北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.028

版面相似中文表单的分类方法研究

王思萌　高良才†　王悦涵　李平立　汤帜

北京大学计算机科学技术研究所，北京　100080; † 通信作者,E-mail:glc@pku.edu.cn

摘要　针对具有相似版面的中文表单，提出一种简单有效的基于距离度量的表单分类方法，该方法对表单的用户填写信息、布局信息和位置偏移分别进行距离度量，并通过 3 种权重有效降低了用户填写信息的随机性、版面相似表单的布局一致性和位置抖动性对表单分类的影响。实验表明，所提方法在多个中文表单图像库上的分类准确率达到 90%以上，相较于目前最新的表单分类方法有明显提高。

关键词　表单分类; 距离度量; 权重计算

中图分类号　TP391

A Study on Classification of Forms with Similar Layout

WANG Simeng, GAO Liangcai†, WANG Yuehan, LI Pingli, TANG Zhi

Institute of Computer Science and Technology, Peking University, Beijing 100080; † Corresponding author, E-mail: glc@pku.edu.cn

Abstract　The authors propose a simple but effective distance based method to identify forms with similar layouts by measuring the user filled-in data, preprinted data and dithering data. The proposed method utilizes three kinds of weight components to mitigate the impact of randomness of user filled-in data, consistency of similar layouts and position dithering respectively. Experimental results show that proposed method can achieve more than 90% identification accuracy on a series of data sets, which is significantly better than the results of the state-of-the-art method.

Key words　form classification; distance metric; weight calculation

目前，在很多业务(如银行、保险和统计等)中，大量的中文表单通过打印/复印等形式生成后，传递给客户进行打印填写或手工填写，因而导致大量的中文表单以纸质形式存在，给后期的表单自动化处理带来许多挑战与困难。为了使办公更加自动化，进而能够从表单中抽取挖掘出有用的信息，对表单自动化处理的需求日益强烈。

表单的自动化处理通常包括纸质表单的扫描、读入、分类、版面分析、识别和编辑等一系列过程。其中表单分类是表单自动化处理流程中关键的步骤，它能够对版面分析和识别过程进行指导，使处理流程更加自动化。表单分类的粒度往往因业务场景而不同，有的仅针对语言进行分类[1]，有的根据是否有印章进行分类[2]。本文研究版面相似中文表单的分类问题，分类目标是将采用不同表单模板的表单区分开，即同一类的表单除用户填写区域外均完全相同。

在银行和保险等机构中，有大量的相似业务存在，如“取款”和“存款”；另外还有隶属于不同银行或保险机构的同一种业务，如不同银行的汇款单。银行等业务部门的分类归档工作通常是由人工完成的，通常业务部门在处理同一客户的需求时，就会有多种相似的表单需要客户和工作人员进行填写和登记，然后再由工作人员人工进行录入操作。因此银行有进行相似表单分类的业务需求。邮局等可以进行电汇的工作部门常常会汇集来自不同银行的电汇单，因此更有将类似的电汇单进行分类的需求。这一类中文表单模板通常有国家的标准设计要求，

国家自然科学基金(61202232)和北京市自然科学基金(4142023)资助
收稿日期: 2014–06–28; 修回日期: 2014–10–22; 网络出版时间: 2014-12-01 09:26

因此他们的版面设计几乎完全相同，差别只体现在表单标题中的业务名称或银行名称和标志上。

目前绝大部分表单分类方法都是从表单图像直接进行特征的提取，有的提取全局的特征，有的提取局部的线条等结构特征。但是这些方法都不适用于版面相似表单的分类，因为对于这类表单，这些方法所提取出的特征，尤其是结构特征，几乎都是相似的特征，区分度很小。所以这一类方法在进行表单分类时，常常会被表单的结构相似性所迷惑，分类效果较差。

在实际应用当中，需要进行分类的中文表单绝大部分是已经由用户填写好(打印填写或手写填写)的表单。由于表单的类别差异仅体现在表单版面部分的差异上，与用户所填信息无关。因此对于表单分类任务而言，可以认为用户填写信息是噪声信息。在此前提下，版面相似表单的分类主要有以下两个挑战。

