****

**软件系统设计说明书**

题 目： 网店工商信息图片文字提取系统

学 院： 计算机科学与工程学院

专 业： 计算机科学与技术

班 级： 2018 级卓越工程师班

学生姓名： 崔梦涛

癿盼盼

杨蕊媛

杨钦颖

指导教师： 代祖华

摘 要

图片中的文本为描述图片内容提供了十分有用的信息,对于构建基于内容的多媒体检索系统具有重要作用。因此,有效地提取和识别这些文字对于图像理解、图片内容分析、基于内容的图像和图片检索等领域具有重要意义。目前商业的OCR技术对于二值图像中文字的提取识别已经趋于成熟,但图片中的文字具有分辨率低、背景复杂、文字形态差异大等特点,这给图片中文字的有效提取带来了极大的挑战,限制了OCR技术的成功应用。

许多图像都包含丰富的文字信息 ,如用作网页设计的以图像形式存在的标语中的字幕。这些文字的自动检测、分割、提取和识别 ,对图像高层语义内容的自动理解、索引和检索非常有价值 ,因此引起国内外众多学者的研究兴趣。为使人们对该领域有一个系统的了解 ,并使该领域研究人员有所借鉴 ,在对目前国内外图像和图片中文字获取技术相关文献综合理解的基础上 ,综述了该领域的发展现状 ,同时从文字检测、抽取和文字识别两个方面 ,重点讨论了其主要的技术方法及应用优缺点 ,并结合当前面临的问题 ,指出今后可进一步研究的方向。

**关键词：**图片；信息提取；EXCEL

目 录

[第一章 引言 1](#_Toc12100)

[第二章 任务概述 3](#_Toc22627)

[2.1 目标 3](#_Toc18946)

[2.2 运行环境 3](#_Toc15775)

[2.3 项目主要的技术方案 3](#_Toc6926)

[2.4 可行性分析 3](#_Toc25823)

[第三章 系统总体结构 5](#_Toc65)

[3.1 系统功能描述 5](#_Toc9736)

[3.2 总体结构设计 5](#_Toc26008)

[3.2.1 模块设计 5](#_Toc2603)

[3.2.2 系统流程图 7](#_Toc20700)

[3.3 接口设计 7](#_Toc1275)

[3.3.1 用户人机接口设计 7](#_Toc8123)

[3.3.2 用户界面设计 8](#_Toc31501)

[第四章 系统结构设计 14](#_Toc13202)

[4.1 软件系统数据库逻辑结构 14](#_Toc771)

[4.1.1 实体和属性的定义 14](#_Toc24679)

[4.1.2 实体之间的联系 14](#_Toc263)

[4.1.3 全局E-R模式 14](#_Toc24815)

[4.1.4 通过Power Designer生成的概念模型CDM 16](#_Toc30544)

[4.2 设计软件系统总体结构 17](#_Toc24469)

[4.3 软件的重用方案 18](#_Toc31612)

[4.4 设计关键类的重点服务 19](#_Toc25816)

[第五章 运用设计 21](#_Toc1)

[5.1 数据字典设计 21](#_Toc15740)

[5.2 安全保密设计 21](#_Toc23059)

[5.2.1 系统安全设计 21](#_Toc16228)

[5.2.2 操作系统平台的安全管理 22](#_Toc28579)

[5.2.3 数据库系统的安全管理 22](#_Toc18952)

[参考文献 24](#_Toc5729)

# 第一章 引言

1.1 编写目的

从图像中提取文字属于信息智能化处理的前沿课题,是当前人工智能与模式识别领域中的研究热点。由于文字具有高级语义特征,对图片内容的理解、索引、检索具有重要作用,因此,研究图片文字提取具有重要的实际意义。又由于静态图像文字提取是动态图像文字提取的基础,故着重介绍了静态图像文字提取技术,总结了几种已提出的算法,并利用计算机语言学方法对提取出的文字进行后期处理,大大提高了文字提取的正确率。

根据国家工商总局《网络交易管理办法》要求对网店营业执照信息进行公示，天猫网店经营者营业执照信息会在天猫店铺上以图片形式进行公示，但图片信息不能进行结构化处理，需要提取出图片中的企业注册号、企业名称形成结构化文档。

如果需要统计某家企业在天猫平台的网店经营情况，就需要先把天猫平台的企业信息采集下来进行结构化处理。参赛者可以用附件提供的天猫平台上采集的50家企业图片作为测试数据，提取出文字信息后汇总进Excel作为交付文件。

1.2 背景

随着计算机技术、多媒体技术和通信技术的飞速发展，数据库中和网页上的数字化图像和视频呈现爆炸性增长，基于数字化图像和视频的网络检索正在迅速成为信息传递与服务的主流。如何从Internet 上准确迅速地找到用户所需的内容成为日益突出和紧迫的需求。同时，现代的图书馆为了满足用户对多媒体内容的查询，也开始收藏图像和视频等资料，因而也需要有效的方法组织和检索这些多媒体内容。另外，传统的基于关键词的检索方式，一方面由于信息量急剧膨胀使得提取关键词变得繁琐而低效;另一方面，单纯依靠关键词的检索结果不完整，不能满足用户的需要，因而基于内容的检索技术应运而生并取代之成为主流。

图像和视频中的文字是帮助理解图像和视频内容高层语义的重要线索。大量的网页图像例如新闻图片、广告画等都通常在其内部嵌以简短的文字来概括所描述的事件和对象，以帮助用户快速理解。如果这些文字能够自动地被检测、分割和识别出来，就可以让机器来自动理解图像的内容并对图像进行分类，这必定会对图像高层语义的自动理解、索引和检索提供很大的帮助。例如，在电子图书封皮中检测到的标题文字，可以为图书的归档节约大量的人力;在新闻视频中检测到的标题文字，可以为新闻事件内容的理解提供直接语义特征。

# 第二章 任务概述

## 2.1 目标

根据国家工商总局《网络交易管理办法》要求对网店营业执照信息进行公示，天猫网店经营者营业执照信息会在天猫店铺上以图片形式进行公示，但图片信息不能进行结构化处理，需要提取出图片中的企业注册号、企业名称形成结构化文档。

如果需要统计某家企业在天猫平台的网店经营情况，就需要先把天猫平台的企业信息采集下来进行结构化处理。参赛者可以用附件提供的天猫平台上采集的50家企业图片作为测试数据，提取出文字信息后汇总进Excel作为交付文件。

## 2.2 运行环境

Web 浏览器：0+、Chrome、Opera、Safari、Firefox及任何支持HTML5标准的浏览器。  
 标准分辨率：1024768、19201080、2K。

## 2.3 项目主要的技术方案

通过对系统功能需求的分析，项目涉及到的主要技术有：

1、利用Python平台上的图像识别标准库对图片进行处理，提取图片中的文字信息。

2、文字识别算法采用光学字符识别(OCR,Optical Character Recognition)。光学字符识别是对文本资料进行扫描，然后对图像文件进行分析处理，获取文字及版面信息的过程。

