在 Trier 和 Taxt 的论证下，得到了比其他全局阈值法性能更好的结论。而 Bernsen 算法也比其他的局部阈值法效果良好。因此，本文将两者进行结合，改进二值化算法，进一步提升其性能，为食品配料表的识别铺平道路。算法详细过程如下：

- 1) 创建集合 S 和 S^* ，将 Prewitt 算子提取的图像边缘加入集合 S^* 中。
- 2) 用 Ostu 阈值法计算阈值 T ，对边缘点以外的各点进行公式 (3-10) 处理：

$$\begin{cases} f(x, y) > (1 + \partial)T \in S \\ f(x, y) < (1 - \partial)T \in S \\ ((1 - \partial)T \leq f(x, y) \leq (1 + \partial)T) \in S^* \end{cases} \quad (3-10)$$

根据公式 (3-10)，如果点 (x, y) 满足 $f(x, y) > (1 + \partial)T$ 或 $f(x, y) < (1 - \partial)T$ ，那么点 (x, y) 属于集合 S ；如果点 (x, y) 满足 $(1 + \partial)T \leq f(x, y) \leq (1 + \partial)T$ ，那么点 (x, y) 属于集合 S^* 。

3) 首先对 S^* 中的各点进行一系列处理。依据公式 (3-11) 对 Bernsen 算法进行改进，以防止个别点的干扰。

$$T_1(i, j) = 0.5 \times (\sec_{\substack{-\omega \leq m \leq \omega \\ -\omega \leq n \leq \omega}} \max f(i + m, j + n) + \sec_{\substack{-\omega \leq m \leq \omega \\ -\omega \leq n \leq \omega}} \min f(i + m, j + n)) \quad (3-11)$$

$$T_2(i, j) = 0.5 \times (\sec_{\substack{-\omega \leq m \leq \omega \\ -\omega \leq n \leq \omega}} \max f(i + m, j + n) - \sec_{\substack{-\omega \leq m \leq \omega \\ -\omega \leq n \leq \omega}} \min f(i + m, j + n)) \quad (3-12)$$

$$T_3(i, j) = \text{avg}_{-\omega \leq m, n \leq \omega} f(x + m, x + n) \quad (3-13)$$

4) 如果 $T_2(i, j) > \partial T$ ，那么我们采用公式 (3-11) 中的 $T_1(i, j)$ 对灰度图像进行二值化；如果 $T_2(i, j) < \partial T$ ，那么我们采用公式 (3-13) 中的 $T_3(i, j)$ 进行二值化。

5) 对集合 S 中的点，直接用全局阈值 T 进行二值化。

图 3-3 为原始输入图片，图 3-4 是通过 Bernsen 算法处理后的图片，图 3-5 是应用 Ostu 算法处理后的图片，图 3-6 为采用本文算法得到的二值化后的图片。

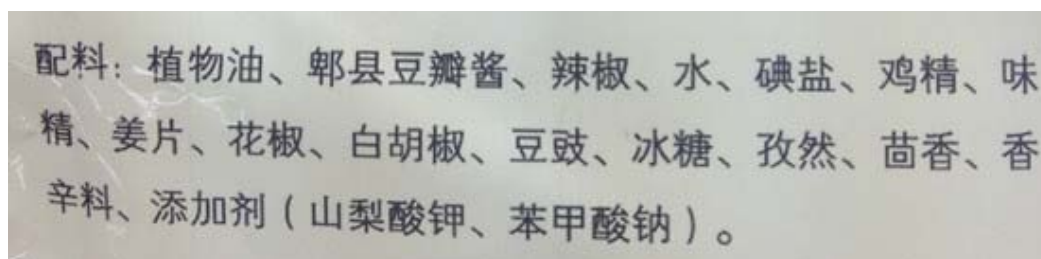


图 3-3 二值化效果实验图——原始图片

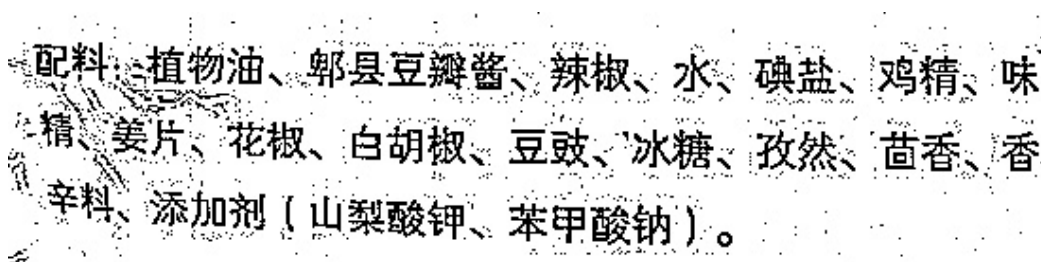


图 3-4 二值化效果实验图——Barnsen 算法

配料：植物油、郫县豆瓣酱、辣椒、水、碘盐、鸡精、味精、姜片、花椒、白胡椒、豆豉、冰糖、孜然、茴香、香辛料、添加剂（山梨酸钾、苯甲酸钠）。

图 3-5 二值化效果实验图——Ostu 算法

配料：植物油、郫县豆瓣酱、辣椒、水、碘盐、鸡精、味精、姜片、花椒、白胡椒、豆豉、冰糖、孜然、茴香、香辛料、添加剂（山梨酸钾、苯甲酸钠）。

图 3-6 二值化效果实验图——改进算法

从实验结果来看，本文采用的改进算法取得了良好的效果。

3.4 倾斜检测与校正

在配料表的拍摄过程中，由于是人工使用手机进行拍摄，难免会有角度问

题，如果拍摄的角度偏差不大，那么可能对结果影响不大，但是如果角度偏差超过 30 度，也许计算机看来就是另外的东西了^[34]。

3.4.1 倾斜检测

1) 基于投影的方法

投影就是对某个方向上黑色像素点进行统计求和的结果^[35]。 $f(x,y)$ 在 x 和 y 方向的投影分别为：

水平投影，见公式 (3-14)：

$$h(y) = \sum_{x=0}^{m-1} f(x, y), (0 \leq y < n-1) \quad (3-14)$$

竖直投影，见公式 (3-15)：

$$v(x) = \sum_{y=0}^{m-1} f(x, y), (0 \leq x < m-1) \quad (3-15)$$

2) 基于 Hough 变换的方法

Hough 变换就是笛卡尔坐标系到参数空间的映射，公式 (3-16) 为其转换公式。

$$\rho = x \cdot \cos\theta + y \cdot \sin\theta \quad (3-16)$$

参数 θ 可以取不同的值，当 θ 变化时，就会出现不同的点被映射到参数空间。如果参数空间中有曲线相交，通过变换，可得知在笛卡尔坐标系出现了直线。通过这种方法，可以发现图像中的直线，进而得到偏角。

根据公式 (3-17) 用参数空间中的值对目标点进行逐一变换。

$$\rho_i = x \cdot \cos\theta_i + y \cdot \sin\theta_i \quad (1 \leq i \leq n) \quad (3-17)$$

Hough 变换的优点是抗干扰能力强，但其缺点也比较明显，就是比较消耗时间和空间，在进行多次变换时，带来很大的计算量。

3) K_最近邻法

K_最近邻法是对连通区域中心点的统计，通过连接相邻的中心点，计算出矢量方向，然后统计计算出页面倾斜角度。

3.4.2 倾斜校正

倾斜校正就是在倾角的测量后，对其进行纠正，以使待识别图像变为水平方向，以方便配料表的识别。但是在旋转的过程中，有可能造成像素点丢失，为了弥补这一问题，一般要进行插值计算^[16-17]。

如图 3-7 所示，以图像中心为原点，建立直角坐标系，进行旋转变换。

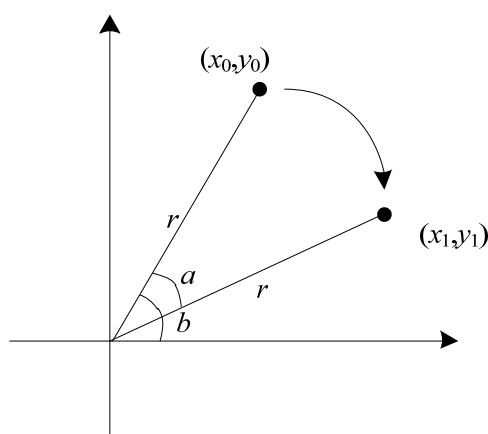


图 3-7 倾斜校正

根据图 3-7 可得旋转前公式 (3-18):

$$\begin{cases} \rho = x \cos \theta + y \sin \theta \\ x_0 = r \cos b \\ y_0 = r \sin b \end{cases} \quad (3-18)$$

旋转后见公式 (3-19):

$$\begin{cases} x_1 = r \cos(b-a) \\ y_1 = r \sin(b-a) \end{cases} \quad (3-19)$$

