

## 基于模糊推理的名片识别后处理方法

马培军，苏小红，王雪峰

(哈尔滨工业大学计算机科学与技术学院，黑龙江哈尔滨 150001，E-mail: wfxjb@etang.com)

**摘要：**针对名片信息的特点和通常的OCR名片识别方法识别率较低的问题，提出一种新的基于模糊推理的名片识别后处理方法。该方法通过OCR识别得到的文本信息和候选文本信息来进行文本内容分析，通过OCR过程中的图像切分参数进行版面分析，在分析中均采用模糊推理的方法，同时，提出一种新的模糊运算的交型算子，应用于模糊推理运算中。最后综合上述内容分析和版面分析的结果得到最终的信息分类结果。实验结果表明，该方法在名片识别和分类正确率方面明显优于其他几种常用名片系统采用的算法，本方法不仅提高了OCR识别的正确率，而且还提高了经后处理以后的识别正确率。

**关键词：**模糊推理；OCR；识别后处理；分类

中图分类号：TP391 文献标识码：A 文章编号：0367-6234(2006)01-0015-04

Post-process method for business card recognition

based on fuzzy reasoning

MA Pei-jun, SU Xiao-hong, WANG Xue-feng

(School of Computer Science and Technology, Harbin Institute of Technology., Harbin 150001, China, E-mail: wfxjb@etang.com)

**Abstract:** In view of the characteristics of card information, a new post - process method for business card recognition based on fuzzy reasoning is proposed to solve the low rate of card recognition with the OCR method. The card information is analyzed according to OCR recognition results and candidate results. The typeset page is analyzed by image-syncopated parameters provided by the OCR system. In the above process, the fuzzy reasoning method is applied to the typeset analysis. Furthermore, a new type of intersection operator is designed which is used in fuzzy reasoning. Then the final results are obtained by fusing the above information. The experimental results demonstrate that this method is effective in improving the accuracy rate of OCR recognition and classification of business cards, including the post-process results when compared with other methods.

**Key words:** fuzzy reasoning; optical character recognition; post-process of recognition; classification

OCR系统可以将图像中的文字信息识别成文本。在很多情况下，只将图像识别成文本是不够的，需要进一步处理识别出来的文本信息，这就需要进行识别后处理。

在名片识别后处理研究方面，文献[1]提出基于先导词的后处理算法，这种方法完全依赖于先导词，在处理过长或过短先导词时效果很差；文献[2]提出自底向上的后处理分类算法，这种方法结合语义分析与版面分析，有一定的效果，但识

别正确率仍然不高。上述方法的一个共同特点就是大多采用了语义分析，但是名片上的信息都是简单信息，并不存在语义关系，因此，对于名片识别问题而言，采用语义分析的方法是不合适的。另外，以上算法都没有考虑OCR识别产生带有噪声的信息<sup>[3,4]</sup>。

实际上，人对名片上的信息进行分类的方法难以用精确的数学公式或模式进行描述，而是用“可能”、“很”等模糊量描述分类的原则和方法，因此，本文提出在名片识别和信息分类中使用模糊推理和模糊分类的一种新的名片识别后处理算法，解决了名片信息分类的描述问题。同时，提出针对名片识别后处理的纠错方法。

## 1 基于模糊推理理论的名片识别后处理算法

### 1.1 识别后处理系统的总体设计

识别后处理算法要完成对 OCR 识别结果的信息分类、纠错、以及与其他应用程序交互。整个算法组成如图 1。设计思想是：对一个识别结果字符串求其对各个信息种类的隶属度，如果其中一个明显高于其他的种类，则将其分为这个信息种类；否则进行相对分类，判断其更可能是哪一个信息种类。对每一个结果执行这些操作，然后对整个结果进行分析，以避免出现一些错误情况（如两个姓名等）。最后进行纠错处理。