1) 用户所填信息的随机性。用户所填信息相对于固定的表单版面完全因人而异，并且同一种类表单的用户所填信息也不尽相同。因此在提取全局特征时，用户所填信息的位置变化以及字体变化等会引起全局特征的变化，造成分类错误。一个典型的例子如图 1 所示，(b)和(c)是属于同一类型的两张中文表单。将两张表单中的用户填写信息叠加在一起，结果如图 1(a)所示，绿色部分是表单(b)的填写信息，红色部分是表单(c)的填写信息。从中可以看出用户填写信息之间具有很大不同：填写的信息内容和字体不同；同时填写信息的位置之间也有较大差异。

2) 中文表单版面部分的区分信息过少。由于版面相似表单的类别差异只体现在表单版面部分的差异信息上(如前文所提到的表单标题和银行标志等)，差异信息非常有限，往往只靠标题中几个字符的差别确定表单的种类，但这些差别又会被用户填写信息的差异所掩盖，导致分类出现错误。因此如何最大化地利用这类有限的差异信息，成为版面相似表单分类的关键和挑战。在这种情况下，使用单纯的欧氏距离进行计算，同类表单之间的距离很有可能大于不同类表单之间的距离。一个具体的例子见图 2，其中(a)和(b)属于同一种类型的表单，而(a)和(c)属于不同类型的表单。由于(a)和(c)的用户填写信息相近，字体和位置均大致相同，而(a)和(b)在用户填写信息的字体、内容和位置上均有较大的差别。(a)和(c)虽然版面区域有明显差异信息(标题中的银行名称)，但此类区分信息过少，因此导致非同类表单的欧氏距离反而小于同类表单的距离，从而出现分类错误。

为应对以上挑战，本文提出一种简单有效的基于加权距离的中文表单分类算法，减少用户填写信息的随机性带来的影响，同时放大表单版面中区分信息的重要性，从而针对版面相似的中文表单取得较好的分类性能。

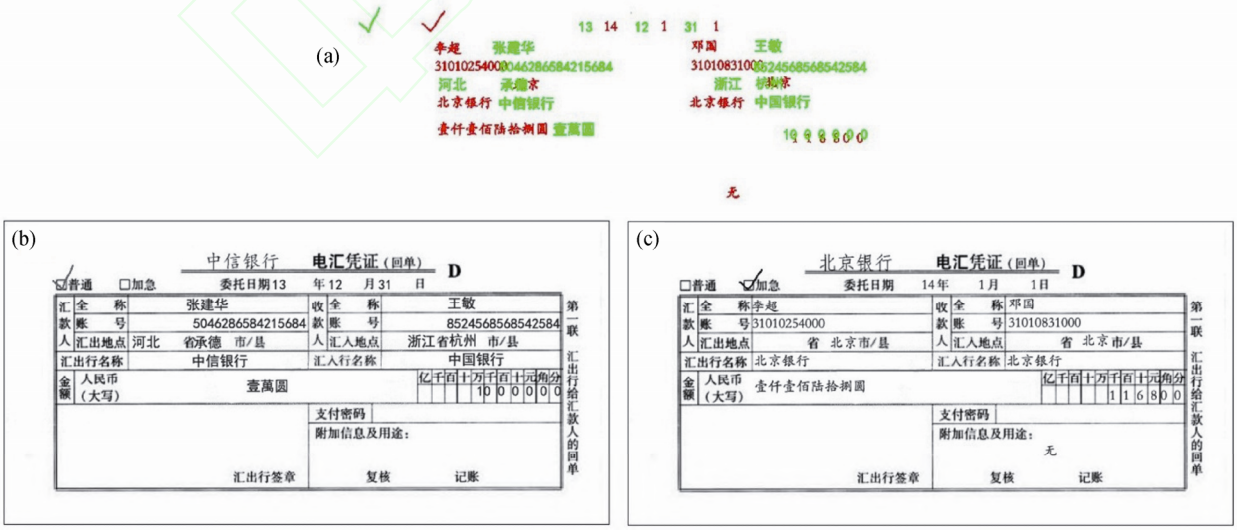


图 1　同种中文表单中的用户填写内容的差异示例

Fig. 1 Illustration of two forms from the same category but with different user filled-in data