根据正弦加法定理和余弦加法定理可得到公式 (3-20):

$$\begin{cases} x_1 = r \cos b \cos a + r \sin b \sin a = x_0 \cos a + y_0 \sin a \\ y_1 = r \sin b \cos a - r \cos b \sin a = -x_0 \sin a + y_0 \cos a \end{cases} \quad (3-20)$$

用矩阵的形式表示为公式 (3-21):

$$[x_1, y_1, 1] = [x_0, y_0, 1] \times \begin{bmatrix} \cos a & -\sin a & 0 \\ \sin a & \cos a & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3-21)$$

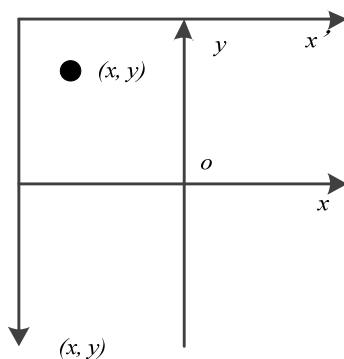


图 3-8 两种坐标的关系

如图 3-8 所示, 设图像的宽为 w , 高为 h , 可以得到公式 (3-22):

$$[x, y, 1] = [x', y', 1] \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ -0.5w & 0.5h & 1 \end{bmatrix} \quad (3-22)$$

则通过逆变换可得公式 (3-23):

$$[x', y', 1] = [x, y, 1] \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ -0.5w & 0.5h & 1 \end{bmatrix} \quad (3-23)$$

则最终变换步骤可为：

- 1) 将坐标系由 $x'o'y'$ 变为 xoy ;
- 2) 在 xoy 坐标系下进行旋转变换;
- 3) 将坐标系 xoy 变回 $x'o'y'$ 。

3.5 基于投影与聚类的版面分析与字符切分

3.5.1 配料表版面分析介绍

配料表的版面识别就是通过分析所拍摄的配料表图像，从中找出文字区域、图片区域、表格区域等。在整个寻找过程中，依靠计算机自动完成^[36-37]。当然，目前也有些系统依靠用户进行手动选择，但是无疑带来了操作的不便利性。

食品配料表的版面，尤其是需要拍摄的版面，一般可能有以下几种结构：

- 1) 纯文本区域：由各种数字、符号、文字等组成的区域。
- 2) 图形区域：配料表中的图案、logo 等非文字部分。
- 3) 表格区域：主要包括一些统计报表类信息。

进行配料表识别，有一个先天优势，就是不需要识别整个商品的版面。使用者可以人为地过滤掉过多无效信息，直接对配料表部分进行识别。

本文又将这四类基本的版面构成分为有效识别区域和无效识别区域，只对文字区域和表格区域进行识别。例如，可以调整摄像头位置，对准要识别的配料，进而应用版面分析，将其从背景及其他配料中分离出来，进行文字识别。

3.5.2 版面分析常用方法

版面分析一般分为从局部到整体以及从整体到局部两种大的方向^[38]。

1) 从整体到局部

这种方法也有一种叫法叫做“自顶向下”，也就是对一个图形进行分割，找出每一个区域分界线。对于印刷的文本处理效果比较好。此类方法的代表有：RLSA 算法、基于 X-Y 切割法、基于投影二分的方法等^[39]。

2) 从局部到整体

这种方法也叫做“自底向上”方法。与上一个方法相反，是从细节处，找到与之类似的区域进行合并，确保不同区域之间没有交叉。这种方法灵活性更

强，但是处理速度较慢。

虽然版面分析算法较多，但基本是根据以上方法引出的投影法和联通域法 [40-41]。

3.5.3 基于投影法的文档图像分析

投影法其实就是一个降维的过程。把二维平面变成一维向量，进而根据一定特性进行切割。对于一幅二值化后的图像，在水平或者数值方向上，统计每一列或者一行上面黑色像素也就是目标像素点的总和，其结果作为投影值。我们定义原始图像的二维函数是 $f(i,j)(i=1,2,\cdots,M; j=1,2,\cdots,N)$ ，对其在两个方向上进行投影可得到 $V(j)$ 和 $H(i)$ ，见公式 (3-24)。

$$\begin{cases} H(i) = \sum_{j=1}^N f(i,j) \\ V(j) = \sum_{i=1}^M f(i,j) \end{cases} \quad (3-24)$$

图 3-9 为原始配料表文档，图 3-10 为对其进行行投影的结果，图 3-11 为对其进行列投影的结果。

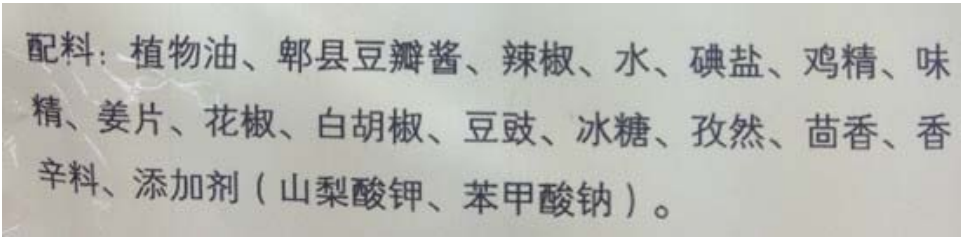


图 3-9 原始文字图片



图 3-10 行投影示意图



图 3-11 列投影示意图

由图 3-9~图 3-11 分析可知，我们可设定阈值 θ ，对投影的结果进行判别，根据新的投影公式（3-25）可得到分割结果。

$$\begin{cases} H(i) = \begin{cases} H(i), & H(i) \geq \theta \\ 0, & H(i) < \theta \end{cases} \\ V(j) = \begin{cases} V(j), & V(j) \geq \theta \\ 0, & V(j) < \theta \end{cases} \end{cases} \quad (3-25)$$

3.5.4 基于聚类分析的文本切割

鉴于配料表的特殊性，一般不涉及表格、公式等，文字独立性较强，可在投影法的基础上进行聚类分析。对相邻两字间的距离应用直接聚类法^[42]。

首先，将所有距离当作一类，再根据距离最小原则，选出一对分类合并成一个新类，这样迭代下去，经过 $m-1$ 次，即可将全部样本归为一类。

找出字间距及词间距，得到最佳阈值来对词语进行切割，如图 3-12：

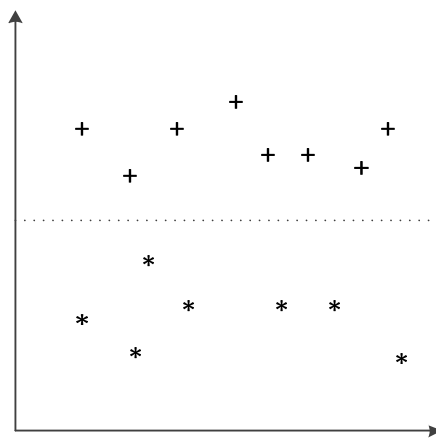


图 3-12 基于聚类分析的阈值确定

这样，根据所得阈值及中心点，提取出所要查询的配料。

水、白砂糖、蛋白质、
果胶、甜味剂、氨基酸、
葡萄糖、石碱汁

图 3-13 原始配料表

对图 3-13 其进行聚类的系谱图为图 3-14：

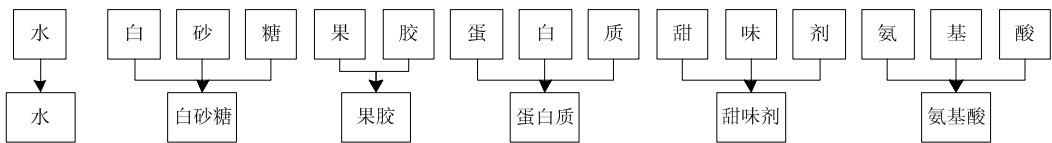


图 3-14 聚类系谱图

得到甜味剂识别结果见图 3-15：

水、白砂糖、蛋白质、
果胶、甜味剂、氨基酸、
葡萄糖、石碱汁

图 3-15 文本切割结果

3.6归一化

对于汉字来说，如果要统计其四周面积，那么图像尺寸越大，面积越大，尺寸越小，面积特征越小，这样就很难进行区分。有两种方案解决这个问题：

- 1) 对于不同大小的原始图像，建立不同的特征库；
- 2) 将所有的字符图像变成统一的大小，也就是所谓地采用归一化^[43]。

显而易见，如果采用方法（1），那么就要建立多个特征库，并且要想拥有较好效果，就需要更多的特征库，是一个很难完成的任务。即使能够完成，也造成了空间和时间的巨大浪费。所以，一般采用方法（2）的归一化的方法。

本文主要采用线性归一化方法^[44]，将原始图像，按照一定的比例进行伸展或者缩小变换，使之大小变为目标大小。设 $f(x,y)$ 为原图像， $g(x,y)$ 为归一化后图