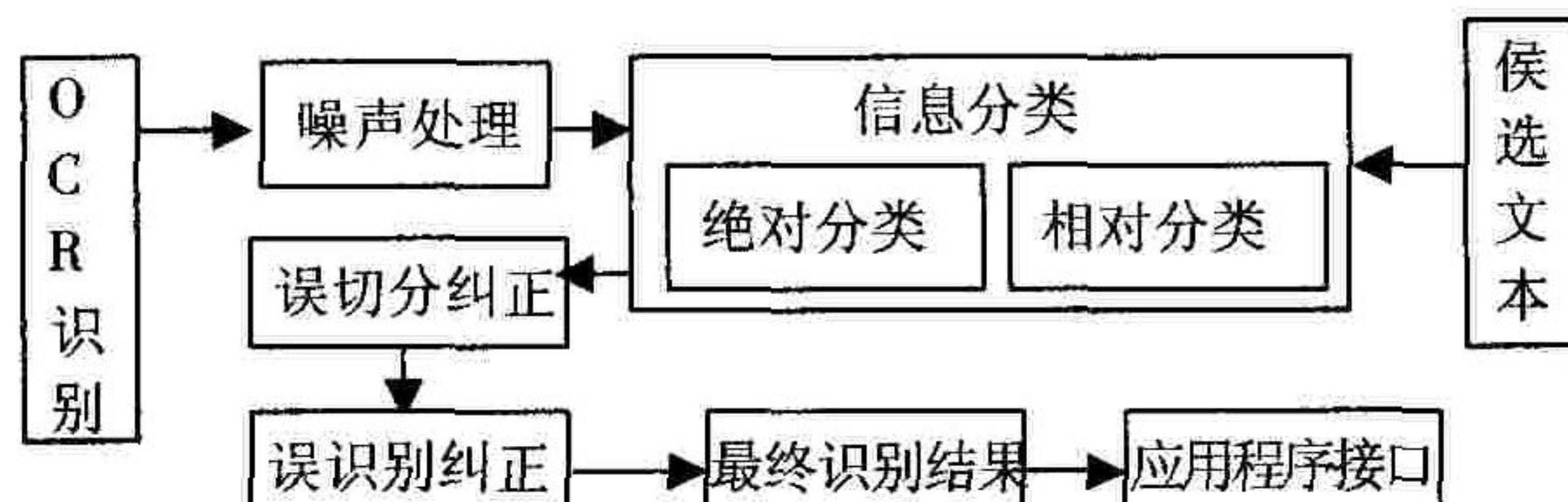


图 1 名片识别后处理系统的数据处理流程

### 1.2 对模糊推理模型的扩展

由于名片后处理中输入信息的特殊性，需要针对这种特殊性对模糊推理的模型进行扩展。首先从模糊推理基本模型获得一个用于信息分类的模糊推理实例。简单模糊推理规则实例：

rule 1 : if  $x$  is short and  $y$  is big, then  $S$  is a name

rule 2 : if  $x$  is long and  $y$  is small, then  $S$  is an address

加权扩展模型：由于在一条多维规则当中，各维在规则中的重要性往往不同，同时，一条模糊推理规则也不能认为百分之百正确，因此，在简单推理模型的基础上引入加权系数和置信度的概念，扩展成为加权推理模型。其规则实例如下：