3、利用图片格式转换工具转换图片格式，转换完成之后，再进行图片文字识别。

4、对图片进行模糊寻找，给出图片文件夹的大致路径，然后程序能够找到正确的图片路径，并能够正确进行文字识别。

5、利用Python中的模式匹配技术从大量文字信息中提取出有效信息。

6、调用Python的openpyxl第三方库，通过add\_sheet()方法中的write()函数将提取出的有效数据写入到excel中，然后使用save()函数保存excel文件。

### 2.4 可行性分析

1、投资必要性

根据目前的市场调查及分析预测的结果，以及有关的产业政策的因素论证项目投资建设的必要性。该项目具有实际用途，可以方便用户对企业工商信息进行汇总管理，具有投资价值。

2、技术可行性

开发环境操作系统:WINDOWS或LINUX。

开发语言:（C＋＋，JAVA，Python）。

3、财务可行性

主要从项目及投资者的角度，设计合理财务方案，从企业理财的角度进行资本预算，评价项目的财务盈利能力，进行投资决策。

4、组织可行性

制定合理的项目实施计划，设立合理组织机构，选择经验丰富的管理人员，建立良好的协作关系，制定合适的培训计划，能保证项目顺利执行。

5、经济可行性

该项目非大型项目，经济支出消耗较少，较少的投资可以获得较高的回报。

该项目在实现区域经济发展目标，有效配置经济资源，增加就业，提高人民生活等方面效益很大。

6、风险因素

主要是对项目的市场风险，技术风险，财务风险，组织风险，法律风险，经济及社会风险等因素进行了评价，制定了规避风险对策，为项目全过程的风险管理提供了依据。

其中，技术风险主要是现今社会知识更新和科技发展的加速，使得新技术的生命周期缩短，竞争者的技术不断改进，这会对我们开发出的现有产品形成冲击，并且市场上可能出现许多替代的产品，不能保证我们的产品在市场中一直具有竞争优势。应对策略是能够经常发布更新以适应各种新的要求，防止被其他竞争者抢占先机。

财务风险主要是现阶段没有专业的财务人员，资金回收策略不当以及在筹资方面资金结构不合理会导致出现财务风险。应对策略是请教专业的财务人员，掌握相关知识，制定合理的资金方案。

# 第三章 系统总体结构

## 3.1系统功能描述

功能需求：

1、程序能够自动读取企业工商信息图片所在的文件夹路径。

2、从图片文件夹路径中顺序取出图片进行识别。

3、因为天猫平台公示的图片内容没有固定格式，所以需要程序能匹配不同格式的图片内容提取信息。

4、能够提取出图片中的企业注册号、企业名称数据项，并保存进Excel中，因为企业注册号、企业名称数据项要进行分析处理，所以需要保证提取信息的准确性，识别准确率需要保证在95%以上。