像，它们的大小分别为 $M_1 \times N_1$ 和 $M_2 \times N_2$ 。那么 $f(x,y)$ 中的任意点 (x_i, y_i) 对应于 $g(x,y)$ 中的点 (h_i, v_i) 见公式 (3-26)：

$$\begin{cases} h_i = \frac{M_2}{M_1} \times x_i \\ v_i = \frac{N_2}{N_1} \times y_i \end{cases} \quad (3-26)$$

公式 (3-26) 适用于图像缩小的情况，如果是放大，则可能产生像素丢失，就是在大图存在没有办法和小图对应的点，要经过插值计算才能够填充。

我们对四种常用的插值计算进行了对比，结果见表 3-1。从中可以看出，最近邻插值有较严重的锯齿现象；线性插值和三次样条插值虽然有效去除了锯齿，但是对于二值图像却造成了很多灰色像素，区域插值效果较为理想，也是本文所采用的方法。

表 3-1 四种不同插值算法归一化放大对比

原图像	最近邻插值	线性插值	三次样条插值	区域插值
				

3.7 汉字细化

汉字细化的本质，是把不同字体的笔画变成单一像素，只保留汉字的骨架，却不保留粗细的信息。细化后的汉字，既包含了不同笔画间的拓扑结果，又减小了信息量，提高处理速度^[45]。良好的细化效果要满足以下四点：

- 1) 细化后的笔画应该保持原笔画布局，最好是中轴线；
- 2) 最好不要断掉，一笔变为两笔；
- 3) 要保持原来笔画间的关系；
- 4) 宽度为单一像素。

p_3	p_2	p_1
p_4	p	p_0
p_5	p_6	p_7

图 3-16 邻域示意图

目标像素点 p 的 8 邻域如图 3-16 所示。令像素值为 1 的点为前景点，像素值为 0 的点为背景点，对目标像素的 8 邻域分别进行算术逻辑运算，根据结果判断像素是否应该删除。其细化过程分为 2 个步骤：

1) 对于前景目标像素 p ，若其 8 邻域点满足以下条件：

a) $2 \leq N(p) \leq 6$

b) $S(p) = 1$

c) $p_0 \cdot p_2 \cdot p_6 = 0$

d) $p_0 \cdot p_4 \cdot p_6 = 0$

则标记点 p 为待删除点，完成本次全部扫描之后将其删除。其中， $N(p)$ 为 p 的 8 邻域中非零点的个数， $S(p)$ 为 8 邻域顺时针像素变化的次数。

2) 若 p 满足条件 a 和 b 以及：

c') $p_0 \cdot p_2 \cdot p_4 = 0$

d') $p_2 \cdot p_4 \cdot p_6 = 0$

同样标记 p 待删除。

反复迭代 1) 和 2)，直至没有可删除的点后，就生成了字符细化后的骨架，如图 3-17：



图 3-17 细化结果

3.8 本章小结

本章主要介绍了一些食品配料表的预处理方法。通过对一些预处理方法如二值化、文本切割等进行改进，达到良好的预处理效果，为后续识别提供基础。

第4章 配料识别

4.1 汉字特征提取

4.1.1 特征提取概述

对于一幅图像，假设长宽分别为 M 、 N ，则其信息量或者说维度为 $M \times N$ ，这无疑是很大的。如果将所有配料表提取出来的字与字库的图像文档进行对比，是极其不合实际的。其实，要想区分不同的配料文字，并不是要把文字的所有信息全部提取出，只需要了解文字的特征即可。如一横，就代表汉字“一”，而不是一个点阵。从专业角度而言，特征就是从高维变化到低维后的信息。通过降维，最大限度地保留特征信息，又尽量减小维度，从而提高识别率和识别速度。因此，我们在设计食品配料表识别系统时，要保证以下几个原则^[46]：

- 1) 要选取能够反映配料表中字符的本质的特征，要求同类字符的特征距离小，不同类距离大，具有较强的分类能力。
- 2) 提取方便，并且维数越低越好。
- 3) 能够适应一定的噪声。在食品配料表的拍摄过程中，难免会有很多的光照和颜色干扰，系统要能够从容应对这些情况。

到目前为止，世界上很多学者对汉字的特征提取都取得了相当有效的研究，也提出了各种方法来进行识别，虽然每个方法都具有其特定的优缺点，但是基本也是分为结构特征和统计特征两大类。

4.1.2 基于统计的特征提方法

该方法将食品配料表图像点阵进行变换提取，这是一种在整体上的特征提取，速度快，算法简单，且不包含局部信息，因此会造成很多信息丢失，造成识别效果较差^[47]。

1) 全局变换特征

把汉字图像作为二维点阵图像，将对其进行的各种变换的变换系数作为特征进行提取。常用的变换有 Fourier 变换、Hough 变换、Rapid 变换。若直接对

这些系数进行求解，计算量一般非常大。通常解决这个问题的方法是对图像进行投影，降成一维，再对一维数据进行全局变换。由于在降维的过程中对方向上的像素点进行了积累，从而增加了抗干扰性。虽然这个方法的抗干扰能力强、算法简单，但是笔画的细微差别如长短、形态，都能给结果带来很大影响。

2) 笔画穿透数特征

汉字的笔画密集程度可用笔画穿透数特征来表示，这一特征也在某种程度上反映了汉字的繁简程度。提取步骤如下：

- a) 将对汉字图像进行归一化，之后分成多等份（如 8 份）；
- b) 统计每一个等份中单个方向上黑白像素的变化次数，也就是经过了多少次汉字笔画；
- c) 按 a~b 步，对水平、竖直、和对角线四个方向进行特征提取。

这种方法速度快的同时，对于笔画粘连的情况效果较差。

3) 四周面积编码特征

鉴于方块字的特性，汉字四周亦包含相当丰富的结构信息。而汉字的四周笔画一般较少，很少出现内部笔画那样的粘连、断裂等干扰，具有很好的抗噪声能力。

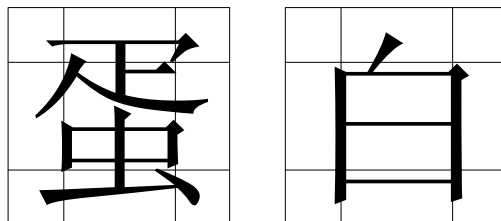


图 4-1 四周面积编码示意图

如图 4-1 中所示，四周面积编码指的是在汉字的四周分割出四个大小合适的边框，计算每个边框内黑色像素数，并依据其格式进行划分量化，得到一个序列，这个序列就是所求的四周面积编码。

4.2 分类器原理

文字识别是食品配料表识别系统中最关键的一步——根据提取的特征进行分类决策。在一般的模式识别算法中，对于分类器的设计有多分类器和单分类

器，而前者是后者的集成^[48]。

对于小规模分类识别，可以选用 SVM 方法、神经网络等单分类器，而对于较多的样本，则只适合选取某个特定的分类器解决某一方面的问题，通过多个分类决策得到最终结果。

要想达到较好的文字识别效果，以前经常进行多个单分类器分别进行识别，然后从中选一个最好的结果，并将其分类器推广使用。这种方法其实只是一个分类器的选择，并没有在识别率上下功夫^[33]。我们可以选择多个分类器相互作用，利用其互补性来进行识别的判断。多分类器的集成在组织结构上分为级联和并联两种形式。在食品配料表的识别系统中，我们选用树形分类器（如图 4-2），主要原因有：

- 1) 将复杂问题简单化，将单一问题多级化。将配料表信息的识别转换成多种参数的识别，针对于每一种特征利用不同的分类器进行分类，再将多级的相互作用结果作为最后的识别结果，其速度快、精度高。
- 2) 汉字字库巨大，即使是食品配料表也有一千到两千字，如果采用单一的分类器，必然导致时间的巨大损耗，给使用者带来很差的体验。

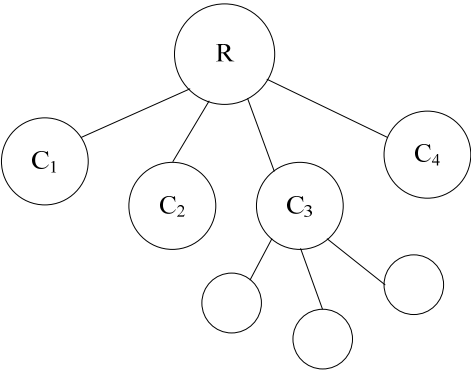


图 4-2 树形分类器

树型分类器的主要目标就是提高速度，但是也不能只考虑速度而忽略其相互之间的各种制约，给结果的精确度造成很大影响。如果分的级别很多，那么由于每个分类都可能造成结果误差，通过多层误差的积累，很难达到好的识别效果。只有通过对多种因素的权衡，选择和设计出一个合理的分类器，才能在速度和准确率上达到一个平衡。