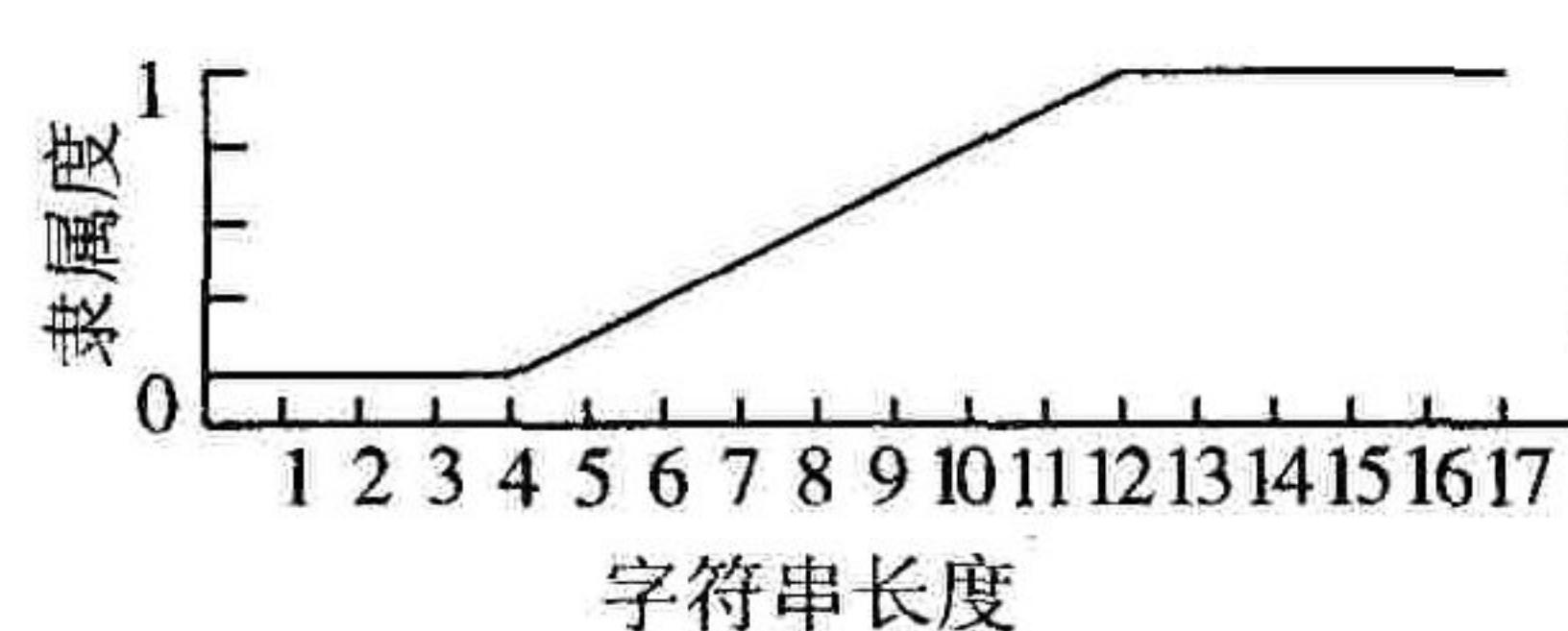


图 2 String is long 的隶属函数

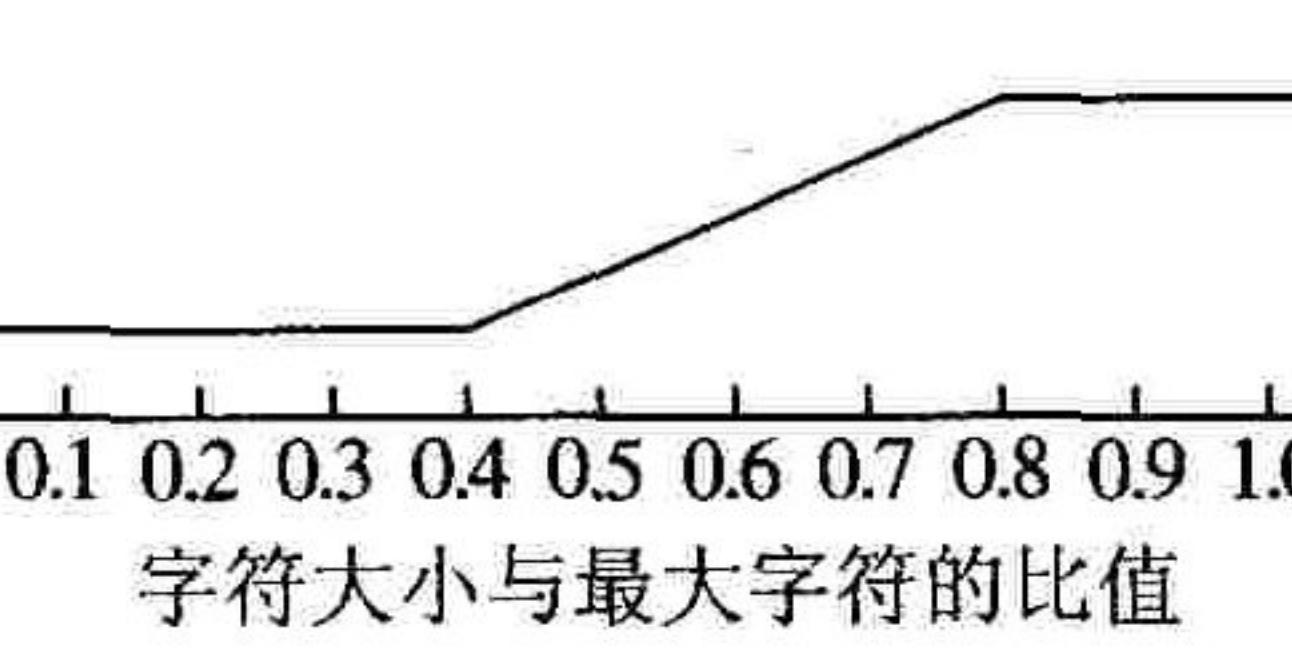


图 3 Font is big 的隶属函数