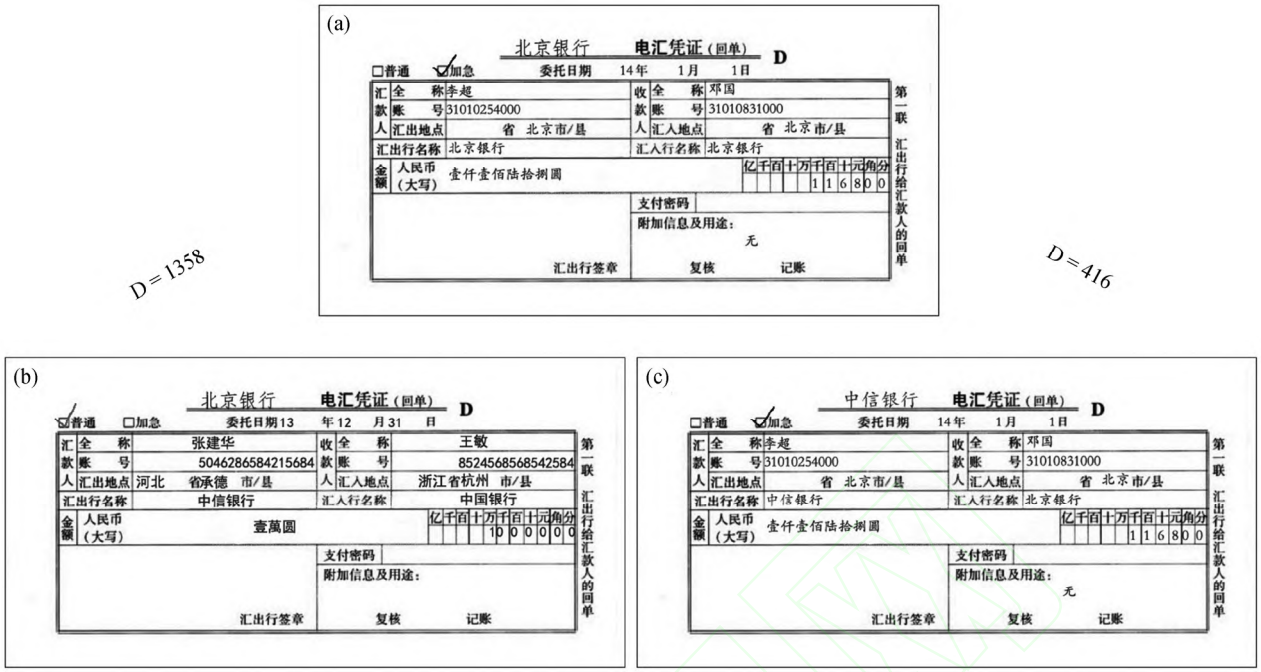


图 2　同类中文表单距离大于异类中文表单距离的示例

Fig. 2 Illustration of distances between forms of the same category and that of different categories

1　相关工作

表单分类作为表单识别的关键步骤, 已经引起研究者的广泛关注, 大量的表单分类方法被相继提出, 其中主要的方法类型包括 3 种: 基于全局特征提取的方法、基于结构特征的版面分析方法和基于分层特征表达的方法。

在全局特征提取方面, 有基于字数、单元格和 Haar 特征等的方法[3]。Sarkar[4]提出一种类 Haar 特征的方法, 并使用潜在条件独立(Latent Conditional Independent, LCI)模型进行表单分类。Shimotsuji 等[5]提出一种点集匹配技术, 将表单中单元格的中心标记成点, 然后对不同的表单进行点集匹配。

对表单结构特征的提取也是表单分类中非常有效的手段。Bynu 等[6]提出了基于表单中线段提取的方法, Ting 等[7]则将线段和文本表示成字符串进行分类。这一类的方法对于具有明显结构性版面的表单, 取得了较好的效果。

对表单特征进行分层表达也是具有较高分类准确率并且计算复杂度较低的方法。Duygulu 等[8]提出一种基于 X-Y 树的分层方法, 表示表单中的矩形结构。Bagdanov 等[9]则将表单的物理版面信息提取成多层 XY 树, 并编码成固定长度的特征向量, 然后使用神经网络模型和多层感知机进行分类。