5、最终的识别结果以一份汇总的Excel交付。

非功能需求：

1、识别速度保持在60秒识别50张图片。

## 3.2总体结构设计

## 3.2.1 模块设计

综合考虑系统的需求之后，系统的结构化设计如图3.2.1。

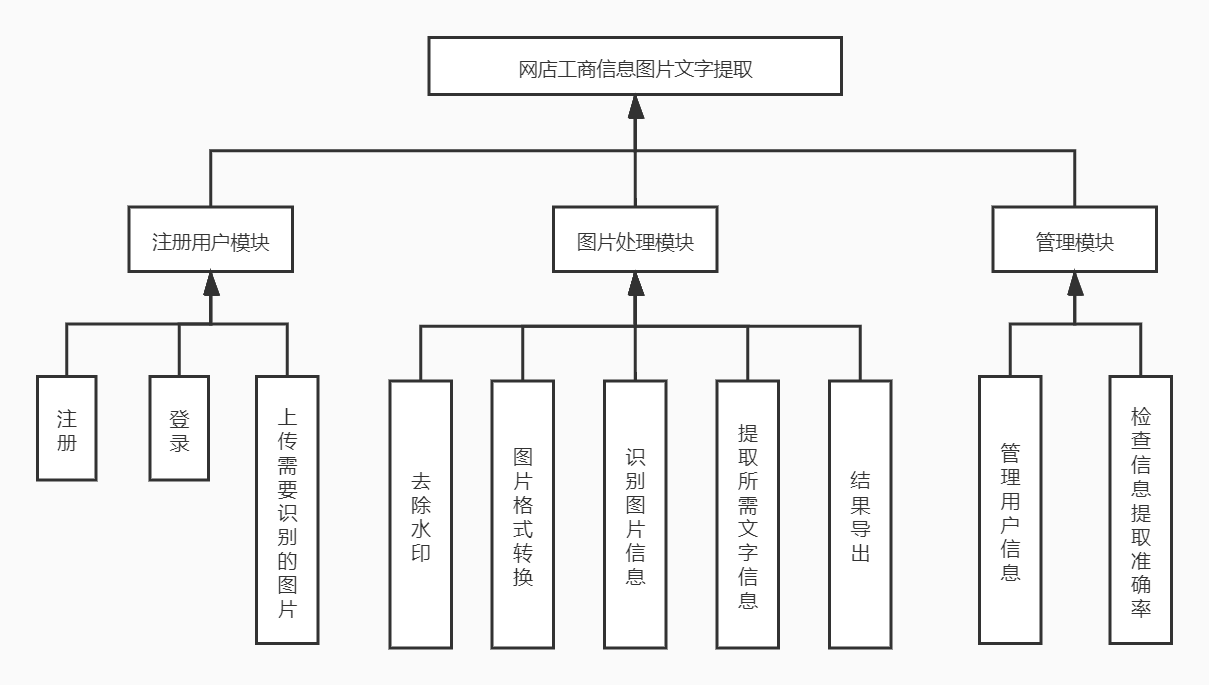


图3.2.1 结构化设计

各模块功能：

1、注册用户模块：

功能有注册、登录、上传需要识别的图片。

其中注册需要手机号、密码和短信验证码；登录需要手机号和密码，用户上传需要识别的图片之后，系统执行图片处理模块。

2、图片处理模块：

功能有去除水印、图片格式转换、识别图片信息、提取所需文字信息、结果导出。

其中前四个功能模块在用户上传完图片后系统自动执行，结果导出模块可由用户指定导出路径。

3、管理模块：

功能有管理用户信息、检查信息提取准确率。

其中管理用户信息模块可执行对用户信息的添加、修改、查询、删除功能。

## 3.2.2 系统流程图

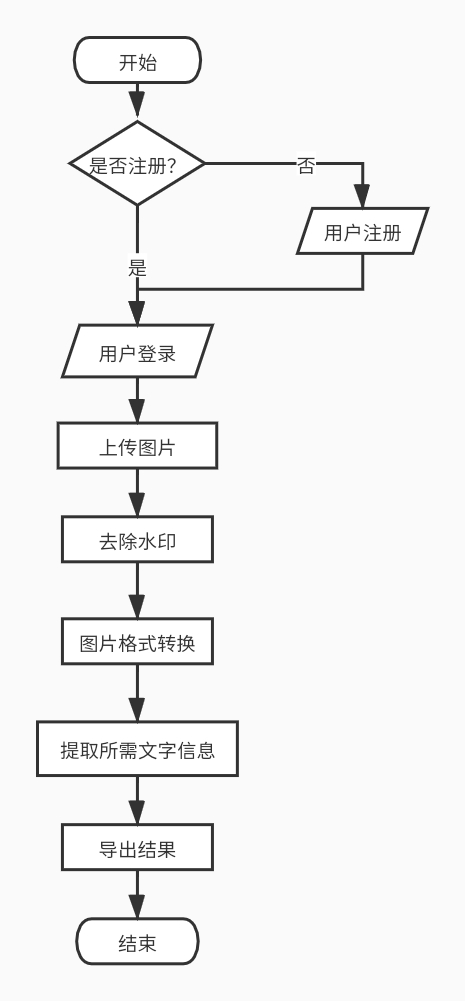


图3.2.2 系统流程图

## 3.3接口设计

## 3.3.1 用户人机接口设计

1.用户模型

用户模型（User Mode）是用户接口设计的基础，提供给用户有关所需处理的对象以及作用于这些对象的处理过程的一个概念性模型。

2.显示屏幕的有效利用

　　为使屏幕得到有效利用，软件设计中应考虑如下几个问题：

　　（1）信息显示的布局合理性。