4.2.1 距离分类器介绍

距离分类器是指对于提取出来的不同特征，在其特征空间内，依据某一规则进行计算，来度量出不同特征间的差别，作为距离进行计算，并依次来判断特征向量属于哪一类。

1) 距离分类器常用算法介绍

设存在 M 个类别 $\Omega_1, \Omega_2, \dots, \Omega_M$ ，每个类别有其训练样本集和识别样本，用 $\{X_1^{(m)}, X_2^{(m)}, \dots, X_{km}^{(m)}\}$ 和 $X_i^{(m)} = (X_{i1}^{(m)}, X_{i2}^{(m)}, \dots, X_{iN}^{(m)})$ 表示。

a) 平均样本法

此方法指的是对训练样本集进行统计计算，找出其中一个最能代表这个类别的样本作为该类的代表。也就是找出一个样本，计算出其到其它样本的距离最短。可以用公式（4-1）的函数表示：

$$T(m) = \frac{1}{K_m} \sum_{i=1}^{K_m} X_i^{(m)} \quad (4-1)$$

由于一个样本集只用一个样本来代替，那么就减小了大量的计算量，使得 $M \times N$ 次计算减少到了 M 次计算。但是此种方法的缺点是有时候效果较差。

b) 平均距离法

公式（4-2）为模式 X 与类别 Ω_i 的平均距离。

$$d(X, \Omega_i) = \frac{1}{K_i} \sum_{j=1}^{K_i} d(X, T_j^{(i)}) \quad (4-2)$$

与平均样本法不同，需要计算出每个类别中每个样本的距离，再计算其平均值，然后根据平均值进行分类，计算量比较大。

c) 最近邻法

最近邻法指的是遍历特征空间中的所有样本点，计算与待识别样本的特征距离，找到特征距离最小的样本点的类别作为识别类别，见公式（4-3）。

$$d(X, \Omega_i) = \min_{1 \leq j \leq k_i} d(X, T_j^{(i)}) \quad (4-3)$$

最近邻法和平均距离法类似，也需要存储所有样本，导致了时间、空间的浪费，并且抗干扰能力较差，容易产生误识。

d) K -近邻法

K -近邻法是根据距离 X 最近邻的 K 个样本点中多数点的类别进行分类，而不是找最近的一个，从而减小了噪声的干扰。

在 K -近邻法中，选择 K 值是一个非常重要的工作。如果 K 值太大，就很难进行分类，如果太小又会受到噪声干扰，特别地，如果 K 为 1 时，就会变成最近邻法。

2) 距离的度量

距离函数的定义形式很多，下面就其中的几种进行一下介绍。设 $X=(x_1, x_2, \dots, x_n)^T$, $Y=(y_1, y_2, \dots, y_n)^T$ 为 n 维空间中的两点。

欧几里德距离(Eucidean Distance), 见公式 (4-4):

$$d(X, Y) = \left[\sum_{i=1}^n (x_i - y_i)^2 \right]^{1/2} \quad (4-4)$$

街市距离(Manhattan Distance), 见公式 (4-5):

$$d(X, Y) = \sum_{i=1}^n |x_i - y_i| \quad (4-5)$$

明氏距离(Minkowski Distance), 见公式 (4-6):

$$d(X, Y) = \left[\sum_{i=1}^n |x_i - y_i|^m \right]^{1/m} \quad (4-6)$$

角度相似函数(Angle Distance), 见公式 (4-7):

$$d(X, Y) = \frac{X^T \cdot Y}{\|X\| \|Y\|} \quad (4-7)$$

其中 $X^T \cdot Y = \sum_{i=1}^n x_i y_i$ 表示了矢量 X 和 Y 之间的内积，而 X 与 Y 之间夹角的余弦则用 $d(X, Y)$ 表示。

4.2.2 识别的策略

一般的识别策略为图 4-3。

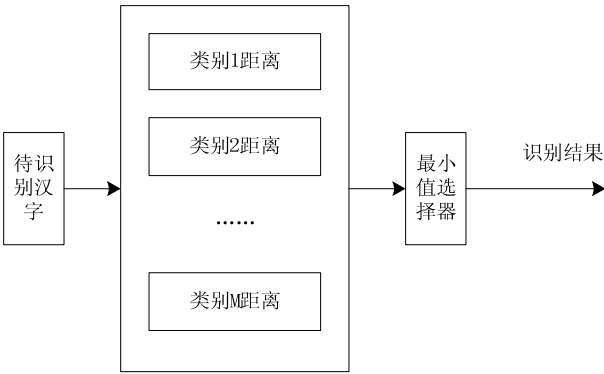


图 4-3 识别过程

4.3 基于智能纠错的分类识别策略

本文统计了配料表中出现的常用字类，属于一个大字符集的中英文混排识别问题。为了既达到较快的识别速度，又得到较高的识别精度，本文在构建系统时采用了基于模板匹配的思想。由于样本较多，若直接进行逐一匹配，会导致识别时间过大，影响用户使用体验。因此，本文采用二级分类策略，如图 4-4，其中一级分类器进行粗分类，给出候选类别，二级分类器进行细分类，给出具体类别^[49]。在识别出配料表后，依据词库对其进行智能纠错，进一步完善识别效果。

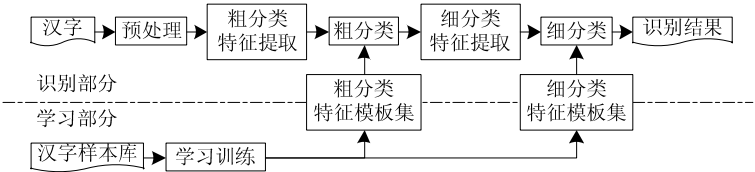


图 4-4 二级分类策略

4.3.1 分类器——粗分类

粗分类的目的是找出一个包含识别样本的小集合，一般有以下要求：

- 1) 正确性与稳定性高。由于两级分类器是级联的，如果粗分类识别错误，那么细分类即使再好也不会得到正确结果；
- 2) 速度快。如果粗分类过慢，则会大量耗时，必须要求算法简单速度快才能保证后面的识别；
- 3) 特征简单。要保证分类字典小，以节约存储空间；
- 4) 要与细分类相协调，粗分类与细分类是一个整体，必须一起解决问题。

在系统进行粗分类时，我们采用笔画边缘特征和笔画密度特征，得到 $48 \times (4+4)=384$ 维向量。接下来进行距离判定，确认输入样本和模板的相似性。该测度距离采用绝对值，其优点是即使特征向量在少数维度上面有较大的差别，只要保证其他维度很接近，也会被判定为两者匹配；其缺点也比较明显，就是不能区分相似字，但本来粗分类的目标也是识别一个子集，而不是确定的字，已经能够满足要求。

绝对值距离测度的表达式为公式（4-8）：

$$d(X, m_i) = \sum_{k=1}^K |x_k - m_i^k| \quad (4-8)$$

其中， $X=(x^1, x^2, \dots, x^K)$ 为待识别样本粗分类特征向量， $m_i=(m_i^1, m_i^2, \dots, m_i^K)$ 为字 ω_i 的粗分类标准模板， K 为粗分类特征维数。

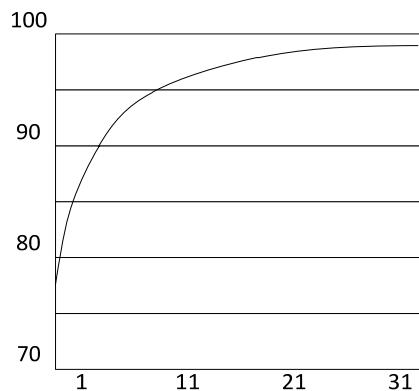


图 4-5 候选类别数——平均识别率

为了使粗分类具有良好的稳定性和一定正确性，我们对细分类个数进行了统计，找出最优的细分类候选字个数。从图 4-5 中可以发现，在细分类候选字个数大于 30 以后，识别率提高很小，却增加了很大的计算量。通过取舍，我们选择 30 作为细分类候选字个数，既保证了处理速度，又保证了识别率。

4.3.2 分类器——细分类

细分类与粗分类不同，细分类最重要的是识别精度，因为只有 30 个候选字，即使算法再复杂，也基本可以接受。因此无论是特征值提取还是识别算法，我们都采用较为精确的方法，以保证识别率。

在细分类过程中，为了获得更多的特征，我们采用 $7 \times 7 + 8 \times 8$ 的二重分割方法，如图 4-6 所示。对文字图像进行交叉分割，以免出现分割边缘不稳定情况的发生。这样在第一层的边缘部分就落在了第二层的中央，有效地保留了各种细微特征。接下来计算每个小区域的穿透特性，得到特征向量。