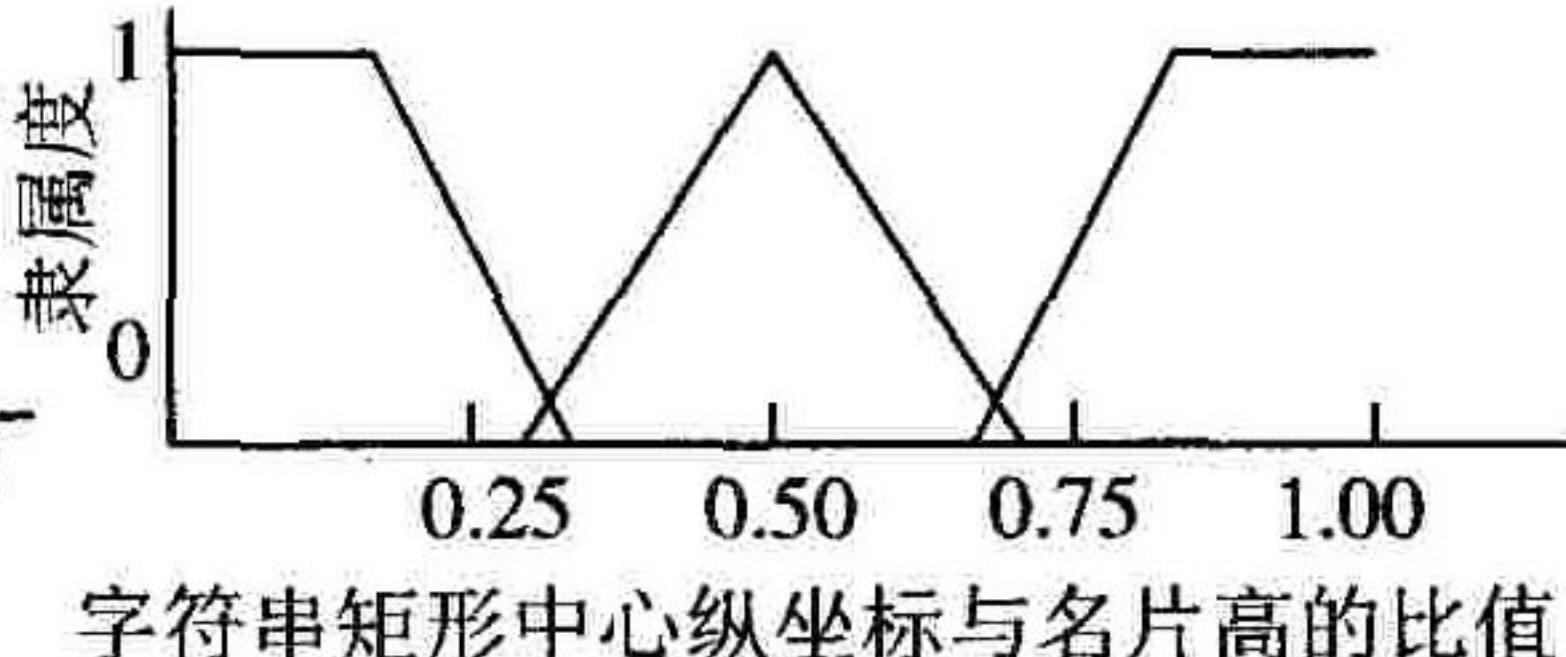


图 4 位置的隶属函数

#### 1.3.2 隶属度的语言值描述

在模糊推理规则中使用数量值的隶属度既不直观，也不便于确定数值的大小，使用语言值描述则直观和方便很多。如“如果字符串很长，他不太可能是姓名”，这里的“很”和“不太”都是对隶属度的语言值描述。