　　 （2）充分而又正确地使用图符（应用图符（application icons）、控制图符（control icons））。

　　（3）恰当地使用各种表示方法进行选择性信息显示。

3.反馈

　　反馈就是动态地显示系统进行中所发生的一些变化，以便更有效地进行交互作用。

4.设计的一致性原则

　　一致性原则是指在设计系统的各个环节中，应遵从统一、简单的规则，保证部出现例外和特殊情况。还应考虑按照用户认为最正常、最合乎逻辑的方式进行设计，减少用户记忆量。

5.回退和出错处理

　　回退（undo）机制包括取消机制、确认机制等。对可能导致错误的一些动作进行预测，设计相应的诊断程序，提供出错消息等。

6.约束机制

　　动作与对象相一致。

7.联机帮助

为用户提供联机帮助（On-Line Help）措施，能在用户操作过程中的任何时刻提供请求帮助。适应不同的用户，提供多种方法使软件能适应不同熟练程度的用户。

## 3.3.2 用户界面设计



图3.3.1 登录界面



图3.3.2 注册界面

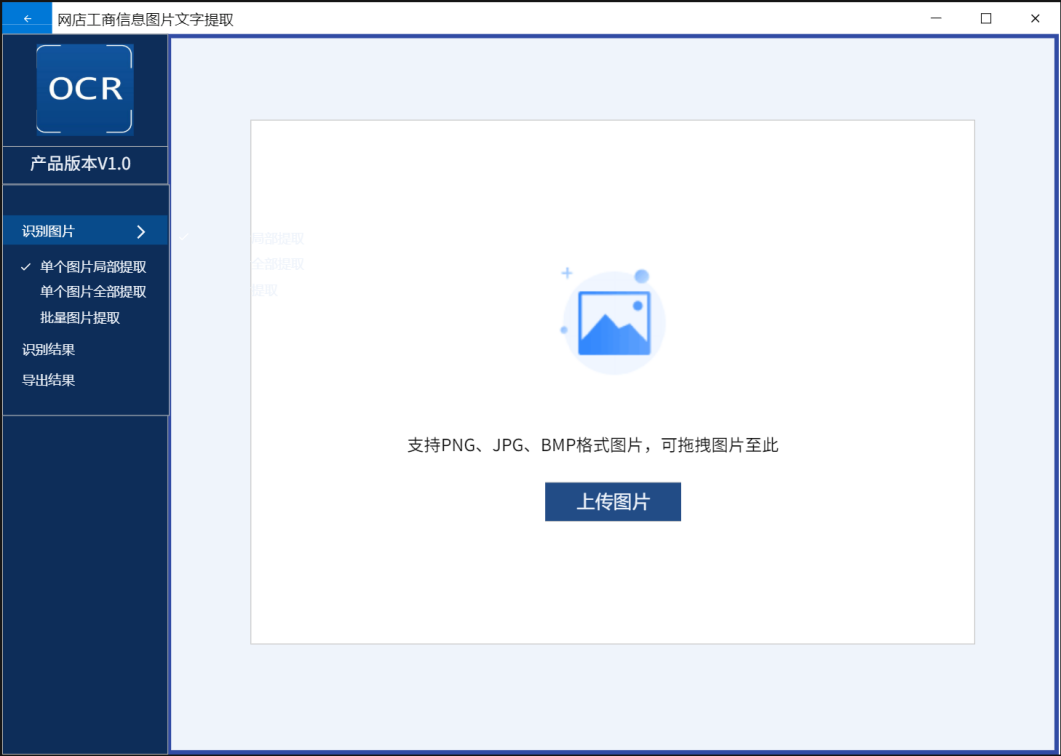


图3.3.3 单个图片上传界面

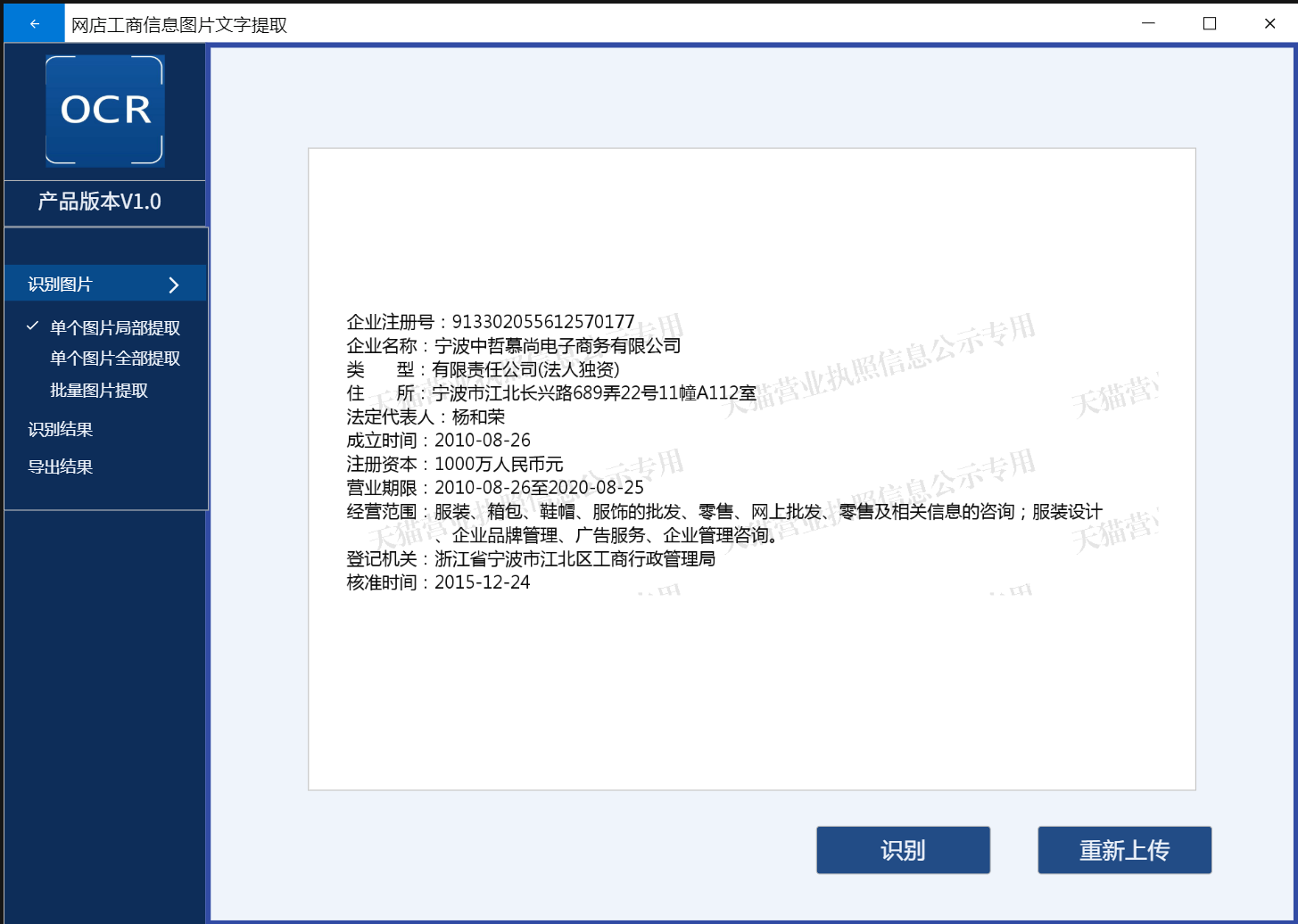


图3.3.4 单个图片上传成功界面

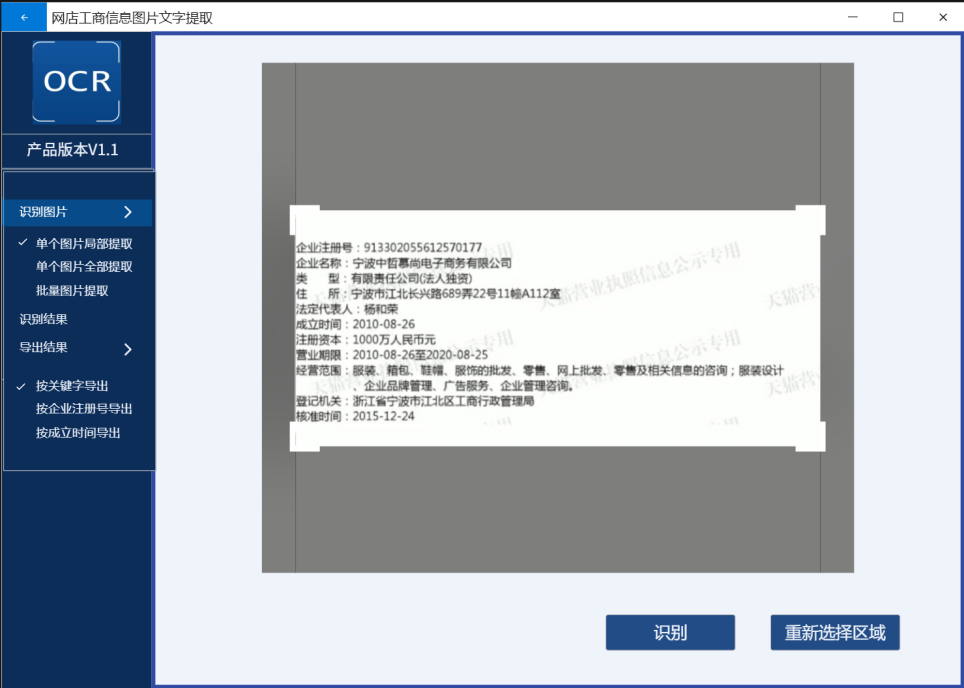


图3.3.5 单个图片上传识别信息局部提取界面

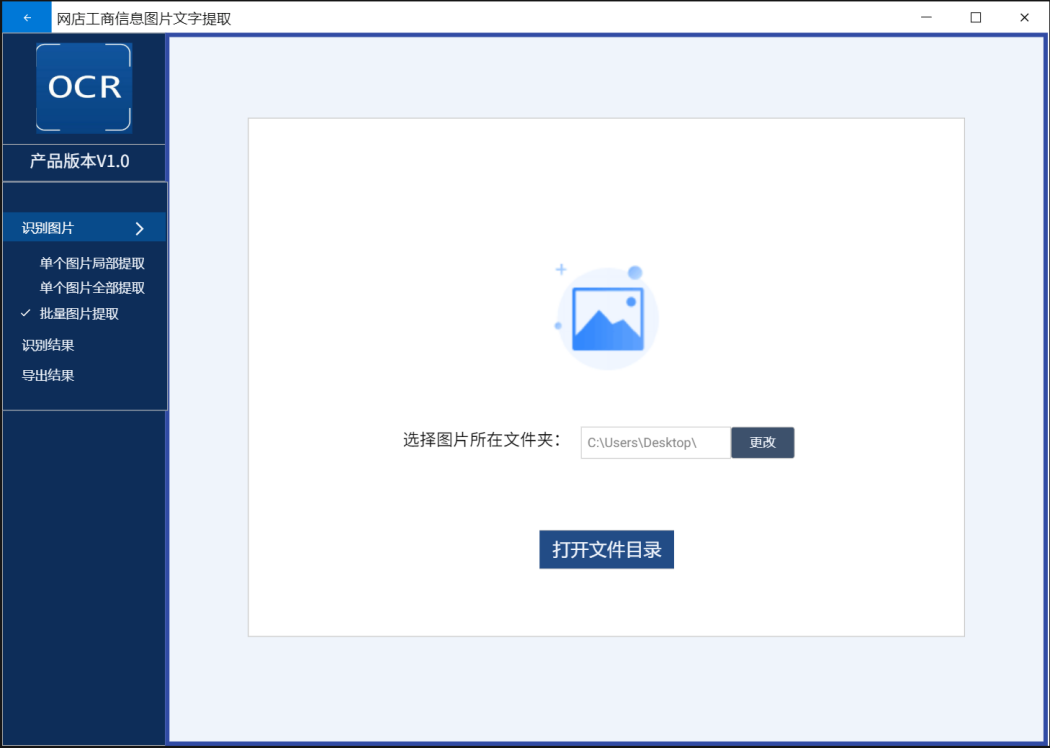


图3.3.6 批量图片提取上传界面