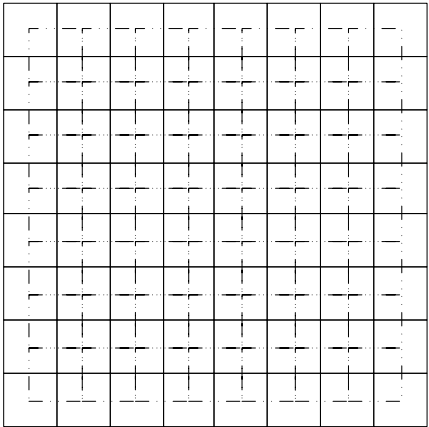


图 4-6 图像分割方式

通过统计汉字特征向量，发现其不仅具有分散性，而且其分散程度有很大不同。因此我们不仅可以由绝对值距离测度获取未知样本与不同字类的相似性，还可以引入标准差距离测度，见公式（4-9）。

$$d(X, m_i) = \sum_{t=1}^T \frac{|x_t - m_i^t|}{\sigma_i^t} \tag{4-9}$$

其中, $X=(x^1, x^2, \dots, x^T)$ 为待识别样本粗分类特征向量, $m_i=(m_i^1, m_i^2, \dots, m_i^T)$ 为字 ω_i 的粗分类标准模板, $\sigma_i=(\sigma_i^1, \sigma_i^2, \dots, \sigma_i^T)$ 为字类的标准差特征, T 为细分类特征维数。若 $d(X, m_j)=\min d(X, m_j)$, 则判决 X 属于 ω_j 类。

4.3.3 基于编辑距离的智能纠错

我们在使用搜索引擎时都发现了一个人性化的功能, 即对拼写错误进行了智能纠错^[50]。当用户在搜索框完成输入后, 搜索引擎会对其进行分析, 检查是否有拼写错误, 若存在错误, 就会给出正确的拼写建议。其实搜索引擎在这个过程中完成了两个工作, 即拼写检查和智能纠错^[51]。

1) 常用纠错法

目前主流的词语纠错方法主要有误拼词典法、N-gram 法和最小编辑距离法。

a) 误拼字典法。这是一种穷举的方法, 通过搜集大规模的拼写错误文本, 建立一个错误集, 每次检查时对比错误集中的错误, 进而给出修改建议。这种方法对时间和空间的耗费都是很巨大的, 是比较原始的方法, 不建议采用。

b) N-gram 法。基于 n 元文法, 通过对大规模的文本进行统计得到一个词间转移概率矩阵、当检测到输入词不在词典中时, 检查转移概率矩阵, 将概率大于阈值的词作为纠错建议。

c) 最小编辑距离法。通过计算输入字符串与某个字符串间最小的编辑距离来对错误输入进行纠正。也就是说要改正错误输入, 最少要进行几次改动。

2) 拼写纠错功能的实现

本文学习搜索引擎的这种方式, 动态维护一个配料表词典^[52]:

首先, 提取 GB2760-2011 中食品添加剂标准写法, 破解搜狗输入法、百度输入法、QQ 输入法等主流输入法的细胞词汇, 将其中的配料词汇整理后, 作为配料词典;

此外, 为用户提供更改功能, 若识别有误, 用户可手动更改正确配料, 并自动提交给服务器, 由服务器将图片样本保留, 将配料加入配料词典;

最后运用最小编辑距离进行智能纠错。

编辑距离一般有插入、修改、删除三种操作。假设 Σ 为有限字符集, Σ^* 是定义在 Σ 上的所有有限长度字符串集合。并设 $X=X_1, X_2, \dots, X_n$ 是 Σ^* 中的一个字符串, 其中 $1 \leq i \leq n$ 。函数 γ 表示编辑操作的代价。若 X 转换为 Y 所经过的编辑序列为 e_1, e_2, \dots, e_m , 这个代价被定义为公式 (4-10):

$$\gamma(S) = \sum_{i=1}^m \gamma(e_i) \quad (4-10)$$

给定 $X, Y, X \in \Sigma^*$, 到 Y 的编辑距离 $\delta(X, Y)$ 可定义为公式 (4-11):

$$\delta(X, Y) = \min \{ \gamma(S) | S \text{ 为 } X \text{ 到 } Y \text{ 的编辑序列} \} \quad (4-11)$$

上式也可递归定义为公式 (4-12):

$$\delta(X_1, X_2, \dots, X_i; Y_1, Y_2, \dots, Y_j) = \min \left\{ \begin{array}{l} \delta(X_1, X_2, \dots, X_{i-1}; Y_1, Y_2, \dots, Y_j) + \gamma(X_i \rightarrow \varepsilon) \\ \delta(X_1, X_2, \dots, X_{i-1}; Y_1, Y_2, \dots, Y_{j-1}) + \gamma(X_i \rightarrow Y_j) \\ \delta(X_1, X_2, \dots, X_i; Y_1, Y_2, \dots, Y_{j-1}) + \gamma(\varepsilon \rightarrow Y_j) \end{array} \right\} \quad (4-12)$$

式中, $\delta(\varepsilon, \varepsilon) = 0$; 如果 $a \neq b$, 则 $\gamma(a, b) = 1$, 否则 $\gamma(a, b) = 0$ 。

根据公式(4-12)的定义, 编辑距离是一个动态规划的问题, 算法的时间复杂度与空间复杂度均为 $O(|X| \cdot |Y|)$ 。如果只关心最后结果, 而不需要获取具体的编辑过程, 则空间复杂度可以减少为 $O(\min(|X|, |Y|))$ 。

例如将 kitten 一字转成 sitting:

sitten (k→s)

sittin (e→i)

sitting (→g)

我们将一个第一个字符串中长度为 i 的一个字串变换到第二个字符串中长度为 j 的字串, 其编辑距离定义为函数 $\text{edit}(i, j)$ 。

显然可以有如下动态规划公式:

if $i == 0$ 且 $j == 0$, $\text{edit}(i, j) = 0$

if $i == 0$ 且 $j > 0$, $\text{edit}(i, j) = j$

if $i > 0$ 且 $j == 0$, $\text{edit}(i, j) = i$

if $i \geq 1$ 且 $j \geq 1$, $\text{edit}(i, j) = \min \{ \text{edit}(i-1, j) + 1, \text{edit}(i, j-1) + 1, \text{edit}(i-1, j-1) + f(i, j) \}$,

当第一个字符串的第 i 个字符与第二个字符串的第 j 个字符不相同, $f(i, j) = 1$; 否则, $f(i, j) = 0$ 。

4.4 基于整词识别的配料表识别策略

4.4.1 分类归一化

由于配料表都是一个个词语组成，目前常用配料大概一千多个，比起单个字的识别，所包含的信息量更大，误识率更低，能够更加有效地进行识别。在单个字的归一化中，我们采用了 48×48 像素的模板，但是在整词归一化中，若依旧采用 48×48 的模板，则可能造成像素累积，当字数较高时，将变成黑块。

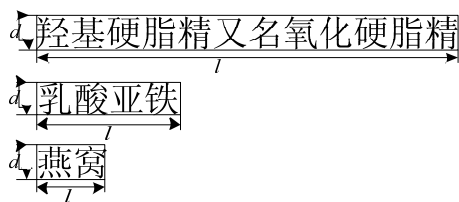


图 4-7 不同长宽比

对于图 4-7 不同长宽比，选择不同的归一化模板进行归一化。

4.4.2 分类器重新设计

根据之前的两级分类器，进行改进，以适应整词识别。将第一级粗分类改为字数信息特征，由字数的不同进行分类，进而在细分类中进行词语分类，依据配料词语特征模板集获取识别结果，见图 4-8：

假设每次识别的正确率为 90%，则整词识别策略的正确率为 $90\% \times 90\% = 81\%$ ，而当前的识别策略在平均 4 个字的配料中，识别率为 $90\% \times 90\% \times 90\% \times 90\% = 59\%$ ，识别率远低于整词识别策略。

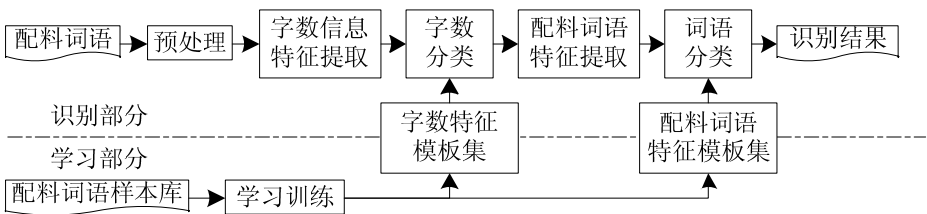


图 4-8 整词识别策略

4.4.3 整词、分字双重识别设计

在系统的实现过程中，采用了整词、分字双重识别策略，见图 4-9。如果两种识别方案结果相同，则认为识别结果正确，如果识别结果不同，反馈给客户端，供用户选择，并将结果返回服务器。通过这种方式，一方面提高识别率，另一方面，可改善用户体验。