#### 1.3.3 合成模糊推理规则

使用 1.2 中扩展的模糊推理模型，结合 1.3.1 中的模糊前提和 1.3.2 中的模糊推论语言

rule 1 : if  $x$  is short with  $ID_1$  and  $y$  is big with  $ID_2$ , then  $S$  is name with  $CF_1$

rule 2 : if  $x$  is long with  $ID_3$  and  $y$  is small with  $ID_4$ , then  $S$  is address with  $CF_2$

其中  $ID$  = Importance Degree,  $CF$  = Certainty of Degree, 各  $ID$  之和为 1.

带噪声加权扩展模型：名片识别信息分类算法的输入是 OCR 系统识别的结果字符串，因此，不能把输入信息认为是纯净无噪声的信息，而是需要把这些输入信息当作带噪声信息进行处理。本文在加权模糊推理规则的基础上引入前提条件中的噪声系数，使之扩展成为带噪声加权扩展模型，其规则实例如下：

rule 1 : if  $S$  with  $NF_1$  contain last name with  $ID_1$  and  $y$  is big with  $ID_2$ , then  $S$  is name with  $CF_1$ .

rule 2 : if  $S$  with  $NF_3$  contain zip code with  $ID_3$  and  $y$  is small with  $ID_4$ , then  $S$  is name with  $CF_2$ .

其中  $NF$  = Noise Factor, 另外，只有字符串的内容会包含噪声信息，而字符串的长度、大小、位置等信息不会包含噪声信息。这一点在上面的规则实例中表现得很清楚。