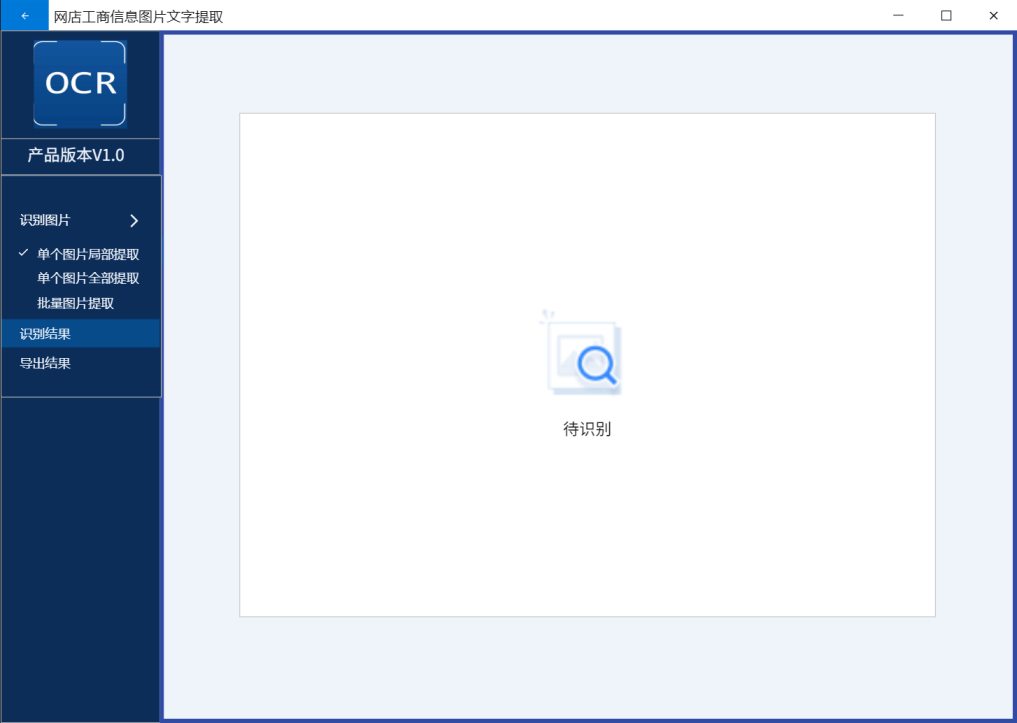


图3.3.7 未选择图片时识别结果界面



图3.3.8 识别成功界面



图3.3.9 按关键字导出界面



图3.3.10 按企业注册号导出界面



图3.3.11 按成立时间导出界面

# 第四章 系统结构设计

## 4.1软件系统数据库逻辑结构

## 4.1.1 实体和属性的定义

用户（手机号，密码）

图片（图片编号，图片）

企业（企业注册号，企业名称）

企业信息（企业注册号，类型，住所，法定代表人，注册时间，注册资本，营业期限，经营范围，登记机关，核准时间）

管理员（账号，密码）

## 4.1.2 实体之间的联系

用户与图片：

一个用户可以上传多张图片，而一张图片只能被一个用户上传，因此用户和图片之间是一对多的关系。

图片和企业：

一张图片上可提取出一个企业，一个企业只由一张图片来显示，因此图片和企业之间是一对一的关系。

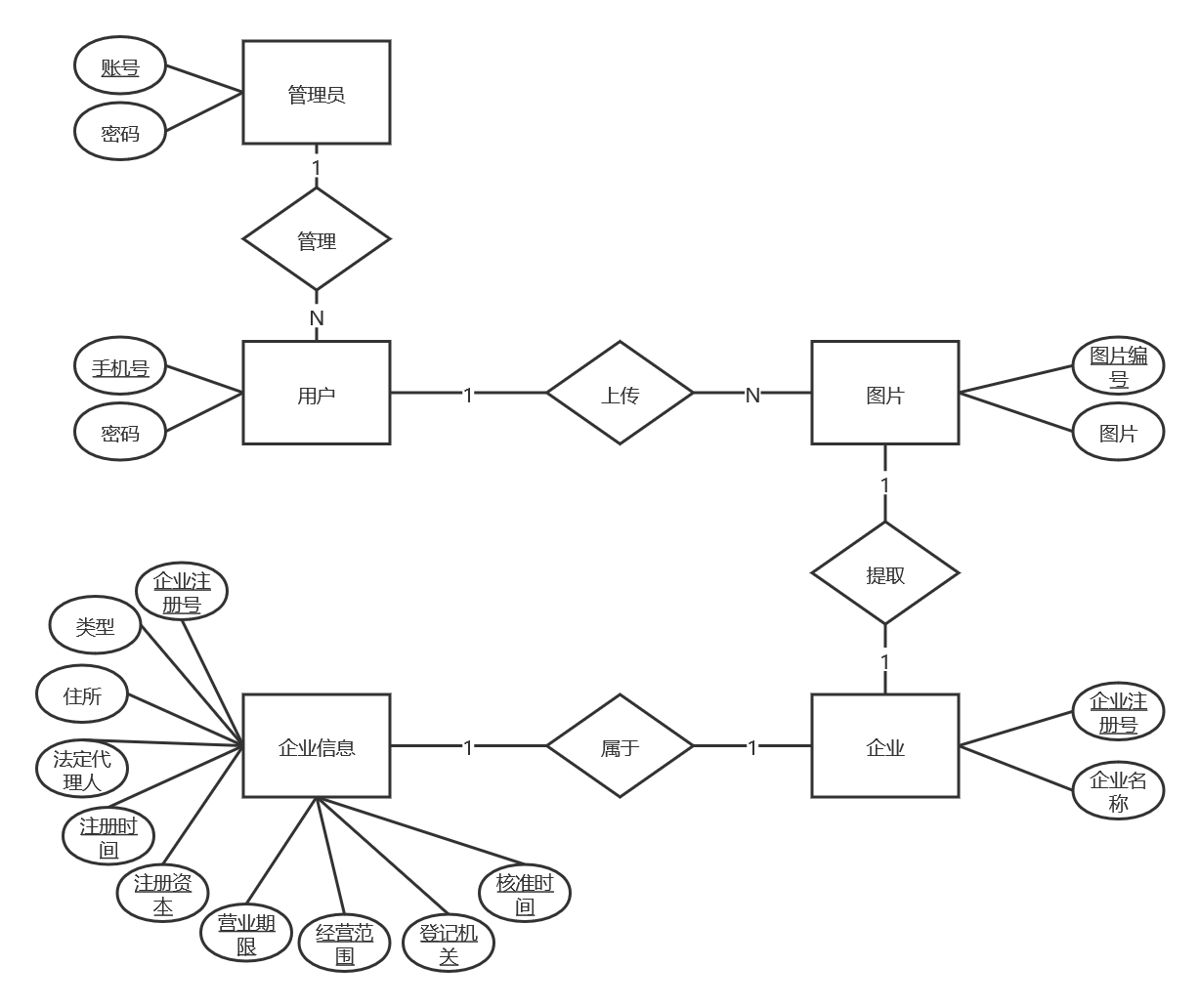
企业和企业信息：

一个企业信息只属于一个企业，一个企业只提供一个企业信息，因此企业和企业信息之间是一对一的关系。

管理员和用户：一个管理员可以管理多个用户，一个用户可以被多个管理员管理，因此管理员和企业之间是一对多的关系。

## 4.1.3 全局E-R模式

数据库的逻辑结构设计就是把概念结构设计阶段设计好的基本实体-关系图转换为与选用的数据库管理系统产品所支持的数据模型相符合的逻辑结构。所以需要先设计实体-关系图，本项目的全局E-R模型如图所示。