然而, 上述方法难以处理相似表单, 从相似表单中提取的特征也非常相似, 因此往往将具有相似版面的表单判断为同一类, 从而导致分类错误。

殷绪成等[10]提出利用 OCR(Optical Character Recognition) 识别标题从而进行表单分类的方法, 并应用于金融票据中, 取得较好的效果。然而, OCR 技术在表单识别中对表单模板有较大的依赖性, 而且错误的识别结果将直接影响分类正确率; 另外基于 OCR 的方法需要识别大量无关信息, 耗时且效率低。

为此, Arlandis 等[11]提出专门针对相似表单的分类算法, 该算法首先检测出相似表单的标志区域, 再用基于距离度量的方法对该区域进行模板匹配。但是该算法要求利用空白表单来提取标志区域。在实际应用中, 由于保密性和安全原因, 空白表单通常难以获取[12], 因此该方法在实际应用时, 适用性较弱。

最新的相关研究中, Bukhari 等[13]提出一种基于 EMD(Earth Mover's Distance)的表单分类方法。该方法利用表单二值化后的连通域面积和连通域像素点位置信息, 将表单灰度图转化为彩色图的一种伪彩色编码算法。实验证明, 该算法对用户所填信

息的位置变化具有很好的鲁棒性，对表单的整体位置偏移也有稳定的分类效果。但是，该算法将表单的标题等区分性信息与其它信息等同处理，未有效利用该类区分性信息，导致最终的分类准确率不高。另外该算法是一种寻优算法，具有较高的时间复杂度。

2 本文方法

本文方法是一个基于距离度量的中文表单分类方法，我们提出了 3 种权重，即 3 个参数，分别用于弱化表单中用户填写信息的随机性、放大表单版面中的区分信息和减少表单位置的抖动性。首先根据训练样本，计算不同类别表单的平均表单，并计算出这 3 种权重的分值；然后针对待分类的表单，计算该表单与不同平均表单的距离，并将这 3 种权重用于距离加权；最后，根据加权计算后的距离，确定最终的表单类别。该表单分类算法的实现过程具体如下。

2.1 预处理

由于本文所提出的表单分类方法是基于点到点的匹配，也就是表单对齐后直接计算两个表单图像矩阵的欧氏距离。因此要求表单之间首先要进行方向与尺度的归一化。扫描的表单图像由于只包含表格和文字等灰度信息，因此边缘检测等预处理工作能够较容易地完成，从而能够直接对表单进行方向校正和尺度归一化处理。预处理工作完成后，对训练的表单图像进行预计算操作。预计算是指计算不同类别表单的平均表单，类似于在人脸合成研究中的平均人脸计算方式，对处于同样位置的不同图像的像素点求平均。同一类表单的平均表单的计算方式如下:

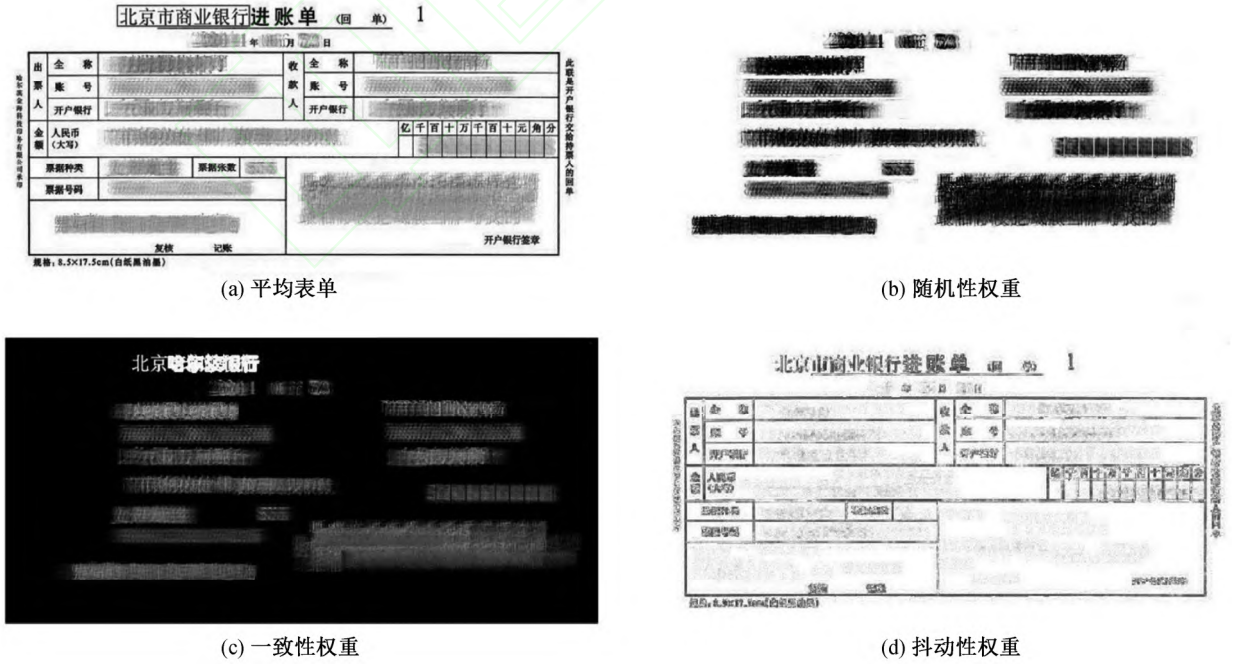
$$\mathrm{Avr}_i = \frac{\sum_{j=1}^{n} T_{ij}}{n}, \tag{1}$$