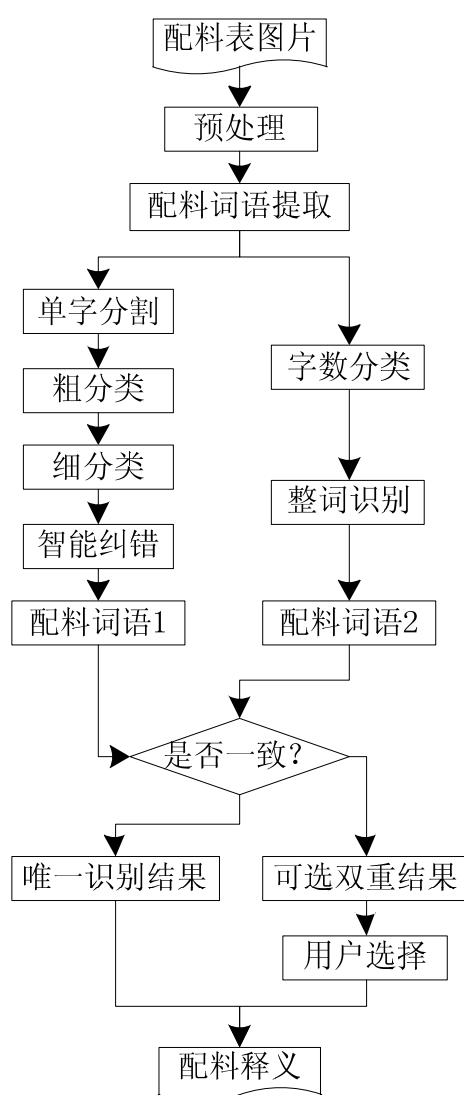


图 4-9 双重识别设计

4.5 本章小结

本章针对配料的识别进行了详细论述。内容涉及汉字特征提取、分类器的设计、智能纠错以及基于整词识别的配料表识别策略。分析了不同特征提取方法的优缺点，以便为分类器的设计提供基础。介绍了在配料表识别过程中分类器集成的较大意义，分析了其使用原因。简单地介绍了距离分类器的常用算法，给出了配料表识别策略。提出了基于编辑距离的智能纠错，改变了传统的识别系统中只识别、不对错误进行处理的做法，能够有效提高识别率。此外，提出了基于整词识别的识别策略，将一个配料词语作为一个整体进行识别，有效地提高了识别率。而整词、分字双重识别及识别反馈机制，都在可持续改进识别率上给出了很好的解决方案。

第5章 系统实现与测试

5.1 系统实现

5.1.1 算法流程

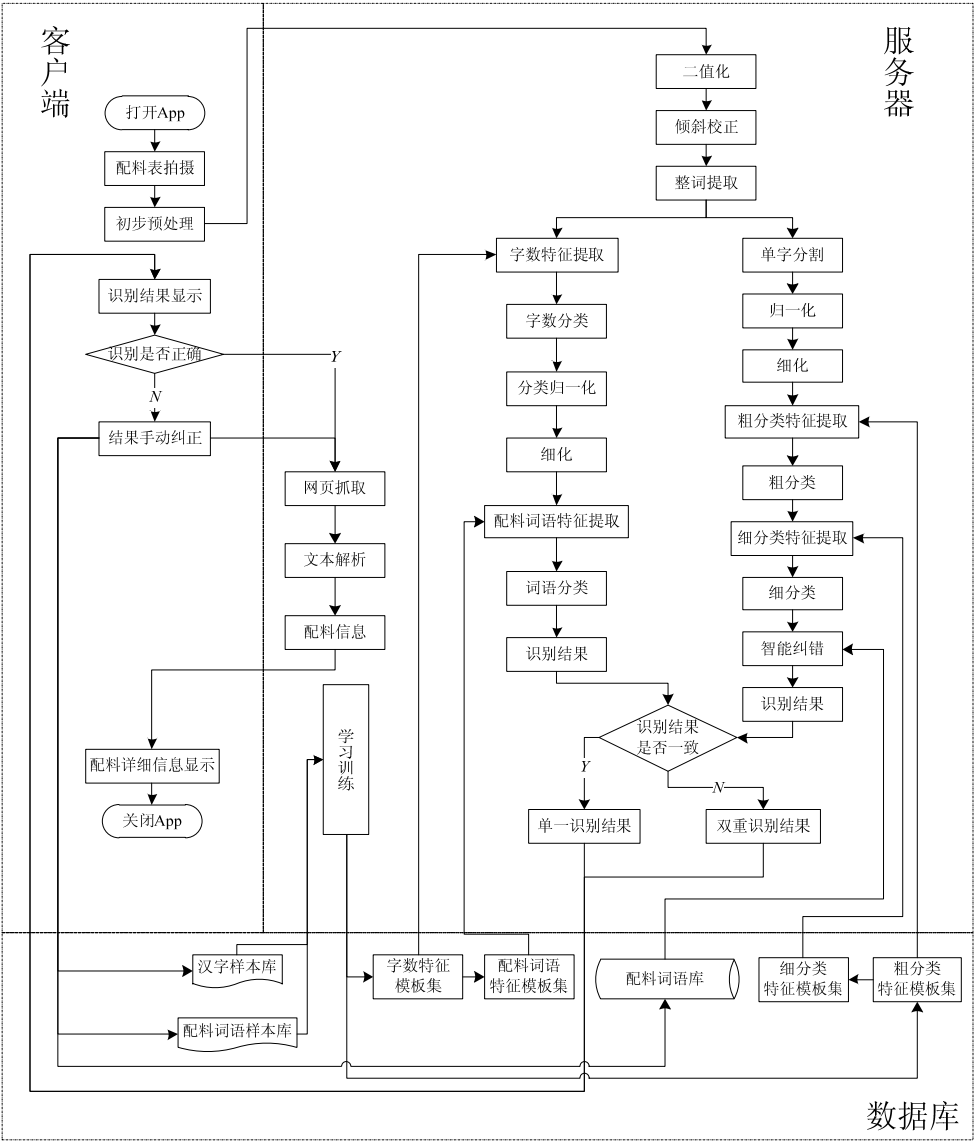


图 5-1 系统详细流程图

经过前文的研究和对比，Android 客户端完成配料表的拍摄，经过简单预处理，上传到服务器。服务器在预处理的过程中，选择文章中描述的二值化算法、基于聚类分析的字符切割算法、归一化、细化等一系列工作，获得良好的特征提取样本。之后，进行整词和单字两种识别策略进行配料词语的识别，进而通过百度百科或维基百科获取配料详细信息，将结果反馈给客户端。

图 5-1 详细描述了整个配料表识别系统的操作及处理流程。通过流程图，我们可以发现本文的几个主要工作：在 Android 客户端，限制了用户的拍摄取景及角度，解决文字识别的“Where”问题；在预处理部分，对多种二值化算法进行了对比，选取了比较适合应用场景的方案；应用本文提出的基于聚类分析的字符分割算法，有效提取整词及单字，为下一步的识别策略打下良好基础；应用智能纠错模块，对识别结果进行了纠正，进一步提高了识别率；创造性地提出了基于整词的识别策略，对配料表进行有效识别。

5.1.2 客户端界面

根据系统需求，设计具体的客户端界面如图 5-2：



图 5-2 客户端主界面

客户端为提供了拍照按钮，之后根据处理结果分别进行 OCR 识别、信息查询等操作。在拍照取景过程中，对取景框进行了限制，以方便获得有效识别区域。

5.1.3 操作步骤

根据用户需求，设计了客户端的操作步骤：

1) 配料表拍摄：



图 5-3 配料表拍摄

如图 5-3 由于在客户端限制了取景框，用户可以方便定位到配料表部分，主动过滤掉比较复杂的环境背景。

2) 识别结果：



图 5-4 识别结果

如图 5-4 图片上传到服务器后，由服务器进行预处理及识别，将结果反馈到客户端。

3) 配料详细信息:



图 5-5 配料信息显示

如图 5-5 通过搜索引擎，可以方便查询到配料详细信息。

5.2测试及对比分析

目前市场上还没有一款能够进行配料识别的专用软件，因此根据用户使用场景，选取了目前市场上应用较多的四款软件，以及传统的浏览器输入查询进行对比。

5.2.1 功能测试对比

针对独特的应用场景，对六款软件（浏览器、四款文字识别软件、本论文食品配料识别系统）操作方便度（以每次使用点击次数作为衡量）、可否进行智能识别、是否可查询配料信息等方面进行对比。对比结果如表 5-1:

表 5-1 功能测试对比

软件	点击次数(如 查询蛋白质)	是否可拍 照识别	是否可查 询信息	效果截图
UC 浏览器	13	×	✓	
图像文本识别	5	✓	×	
OCR Dialer	1	✓	×	
云脉拍照	5	✓	×	
文字识别助手	1	✓	×	
配料表查询系 统	4	✓	✓	

通过表 5-1 可知，大部分的智能识别软件都只是对文字进行识别，并没有进一步处理，如用户想要了解详细信息，还是需要复制文本，打开浏览器进行查询，造成了操作的麻烦。而直接应用浏览器查询却使操作增加，如遇到较多的字、中英混合或者生僻字，会进一步带来麻烦。而本文提出的配料表识别系统，则可通过简单的拍照操作后，获取配料详细信息，简单方便，为用户带来极大的便利。