图4.1.3 E-R图

User(Uphone,Upassword)

用户（手机号，密码）

Picture(Pid,Ppicture)

图片（图片编号，图片）

Company(Cid,Cname)

企业（企业注册号，企业名称）

Information(Iid,Itype,Iaddress,Ipeople,Iregistertime,Imoney,Ideadline,Iscope,Ioffice,Ichecktime)

企业信息（企业注册号，类型，住所，法定代表人，注册时间，注册资本，营业期限，经营范围，登记机关，核准时间）

Upload(Uphone, Pid)

上传（手机号，图片编号）

Extract(Pid,Cid)

提取（图片编号，企业注册号）

Own(Cid, Iid)

属于（企业注册号，企业注册号）

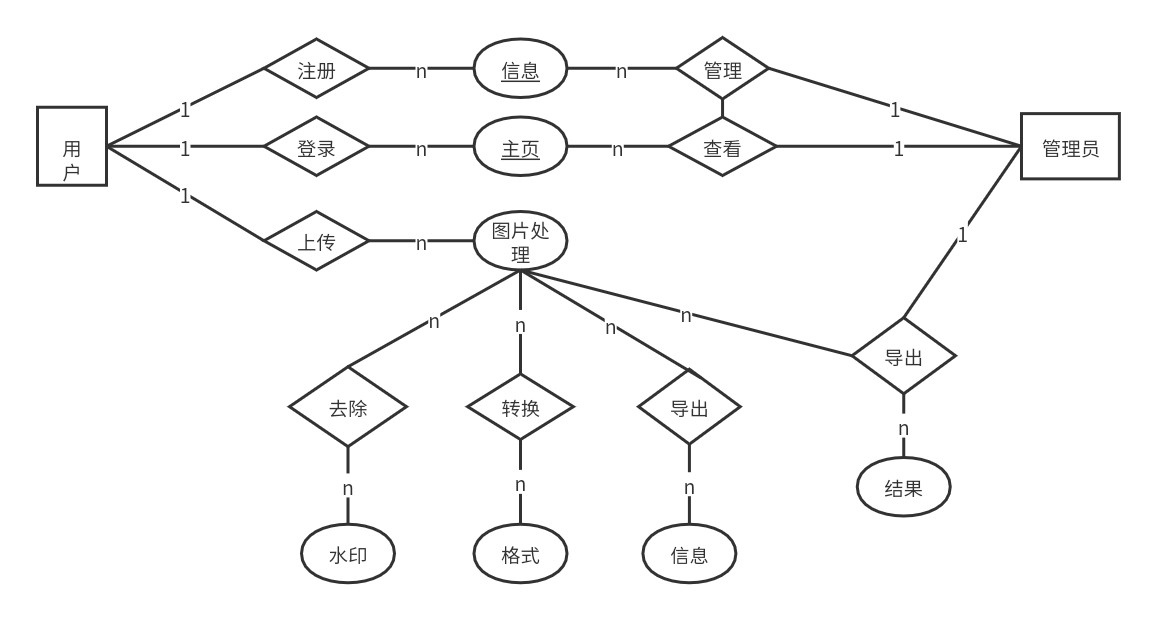


图4.1.3 E-R模型

## 4.1.4 通过Power Designer生成的概念模型CDM

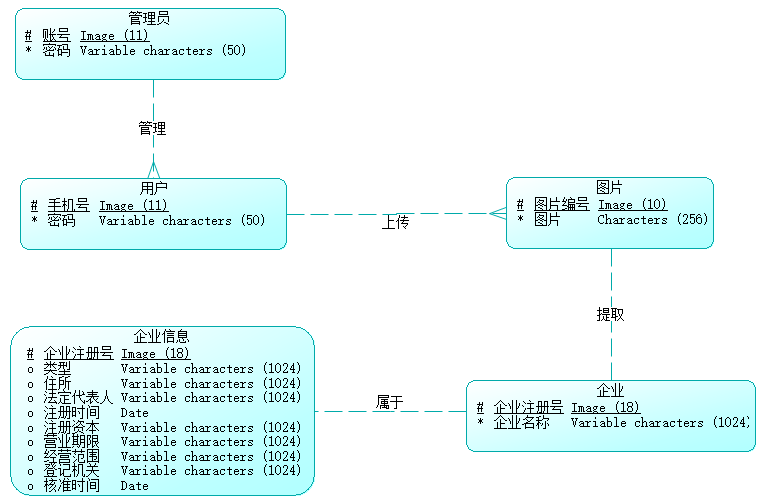


图4.1.2 概念模型CDM

## 4.2 设计软件系统总体结构

本项目采用SSM（Spring MVC+Spring+Mybatis）框架开发，是标准的MVC模式，将整个系统划分为View层，Controller层，Service层，DAO层四层。其中，Spring MVC负责请求的转发和视图管理，Spring实现业务对象管理，Mybatis作为数据对象的持久化引擎。整个系统架构运行流程如下图所示：

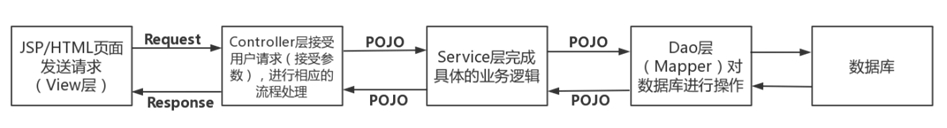


图4.2.1 系统架构运行流程

数据流程图（Data Flow Diagram，DFD/Data Flow Chart），是一种能全面地描述系统数据流程的主要工具，它用一组符号来描述整个系统中信息的全貌，综合地反映出信息在系统中的流动、处理和存储情况。它有两个特征：抽象性和概括性。综合这两个特征对该项目的数据流向进行分析，本项目的整体数据流程如图所示：