其中 Avr_i 是对第 i 类表单的训练图像求平均后的结果，T_{ij} 是第 i 类的第 j 张训练图像，n 是第 i 类的训练图像数量。一个平均表单图像如图 3(a)所示。

2.2 权重计算

2.2.1 随机性权重

为了降低用户填写信息的随机性带来的影响，本文提出随机性权重的计算方式：对同一类中文表单的训练图像在同一位置求方差，计算出该类表单的方差图像，再根据方差值求得所在位置的随机性权重。可以预见，用户填写区域由于随机性而方差较大，但表单的版面部分则由于同一类表单的版面信息基本相同而方差较小。在进行表单间的距离计算时，具有较大方差的用户填写区域作为噪声会干扰表单的分类，因此需要降低此处的权重；而具有较小方差的版面区域则应保持或放大权重。因此，



(a) 平均表单　(b) 随机性权重

(c) 一致性权重　(d) 抖动性权重

图 3　平均中文表单和 3 项权重示意图

Fig. 3　Illumination of the average form and the three weights

可以采用对方差求倒数的方式来定义某一位置的随机性权重, 计算方式如下:

$$I_{ik}^{\text{random}}=\frac{1}{\sigma_{ik}^{\text{random}}}, k=1,2,...,N, \tag{2}$$

其中 I_{ik}^{random} 是第 i 类表单在第 k 个像素点的随机性权重, $\sigma_{ik}^{\text{random}}$ 是第 i 类表单在第 k 个像素点的方差值, N 是表单图像的像素点数(训练表单图像经过归一化具有同样的图像尺寸)。

然而, 表单的版式部分几乎一致, 因此其方差值可能为零, 从而导致权重计算公式的分母为零, 计算不可逆。为了避免这种计算不可逆的情况发生, 本文引入常数项λ, 类似拉普拉斯平滑因子的作用。因此, 随机性权重的计算方式调整如下:

$$I_{ik}^{\text{random}}=\frac{1}{\sigma_{ik}^{\text{random}}+\lambda^{\text{random}}}, k=1,2,...,N, \tag{3}$$

λ_{random} 表示随机性常数, 在本文中, 其取值设置为方差 $\sigma_{ik}^{\text{random}}$ 在 N 个像素点的均值。

通过对随机性权重的计算, 降低了后续距离计算中用户填写区域的权重。随机性权重所对应的图像如图 3(b)所示。黑色部分表示较低的权重, 在图中主要对应用户填写区域。