#### 1.3 模糊推理规则的设计

本算法中的模糊推理规则的设计包括模糊前提的提出、模糊推论的提出，以及融合模糊前提与模糊推论成为模糊规则 3 部分。

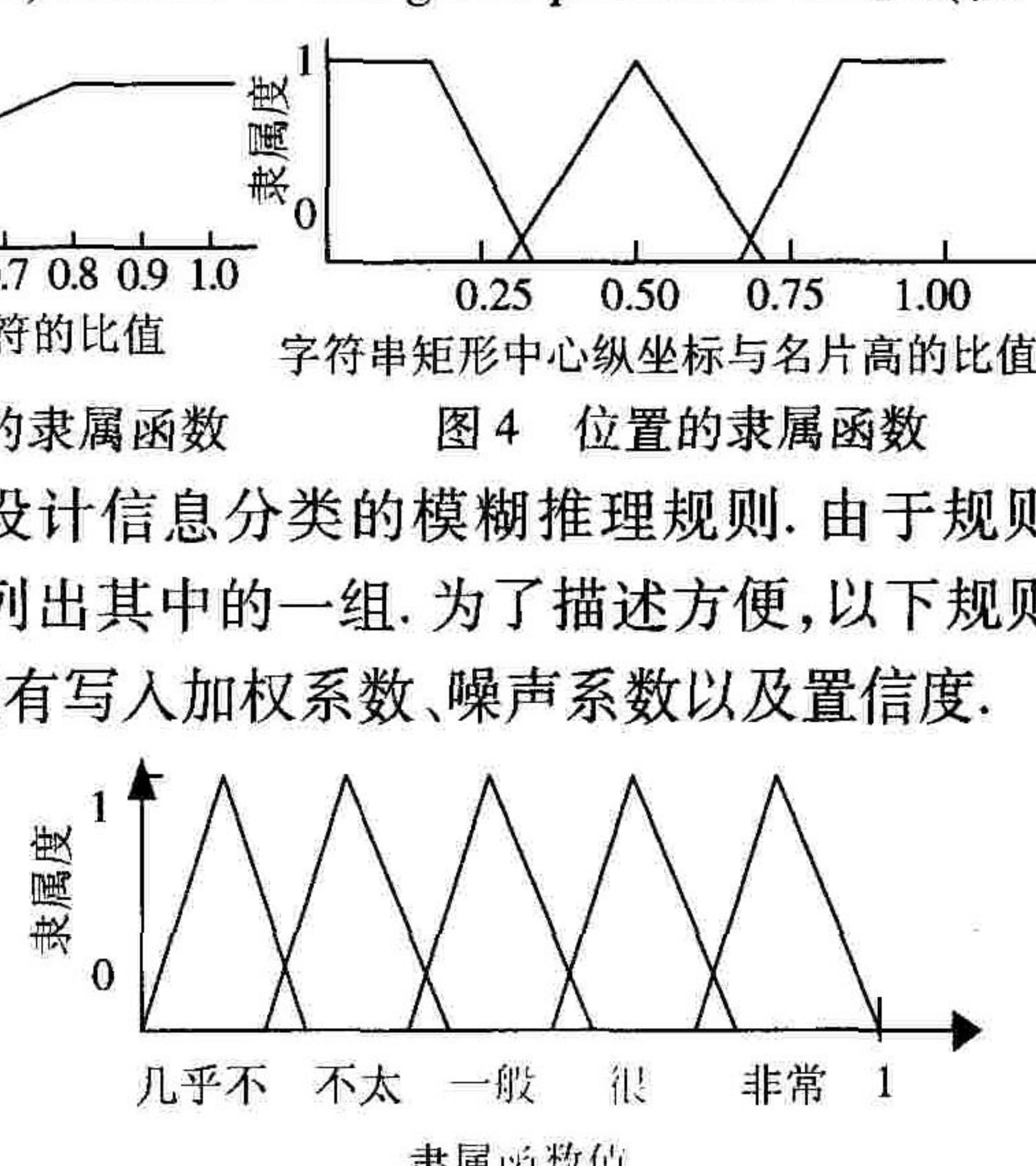
##### 1.3.1 模糊前提

以下是部分模糊前提的隶属函数图像。

1) String is long (与之相对应的模糊前提为 String is short) (图 2)

2) Font is big (与之相对应的模糊前提为 Font is small) (图 3)

3) Position of string is top/middle/bottom (图 4)



规则1: 如果一个字符串非常可能包含一个姓氏而且长度很短, 那么它非常可能是一个姓名.

规则2: 如果一个字符串很可能包含一个姓氏并且字体很大, 那么它很可能是一个姓名.

规则3: 如果一个字符串很短并且字体非常大, 那么它有可能是一个姓名.

规则4: 如果一个字符串在名片的中部并且字体非常大, 那么他很可能是一个姓名.

由于在处理字符串的各种属性时, 只认为字符串的内容包含噪声, 而认为字符串的位置、字体大小等属性不包含噪声信息. 因此, 对字符串的这两类属性分别采用1.2中提到的后两种扩展模糊推理模型. 以下具体是使用这两种推理模型计算隶属度语言值的方法.

### 1) 使用无噪声系数的扩展模糊推理模型

$F(\bigwedge_A (F_i(A_j(p_k)) * ID_l) \cap CF \cap C(l))$  规则中各个条件是“与”关系.

$F(\bigvee_A (F_i(A_j(p_k)) * ID_l) \cap CF \cap C(l))$  规则中各个条件是“或”关系.

其中,  $F$  为从隶属函数值到语言值的映射;  $A$  为规则中前提的集合;  $F_i, i = 1, 2, 3, 4, 5, 5$  个语言值的隶属函数, 分别对应1.3.2中的5个语言值;  $A_j$  为各个属性的隶属函数;  $p_k$  为属性值;  $CF$  为规则的置信度;  $C$  为从语言值到隶属函数值的映射;  $ID$  为条件的加权系数;  $l$  为规则中推论部分的语言值;  $\wedge$ : 交型算子1;  $\vee$ : 并型算子;  $\cap$ : 交型算子2.

### 2) 使用有噪声系数的扩展模糊推理模型

$F(\bigwedge_A (F_i(A_j(p_k)) * (1 - NF)) * ID_l) \cap CF \cap C(l))$  各个条件是“与”关系.

$$F_i(A_j(p_k)) \wedge F_m(A_n(p_q)) =$$

$$\begin{cases} \min(F_i(A_j(p_k)) + |A_n(P_q) - C(l)|, 1) & |A_n(P_q) - C(l)| \leq 0.2 \& F_i(A_j(p_k)) \geq 0.9 \\ F_i(A_j(p_k)) * F_m(A_n(p_q)) & \text{else} \end{cases}$$

注意, 上述公式只是描述了两个条件的情况.