5.2.2 性能测试对比

查询系统的性能无非就是速度和正确率，正确度取决于系统的算法，而速度则会受到硬件的影响。本文基于三星 GALAXY SIII 进行测试，硬件配置为：

- CPU 型号：三星 Exynos 4412
- CPU 频率：1433MHz 四核
- 摄像头像素：800 万像素
- 操作系统：Android OS 4.0
- RAM 容量：1GB

对系统的英文识别正确率、中文识别正确率、中英混合识别正确率、识别速度四个参数进行测试。

1) 正常情况测试

- 光线：柔和平均光照
- 配料表：平整
- 配料表背景：简单，易区分
- 样本数：1000
- 总字符数：5239

表 5-2 正常情况性能测试对比

软件	正确率				速度
	英文	中文	数字	中英混合	
UC 浏览器	手工输入	手工输入	手工输入	手工输入	>30s
图像文本识别	92%	86%	98%	80%	<10s
OCR Dialer	5%	0	6%	2%	<3s
云脉拍照	93%	95%	93%	92%	<15s

软件	正确率				速度
	英文	中文	数字	中英混合	
文字识别助手	3%	0	2%	2%	<3s
配料表查询系统	98%	97%	99%	97%	<10s

从表 5-2 中可以看到，通过对比几款软件，由于浏览器需要手工输入，故不存在正确率问题，但是代价是相对较慢；OCR Dialer 设计出来后虽然速度很快，但基本只能识别英文甚至是只能识别数字，效果很差。图像文本识别和云脉拍照均有不错的识别效果，对于中英文支持度均达到 90%以上，美中不足是识别后不能进行查询操作。本文提出的配料表查询系统中英文识别率均达到 97%，处理速度虽然较慢，但在可接受范围内。

2) 强光照射测试

- 光线：强光束照射
- 配料表：平整
- 配料表背景：简单，易区分
- 样本数：1000
- 总字符数：5304

表 5-3 强光照射性能测试对比

软件	正确率				速度
	英文	中文	数字	中英混合	
UC 浏览器	手工输入	手工输入	手工输入	手工输入	>30s
图像文本识别	85%	82%	83%	78%	<10s
OCR Dialer	5%	0	2%	2%	<3s
云脉拍照	88%	91%	90%	89%	<15s
文字识别助手	3%	0	2%	2%	<3s
配料表查询系统	97%	95%	97%	95%	<10s

由于有强光束的照射，导致配料表部分有高光情况出现，给识别结果带来影响。表 5-3 中可以发现，三款识别效果良好的系统均有不同比例下降，但配料表查询系统的结果还比较满意。

3) 昏暗光线测试

- 光线：昏暗光线照射
- 配料表：平整
- 配料表背景：简单，易区分
- 样本数：1000
- 总字符数：5240

表 5-4 昏暗光线性能测试对比

软件	正确率				速度
	英文	中文	数字	中英混合	
UC 浏览器	手工输入	手工输入	手工输入	手工输入	>30s
图像文本识别	89%	86%	87%	75%	<10s
OCR Dialer	5%	0	2%	2%	<3s
云脉拍照	88%	87%	90%	83%	<15s
文字识别助手	3%	0	2%	2%	<3s
配料表查询系统	94%	91%	95%	91%	<10s

在昏暗光线下，由表 5-4 中发现，还是出现了不同程度的误识情况，但也都保持在 80%以上，配料表查询系统基本保持在 90%以上，效果比较良好。

4) 褶皱测试

- 光线：柔和平均光照
- 配料表：褶皱曲面
- 配料表背景：简单，易区分
- 样本数：1000
- 总字符数：5001

表 5-5 曲面性能测试对比

软件	正确率				速度
	英文	中文	数字	中英混合	
UC 浏览器	手工输入	手工输入	手工输入	手工输入	>30s
图像文本识别	82%	76%	80%	72%	<10s

软件	正确率				速度
	英文	中文	数字	中英混合	
OCR Dialer	5%	0	2%	2%	<3s
云脉拍照	80%	85%	82%	82%	<15s
文字识别助手	3%	0	2%	2%	<3s
配料表查询系统	88%	87%	87%	87%	<10s

在圆柱型包装，或者出现褶皱的包装上面，由于字符变形，使得识别效果出现了较大的退步。见表 5-5，基本所有的系统识别率都低到了 90%以下，若想有较好的识别效果，还是需要使用者对被拍配料表进行一定人工整理。

5) 复杂背景测试

- 光线：柔和平均光照
- 配料表：圆弧曲面
- 配料表背景：简单，易区分
- 样本数：1000
- 总字符数：4987

表 5-6 复杂背景性能测试对比

软件	正确率				速度
	英文	中文	数字	中英混合	
UC 浏览器	手工输入	手工输入	手工输入	手工输入	>30s
图像文本识别	87%	82%	84%	83%	<10s
OCR Dialer	5%	0	2%	2%	<3s
云脉拍照	83%	81%	84%	81%	<15s
文字识别助手	3%	0	2%	2%	<3s
配料表查询系统	93%	92%	95%	91%	<10s

在复杂背景下，三款比较优秀的系统还是能够进行不错的识别，见表 5-6。而配料表查询系统依然可以保持识别率在 90%以上，有很好的用户体验。

5.3本章小结

对系统整体算法流程进行了介绍，与现有系统进行了功能和性能测试，结果表明，本文提出的基于 **Android** 的食品配料表识别系统对于食品配料表的信息查询无论是操作方便性、处理速度、识别正确率均很高，具有很好的现实意义。

第6章 总结与展望

6.1 本文工作总结

本文从总体上介绍了食品配料表的识别过程。从图片拍摄到识别的主要环节均进行了阐述。对食品配料表图像的预处理进行了介绍，提出了不同的分类器集成方案，对主要的原理进行了详细阐述，并进行了实践。本文的主要工作如下：

- 1) 鉴于配料表基本由多个配料词语组成的特殊形式，提出了基于聚类分析的版面切分，通过简单聚类，找出不同词语间距的阈值，进行切割，效果良好。
- 2) 基于编辑距离对识别出的配料词语进行了智能纠错，有效地弥补了识别率无法达到 100% 的缺陷，进一步提高了系统的识别率。
- 3) 提出了反馈机制，如果识别结果有误，用户可以手动输入结果，系统自动捕捉并将其纳入词典，并对新样本加入样本集重新进行训练，以期达到更好的识别效果。
- 4) 提出了基于整词的识别策略。通过搜集与统计，发现常用的配料词语大概一千多个，因此提出了基于词语的识别策略，较传统的基于单字的识别，所携带的信息更多，特征更易于提取，更有效抑制了由于形近字造成的错识率，达到较好识别效果。

6.2 展望

经过几十年的发展，印刷体文字识别技术已经比较成熟，市场上也出现了不少 OCR 识别系统。虽然目前的 OCR 系统能够达到满意的识别率和速度，但是依旧存在很强的局限性，如多语言处理效果不好，不能进行实时详细信息查询等。

随着智能手机的广泛应用，基于手机客户端的 OCR 识别，尤其是专业领域识别将越来越普遍，对于配料表的识别，也还存在很大的发展空间。

- 1) 因拍摄角度或者强光束的干扰，使得拍摄的食品配料表图像质量较差，

不能进行良好的识别。尤其很多包装材质为塑料，更容易反光，给识别造成干扰。因此，有待于提出更好的预处理算法来有效解决这个难题。

2) 文中的系统，文字识别部分主要在服务器端完成，需要有网络的支持。但是有些手机用户对流量使用比较保守，不愿耗费流量传输数据。希望随着手机硬件的发展，并改进原有算法，提高处理速度，降低资源消耗，以期能够在手机端完成识别任务。