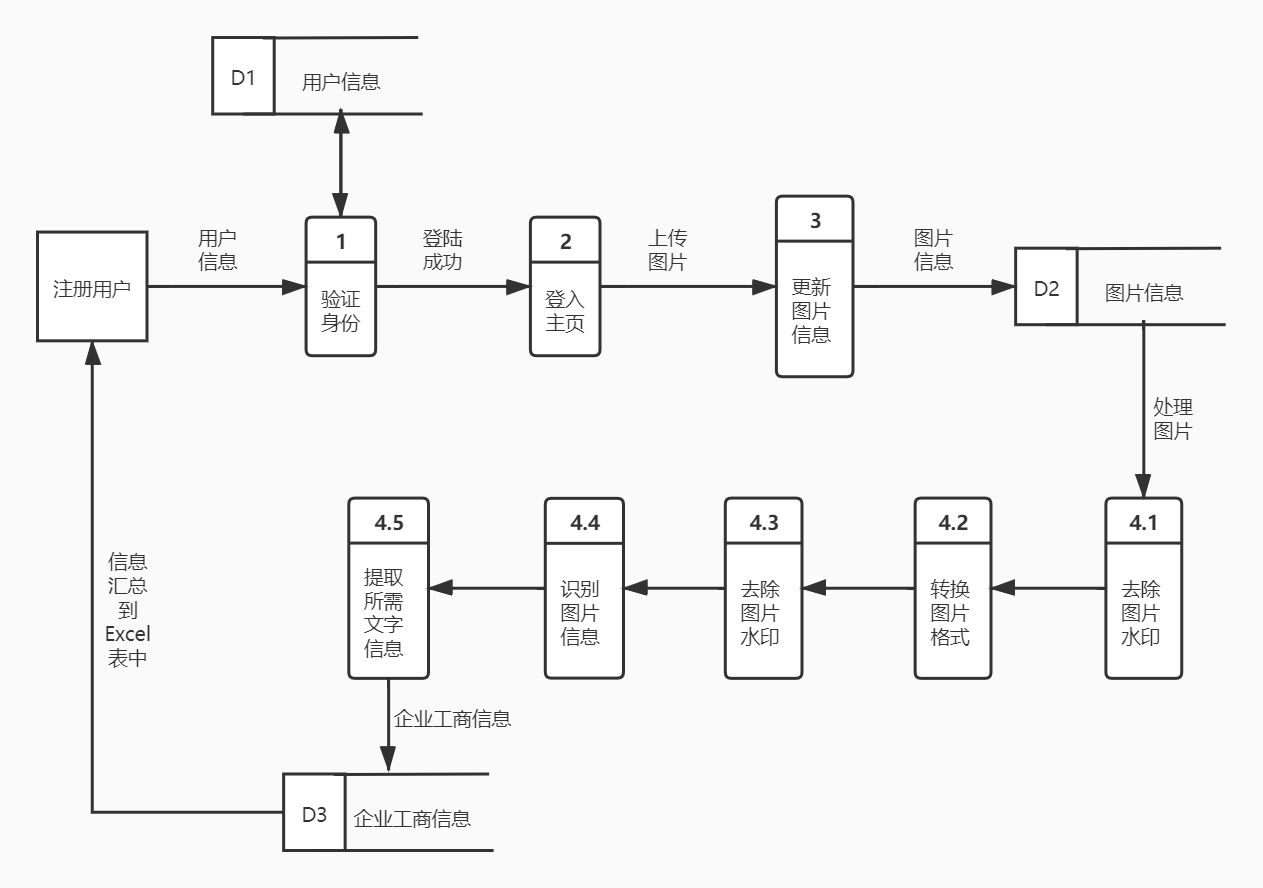


图4.2.1 数据流图

综合考虑系统的需求之后，系统的WBS如图4.2.2。

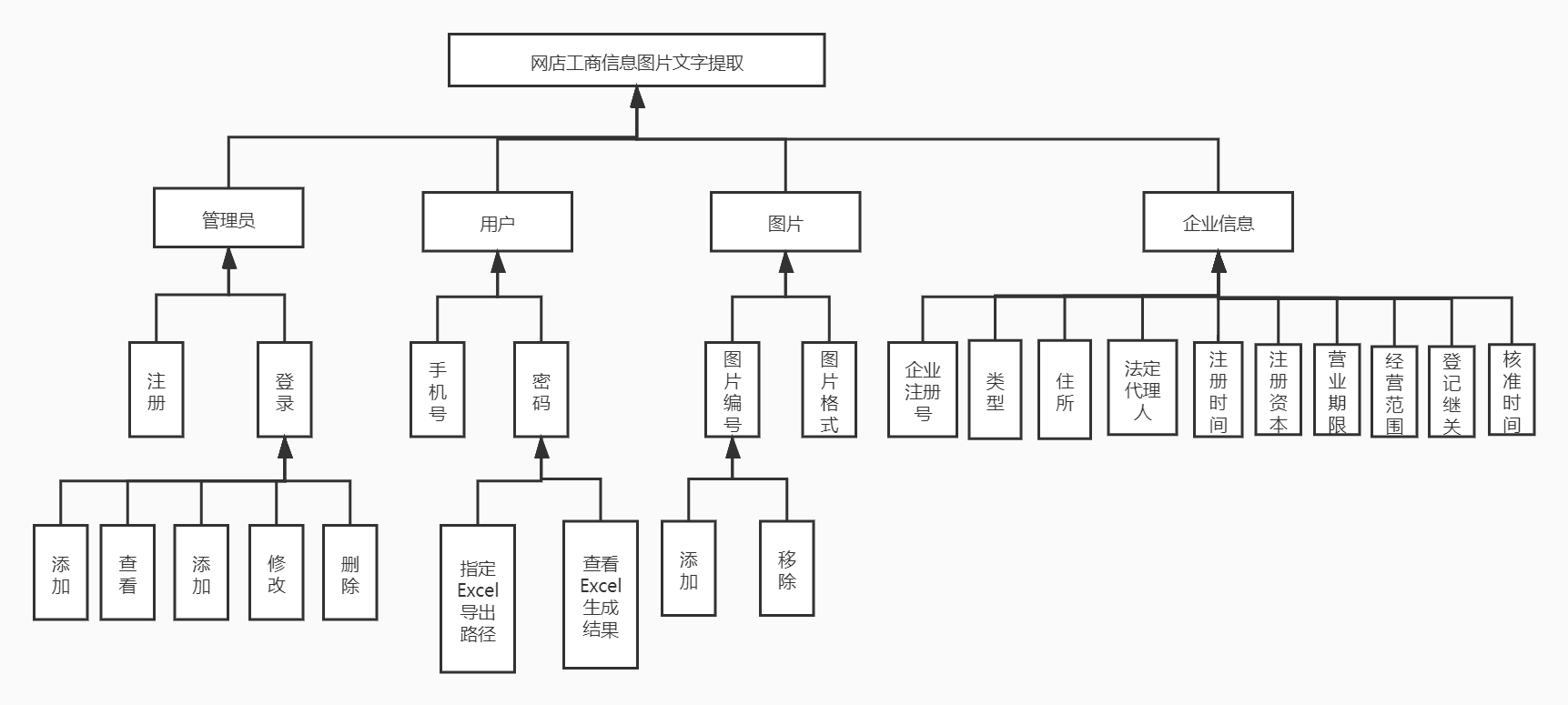


图4.2.2 WBS

## 4.3软件的重用方案

（1）代码的重用性  
   将代码模板化，我们将一些重复性高的代码包装在其自己的组件中，而不是在周围到处复制和粘贴代码。因为在开发软件的过程当中，我们采用的软件代码都是比较简单的，所以在某种程度上来说，如果有人想在我们的代码上继续研究开发，可以继续使用我们的代码去开发，但是如果他们想用自己的代码去实现自己的软件目标，也可以采用我们软件的相关类，以及对象的重用，这些都是相对来说比较简单的，开发者可以简单的读懂代码的类以及对象，然后可以根据他们的想法去开发自己的项目。  
（2）项目的重用性  
   在开发项目的过程中，因为我们设计的是较为简单的一个提取图片上关键信息的网站，所以在各项功能方面相对于一些有较高要求和精准度的关键信息提取网站来说，可能还达不到一定的标准，但是在较小范围内的提取关键信息还是可以用的，在后期的开发研究的过程当中，我们还是可以继续在这个项目的研究上进行开发，当然，一些同行要是想在我们的网站上继续进行开发，也是相对来说可以的，因为在一些功能的设计上，可能还是不太健全，但是在经过后期的修改之后，这样的情况相对来说是可以改变的，然后将网站的功能完善，应用与一些内容很庞大的信息提取和和精准度高的提取关键信息网站。  
（3）项目的领域上  
   在项目的领域方面，我们的项目是跟着社会的发展走，未来的提取关键信息网站肯定是很被社会所需要的，在这一方面肯定是毋庸置疑的，而且，这些技术也会越来越成熟高效。所以在目前，我们现在设计的这种网站的功能可能是不够健全的。但是我们可以在这个提取关键信息技术的领域，不停的将我们这个项目进行重用，进行一个项目的改进，最终让这个项目的功能越来越完善，以至于最后可以用于一些大型的要求精准度高的提取关键信息技术需要上。

## 4.4设计关键类的重点服务

（1）图片信息提取  
   本项目提供多种方式的图片信息提取，因为每个人他所需要的图片信息是不一样的，所以我们设置了不同提取方式的图片识别，有单个图片局部信息提取、单个图片全部信息提取和批量图片信息提取。这样就大大的方便了大家上传图片后更加高效的提取自己所需要的信息。  
（2）关键信息导入EXCEL  
   使用此功能可以将提取出来的信息导入到EXCEL表中，在导入过程中为了更加高效的导入关键信息，我们有设置了不同方式的导入方式，有按关键字导入、按企业注册号导入和按企业注册时间导入。其中，在按关键字导入时我们需要选择填入所需要的关键字、保存路径和保存格式（如TXT文件）；在按企业注册号导入和按企业注册时间导入时，我们又设置了不同顺序的导入方式，如按年、月、日从小到大或从大到小排列导入，还包括保存路径和保存格式（如TXT文件）；这样不仅帮助大家在导出关键信息时能快速的归类，还能帮助大家更加快速的获取关键信息。

# 第五章 运用设计

## 5.1数据字典设计

表User的结构：