该算子的意义是: 在各条件是“与”关系的推理规则中, 如果有超过50%的条件的匹配非常高( $>0.9$ ), 且这些条件权重较大, 那么其余的条件在匹配不是很差的情况下(隶属度绝对值之差 $<0.2$ )不会造成整个前提匹配度的下降, 而是使整个前提的匹配度略有上升.

## 1.4 字符串的分类算法

对于一个输入字符串的分类可以分为两步: 第一步是使用1.2中描述的扩展模糊推理模型对字符串分类, 称为绝对分类; 第二步是根据第一步分类的结果, 决定是否进行进一步的分类, 称为相

$F(\bigvee_A (F_i(A_j(p_k)) * (1 - NF)) * ID_l) \cap CF \cap C(l))$  各个条件是“或”关系.

$NF$  为噪声系数, 其获得方式为: 噪声长度 / 字符串总长度.

在以上的公式中,  $F_i$  要根据规则中的条件进行选择,  $F_i$  得到的是与规则中的一个条件的匹配程度, 将各个条件的匹配程度与加权系数相乘后, 根据条件之间的关系(“与”或“或”关系)分别通过“交型”或“并型”的模糊集运算获得推理规则的整个前提的匹配程度. 然后根据匹配程度, 用“交型”模糊算子调整规则中结论部分的隶属函数值, 这里采用了一个函数对语言值去模糊, 再用函数  $F$  模糊化的方法. 这样, 如果前提的匹配程度不高, 可以相对地降低结论的可能性, 避免推理错误的发生.

使用概率和算子  $\tilde{A} \vee \tilde{B} = \max\{\mu_{\tilde{A}}(x) + \mu_{\tilde{B}}(x) - \mu_{\tilde{A}}(x) * \mu_{\tilde{B}}(x)\}$ .

使用实数乘算子  $\tilde{A} \cap \tilde{B} = \mu_{\tilde{A}}(x) * \mu_{\tilde{B}}(x)$ .

$\wedge$ : 一般使用的交型算子都是使隶属度减小, 但是由于本文中使用的规则有一些特殊性, 在任何情况下都减小隶属度是不合适的, 如以下一条规则: 字符串包含一个姓氏并且字体非常大, 那么它非常可能是一个姓名. 使用一般的交型算子会出现如下情况: 字符串包含姓氏, 字体很大, 显然“字体很大”与“字体非常大”的匹配度 $<1$ , 也就减小了整个前提的匹配程度; 实际上, 在字符串包含姓氏的情况下, 如果字体很小, 应该减小整个前提的匹配程度, 但字体很大的(不是非常大)应该适当增加前提的匹配度而不是减小. 因此, 本文提出了一种新的交型算子.

对分类.

### 1.4.1 绝对分类

绝对分类是将待分类的字符串作为1.3.3中规则的输入参数求其对于各个信息种类的隶属度, 得到该字符串是各个信息种类的可能性. 如果输入字符串是某一个信息种类  $m$  的可能性明显高于它是其他信息种类的可能性, 则可确定  $m$  是这个字符串的分类结果. 以下是一个绝对分类的实例: 设名片上只有姓名、地址和电话号码3个信息种类. 把字符串作为模糊推理规则的输入, 得到隶属度: 该字符很可能是姓名, 不太可能是地址和电话. 则可直接认为该字符串的信息种类是姓名.

### 1.4.2 相对分类

如果得到的字符串属于某两个信息种类的可能性比较接近(由于名片信息种类的有限性,不会出现某个字符串属于两个以上信息种类的可能性相近的情况),则不能直接认为该字符串的分类是可能性较高的,而需要针对这两个特定的信息种类进行进一步的推理,以确认该字符串是哪一个信息种类.由于这种推理是在对比两个信息种类的情况下进行的,因此,称为相对分类.