参考文献

- [1] 王明强,陈顺浩,浦绍飞.预防和控制食品添加剂对食品安全的影响及防止对策[J].中国调味品,2012(4):19-24+28.
- [2] 钟凯,韩蕃璠,姚魁,等.中国食品安全风险交流的现状、问题、挑战与对策[J].中国食品卫生杂志,2012,24(6):578-586.
- [3] Jasminka GIACOMETTI,Djuro JOSIC.Foodomics in microbial safety[J].Trends in Analytical Chemistry,2013,52(3):1024-1029.
- [4] Tomoki TATEFUJI,Miyako YANAGIHARA,Shinobu FUKUSHIMA.Safety assessment of melinjo (Gnetum gnemon L.) seed extract: Acute and subchronic toxicity studies[J].Food and Chemical Toxicology,2014(4):2341-2347.
- [5] 卢斌斌,姚玮华.食品添加剂的安全管理[J].中国调味品,2011(11):4-8.
- [6] 承明华,张海波.试论完善我国食品安全监管工作的对策与出路[J].中国食品卫生杂志,2013(5):60-66.
- [7] Brian WYNNE.Social Identities and Public Uptake of Science[J].Radioactivity in the Environment,2013,19(2):109-115.
- [8] 齐敏,李大健,郝重阳.模式识别导论[M].北京:清华大学出版社,2009.
- [9] 郭军,马跃,盛立东,等.发展中的文字识别理论与技术[J].电子学报,1995,23(10):184-186.
- [10] Maurice MONIQUE,Gioanni HENRI,Abourachid ANICK.Influence of the Behavioural Context on the Optocollic Reflex (ocr) in Pigeons (columba Livia).[J].The Journal of Experimental Biology,2006,209(Pt 2):1345-1350.
- [11] Farjana Yeasmin OMEE,Shiam Shabbir HIMEL,Abu Naser BIKAS.A Complete Workflow for Development of Bangla Ocr[J].International Journal of Computer Applications,2011,21(9):1024-1030.
- [12] 罗军舟,吴文甲,杨明.移动互联网[J].计算机学报,2011(11):5-27.
- [13] 何耘娴.印刷体文档图像的中文字符识别[D].秦皇岛:燕山大学,2011.
- [14] 李俊.印刷体文字识别系统的研究与实现[D].成都:电子科技大学,2011.
- [15] Yushuang TIAN,Kim-Hui YAP,Yu HE.Vehicle license plate super-resolution using soft learning prior[J].Multimedia Tools and Applications,2012,60(3):2340-2347.
- [16] Miriam SCHMIDT,Günther PALM,Friedhelm SCHWENKER.Spectral graph features for the classification of graphs and graph sequences[J].Computational Statistics,2014,29(1):1340-1349.

- [17] Kaushik DEB,Ibrahim KHAN,Anik SAHA,等.An Efficient Method of Vehicle License Plate Recognition Based on Sliding Concentric Windows and Artificial Neural Network[J].Procedia Technology,2012,4(1):1326-1331.
- [18] Tobias BLANKE,Michael BRYANT,Mark HEDGES.Open source optical character recognition for historical research[J].Journal of Documentation,2012,68(5):1570-1573.
- [19] 王科俊,冯伟兴.中文印刷体文档识别技术[M].北京:科学出版社,2010.
- [20] Davide Di RUSCIO,Patrizio PELLICCIONE.Simulating Upgrades of Complex Systems: the Case of Free and Open Source Software[J].Information and Software Technology,2014(3):1230-1237.
- [21] Nishiyama K,Nakaseko M,Hosokawa M. comparison of Muscular Load Between Normal Handwriting and Constrained Writing Forced By Optical Character Reader (ocr) Cards[J].Nihon Eiseigaku Zasshi. Japanese Journal of Hygiene,1990,44(6):1572-1578.
- [22] GB2312-80.信息交换用汉字编码字符集基本集[S].
- [23] GB2760-2001.食品安全国家标准食品添加剂使用标准[S].
- [24] GB7718-2011.食品安全国家标准预包装食品标签通则[S].
- [25] 谭同德,王三刚.基于 OpenCV 的车牌定位方法[J].计算机工程与设计,2013,34(8):2816-2820.
- [26] 陈利华,董志学.基于 Android 的裂缝宽度检测系统设计实现[J].计算机工程与设计,2013,34(9):3195-3199.
- [27] Zheran FANG,Weili HAN,Yingjiu LI.Permission Based Android Security: Issues and Countermeasures[J].Computers & Security,2014(2):1450-1457.
- [28] Christoforos NTANTOGIAN,Dimitris APOSTOLOPOULOS,Giannis MARINAKIS, et al.Evaluating the Privacy of Android Mobile Applications Under Forensic Analysis[J].Computers & Security,2014(2):2400-2404.
- [29] G. CALARCO,M. CASONI.On the Effectiveness of Linux Containers for Network Virtualization[J].Simulation Modelling Practice and Theory,2013,31(2):790-795.
- [30] A. K. PANDEY,Pankaj BHARGAVA.Effects of harvesting intensities and techniques on re-growth dynamics and quality of Terminalia bellerica fruits in central India[J].Journal of Forestry Research,2014,25(1):981-987.
- [31] 曾凡锋,高艳云,付晓玲.文本图像的去噪算法应用研究[J].计算机工程与设计,2012(7):189-193.
- [32] Nazif DEMOLI,Iva MRĚELA,Kristina ŠARIRI.Correlation and image moment approaches to analyze the Glagolitic script carved in stone tablets[J].Optik - International Journal for Light and Electron Optics,2012(2):1029-1035.
- [33] Antonio FERNÁNDEZ-CABALLERO,María T. LÓPEZ,José Carlos CASTILLO.Display text

segmentation after learning best-fitted OCR binarization parameters[J].Expert Systems With Applications,2011,39(4):555-559.

- [34] 刘旭,巫玲,陈念年,等.基于光栅投影序列图像融合的倾斜校正算法[J].计算机应用,2013,33(11):3209-3212.
- [35] W.A.J.P. WIJESINGHE,Eun-A KIM,Min-Cheol KANG.Assessment of anti-inflammatory effect of 5 β -hydroxypalisadin B isolated from red seaweed *Laurencia snackeyi* in zebrafish embryo in vivo model[J].Environmental Toxicology and Pharmacology,2014,37(1):549-558.
- [36] 刘赛,王江晴,张振绘.一种用于脱机手写体女书字符切分的方法[J].计算机应用研究,2011,28(3):1187-1190.
- [37] 王琰滨,蒋龙泉,冯瑞.一种低图像质量车辆牌照的字符分割方法[J].计算机应用与软件,2013,30(3):108-110+117.
- [38] 刘昱.印刷体表格识别的研究[D].哈尔滨工程大学,2013.
- [39] Morteza ZAHEDI,Saeideh ESLAMI.Farsi/Arabic optical font recognition using SIFT features[J].Procedia Computer Science,2011,3(1):1030-1035.
- [40] C.K.H. LEE,K.L. CHOY,Y.N. CHAN.A knowledge-based ingredient formulation system for chemical product development in the personal care industry[J].Computers and Chemical Engineering,2014(1):734-739.
- [41] SungHoo CHOI,Jong Pil YUN,Keunhwi KOO.Localizing slab identification numbers in factory scene images[J].Expert Systems With Applications,2012,39(9):109-116.
- [42] 张建萍,刘希玉.基于聚类分析的 K-means 算法研究及应用[J].计算机应用研究,2007(5):94-102.
- [43] 万金娥,袁保社,谷朝,等.基于字符归一化双投影互相关性匹配识别算法[J].计算机应用,2013,33(3):49-51+120.
- [44] Fadoua DRIRA, Frank LEBOURGEOIS, Hubert EMPTOZ. A new PDE-based approach for singularity-preserving regularization: application to degraded characters restoration[J].International Journal on Document Analysis and Recognition (IJDAR),2012,15(3):237-245.
- [45] 李钦瑞,都云程,刘坤,等.基于模板匹配及曲线拟合的视频字幕细化研究[J].计算机应用与软件,2014,31(1):144-147.
- [46] 任俊玲.脱机手写汉字识别若干关键技术研究[M].北京:北京邮电大学出版社,2013.
- [47] 黄襄念,程萍.文字识别原理与策略[M].成都:西南交通大学出版社,2002.
- [48] 刘昊,方雯逸.基于 BP 神经网络的人脸朝向分类的新思路[J].计算机科学,2012,39(11):366-368+374.
- [49] 付涛.基于高阶神经网络的文字识别算法研究[D].长春:东北师范大学,2010.

- [50] 袁哲.人工智能在拼音输入法中的应用[J].软件导刊,2010(6):12-14.
- [51] 王永景.面向文本识别流的自动校对算法研究[D].上海:上海交通大学,2008.
- [52] Josefa GARCÍA-ROMERO,Rafael GINÉS,Ruth VARGAS.Marine and freshwater crab meals in diets for red porgy (*Pagrus pagrus*): Digestibility, ammonia-N excretion, phosphorous and calcium retention[J].Aquaculture,2014(1):1049-1054.

致谢

转眼间，在武汉理工大学已经度过了我的第七个年头，也即将走向工作岗位，告别学生生涯。

首先，要感谢我的导师肖攸安教授。肖老师在治学态度上极其严谨，对我们的要求也非常严格，尤其是他渊博的知识体系是我奋斗的目标。正是在肖老师的督促和帮助下，我才能够顺利完成毕业论文。

其次，我要感谢我的父母在这七年中给我的默默支持，是他们的无私奉献和谆谆教导支撑着我走过这七年最美好的年华。还要感谢我的女朋友李梦玉，她对我学业的监督和帮助也让我进步良多。

最后，要感谢我的小伙伴老焦、王伟、大胖、龚老师、涵老师、谢大神、大黄、小菜鸟、二姐等，正是由于有了你们，我的研究生生活才多姿多彩。

谢谢大家！

攻读学位期间获得的科研成果

- [1] QIAO Shuang, XIAO You-an. Data Analysis of Aircraft Take-off Performance[C]//2012 4th Electronic System-Integration Technology Conference,2012:684-687.