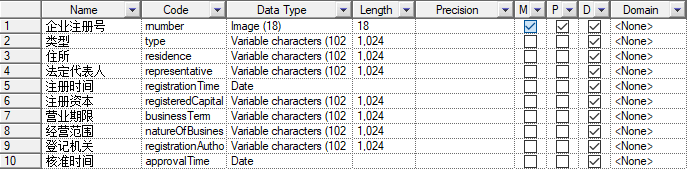
表Picture的结构：



表Company的结构：



表Information的结构:



表管理员的结构：

IMG_256

## 5.2安全保密设计

## 5.2.1 系统安全设计

由于在网络环境下，任何用户对任何资源包括硬件和软件资源的共享，所以必须通过制定相应的安全策略来防止非法访问者访问数据资源，对数据资源的存储以及传输进行安全性保护。在网店工商信息图片文字提取系统中，参考OSI的七层协议，从网络级安全、传输级安全、系统级安全和应用级安全等几方面进行考虑，主要遵循下面的设计原则:

1. 标识与确认

任何用户访问系统资源，必须得到系统的身份认证以及身份标识，如用户的数据证书、用户号码、密码。当用户信息与确认信息一致时，才能获准访问系统。

2、授权

对系统资源，包括程序、数据文件、数据库等，根据其特性定义其保护等级:对不同的用户，规定不同的访问资源权限，系统将根据用户权限，授予其不同等级的系统资源的权限。  
 3、日志

为了保护数据资源的安全，在系统中对所保护的资源进行任何存取操作，都做相应的记录，形成日志存档，完成基本的审计功能。  
 4、加密

为了保护数据资源的安全，在系统中对在网络中传输的信息必须经过高强度的加密处理来保证数据的安全性。

## 5.2.2 操作系统平台的安全管理

在操作系统平台上，应进行如下设置:

1、系统的超级用户口令应由专人负责，密码应该定期变换。  
 2、建立数据库的专用用户，系统在与数据库打交道时，应使用专用用户的身份，避免使用超级用户身份。  
 3、在系统的其他用户的权限设置中，应保证对数据库的数据文件不能有可写、可删除的权限。  
 4、选用较高安全级别的操作系统，时刻了解操作系统以及其他系统软件的动态，对有安全漏洞的，及时安装补丁程序。

## 5.2.3 数据库系统的安全管理

数据库系统是整个系统的核心，是所有业务管理数据以及清算数据等数据存  
放的中心。数据库的安全直接关系到整个系统的安全。在本系统中对此考虑如下:  
 1、数据库管理员(SA)的密码应由专人负责，密码应该定期变换。  
 2、客户端程序连接数据库的用户绝对不能使用数据库管理员的超级用户身份。  
 3、客户端程序连接数据库的用户在数据库中必须对其进行严格的权限管理，控制对数据库中每个对象的读写权限。  
 4、利用数据库的审计功能，以对用户的某些操作进行记录。  
 5、充分使用视图以及存储过程，保护基础数据表。  
 6、对于不同的应用系统应建立不同的数据库用户，分配不同的权限。

# 参考文献

1. 袁欣.计算机图像识别的智能化处理技术瓶颈与突破[J].电脑编程技巧与维护,2016,（8）：83-84.
2. 吉文帅.计算机图像识别智能化处理技术的分析[J].山东工业技术,2017,（15）：100.
3. 唐维.刍议计算机图像识别的智能化处理技术瓶颈与突破[J].电脑知识与技术,2018,（5）：196-197.
4. 刘海艳,陆映峰.计算机图像识别的智能化处理技术分析[J].建筑工程技术与设计,2017,（17）：219.