简单考虑的话,任何不同的两个信息种类之间都需要一组相对分类规则,也就是说如果信息种类有 $N$ 种,那么相应地要有 $(N-1) * N/2$ 组分类规则.但实际上,很多信息种类不可能同时成为一个字符串的候选分类.例如,任何一个字符串都不会在绝对推理中获得对电话号码和姓名这两个信息种类相近的隶属度.因此,实际需要的相对分类规则并不多.以下是实际采用的一组针对电话和传真的相对分类规则.

- 1) 如果当前没有电话号码,则该字符串分类为电话号码;
- 2) 如果当前已经有电话号码,没有传真号码,且字符串属于电话号码的可能性高于传真号码,则该字符串分类为电话号码;
- 3) 如果当前已经有电话号码,没有传真号码,且字符串属于电话号码的可能性不高于传真号码,则该字符串分类为传真号码.

### 1.5 纠错方法

#### 1.5.1 误识别的纠错方法

1.5.1.1 可发现的误识别 这种误识别会导致字符串存在不合理性,如字符串(03)2810-S730,根据名片的特点,可以确认该字符串是电话号码或传真号码,因此,可以进一步确认该字符串中的‘S’是字符‘5’被误识别的结果.这种错误一般可以通过统计的方式进行纠错.

1.5.1.2 不可发现的误识别 这种误识别不会导致字符串存在不合理性,如将姓名“三上真司”误识别为“三上直司”.这种错误很难通过机器发现,纠错也就变得非常困难.对于此类信息的纠错主要依靠第三方库,如姓名库和地址库等.

#### 1.5.2 错误切分的纠错方法

OCR系统的工作流程是先进行图像切分,然后进行模式识别.针对名片识别这个特定的需要,如果切分错误,那么无论识别率有多高,最终的信息也都是不准确的,因此,判断错误切分必须成为后处理算法的一个部分.

名片在排版上具有多样性,OCR切分过程中

很可能出现错误.如图6中的Fax信息字段,Fax和后面的具体号码之间空隙较大,可能被OCR切分成两个字符串;同样,如果两部分信息之间的空白部分过小,就可能被OCR的图像切分模块忽略,导致两部分信息被切分为一段信息.

**Canon**

E&D開発センター  
E&D第一開発部  
E&D14設計室  
主任研究員 高野 隆夫  
キヤノン株式会社  
〒146-8501 東京都大田区下丸子3-30-2  
Tel: (03)3757-6754 (直通) (03)3758-2111 (代表)  
Fax: (03)3757-8830  
E-mail: takano@cl.d.canon.co.jp

图6 名片图像实例

在目前实验中的切分错误都发生在电话号码和传真字段.针对上面的两个例子中出现的情况提出以下解决方案:1)如果某个被判断信息种类为电话的信息段过长(通过模糊推理判断),则切分该信息段.2)如果发现类似的关键词(TEL, Tel, FAX等),但是没有后续信息,则把该信息段与其后的一个信息段合并.

结果纠错可进一步提高最终结果的正确率,但是并不能完全解决OCR系统的误识别问题,最终结果中的错误大部分由误识别引起.另外,名片中经常出现一些图标,它们在被识别为字符之后很难处理.系统测试结果也证明了这一点.

## 2 实例测试结果与分析

使用基于模糊推理的信息分类算法构建一个名片自动识别系统,该系统接受名片图像作为输入,可以将识别结果直接存储到数据库的相应字段中或者将识别结果作为XML文档返回.通过使用该系统对300张日文名片进行测试,结果如表1、2所示.另外,将本系统的实验结果与一些已有系统的测试结果<sup>[5]</sup>进行了比较,结果表明本文提出的识别后处理算法是非常有效的.

表1 识别测试结果

	字符总数	正确识别数量	正确率/%
汉字与日文假名	1 545	1 454	94.11
字母与数字	3 220	3 184	98.88

表2 分类测试结果

有用信息数量/条	分类正确数量/条	分类正确率/%	错误分类数量/条	分类错误率/%	拒绝分类数量/条
323	316	97.83	0	0	7

从表3可以看出,本文提出的算法明显优于其他系统采用的算法.另外,其他系统的最终正确率都明显低于OCR系统的识别正确率,而本文

(下转第129页)