2.2.2 一致性权重

一致性权重是为了减少不同种类中文表单之间的一致性对分类的影响。由于相似版面中文表单的版面区域存在大量的相同内容和结构, 而该区域具有的区分信息却非常有限。因此, 在待分类表单与不同种类的平均表单进行比较时, 需要减弱版面相似内容的权重, 而放大版面区分信息的权重。为此, 本文提出一种与式(3)相似的计算方法, 将版面相同区域的权重降低, 突出版面区分信息(即上文提到的表单标题部分)。计算方差时, 借助预处理中得到的每一类的平均表单, 计算平均表单相同位置之间的方差。用户填写区域在平均表单中虽然具有一定的差异和随机性, 但平均化处理后这种差异被极大的弱化了; 同时, 平均化处理并没有改变不同类型表单之间的版面区分信息。因此在利用平均表单计算方差时能够突出该区域, 强调版面中的类别区分信息。一致性权重的计算方式如下:

$$I_{k}^{\text{cons}}=1-\frac{\lambda^{\text{cons}}}{\sigma_{k}^{\text{cons}}+\lambda^{\text{cons}}}, k=1,2,...,N, \tag{4}$$

其中 I_{k}^{cons} 是表单图像在像素点 k 的一致性权重, σ_{k}^{cons} 是所有平均表单在像素点 k的像素方差, λ^{cons} 与式(3)相似, 同样是为了避免计算不可逆而引入的一致性常数, λ^{cons} 被设置为方差 σ_{k}^{cons} 在 N 个像素点的均值。可以看出, 所有类别的表单图像在同一像素点拥有同样的一致性权重, 而上文的随机性权重和下文的抖动性权重则与训练表单类别相对应。

如图 3(c)所示, 表单图像中较亮的区域表示较大的一致性权重, 这些区域主要是中文表单的用户填写区域和表单版式区分信息, 其中前者的亮度明显小于后者, 说明表单中的版式区分信息相比于用户填写区域更加突出, 同时表单的版面相似区域, 由于具有较高的一致性, 因此一致性权重较小, 在表单图像中表现为较暗的区域。

2.2.3 抖动性权重

为了避免预处理中方向校正和尺度归一化误差、表单版面本身打印偏离等所引起的表单位置偏差情况, 本文引入抖动性权重, 基于平均表单进行该权重的计算。像素点与其八邻域像素的方差越大, 说明抖动偏差所带来的影响越大, 需要降低该区域的权重。抖动性权重的定义如下:

$$I_{k}^{\text{dither}}=\frac{\lambda^{\text{dither}}}{\sigma_{ik}^{\text{dither}}+\lambda^{\text{dither}}}, k=1,2,...,N, \tag{5}$$

其中 I_{k}^{dither} 表示第 i 类表单图像在第 k 个像素点的抖动性权重, $\sigma_{ik}^{\text{dither}}$ 表示第 i 类平均表单在像素点 k 的方差, 该方差是由均值图像中的第 k 个像素点与其周围八邻域像素点共 9 个像素值的方差, λ^{dither} 是与式(3), (4)类似的常数, 设置为方差 $\sigma_{ik}^{\text{dither}}$ 在 N 个点均值的两倍, 避免出现方差计算不可逆的情况, 同时避免该权重减弱差异化信息的重要性。

图 3(d)为抖动性权重在中文表单图像中的表现, 从中可以看出, 抖动性权重主要对文字和表格边界区域进行了权重限制, 避免因为图像校正、归一化和打印时的像素点位置偏差引起分类错误。

3 种权重的计算方式相似, 由于一致性权重和另外两种权重是相反的效果, 因此在另两种权重的基础上, 一致性权重的计算方式做了一些变化。最终 3 种权重的效果均达到像素点的置信度越高, 权重越大的效果。

2.3 中文表单分类

本文提出的中文表单分类算法基于以上 3 个权重的距离度量方法, 在计算欧氏距离的基础上, 引入每个像素点所对应的权重, 具体的计算式为

$$D(C,\mathrm{Avr}_i)=\sqrt{\sum_{k=1}^{N}(C_k-\mathrm{Avr}_{ik})^2\cdot I_{ik}^{\mathrm{random}}\cdot I_k^{\mathrm{cons}}\cdot I_{ik}^{\mathrm{dither}}},\quad (6)$$

其中,$D(C,\mathrm{Avr}_i)$是待分类表单 $\boldsymbol{C}$ 和第 i 类平均表单 Avr_i 的距离, I_{ik}^{random} 、I_k^{cons} 、I_{ik}^{dither} 分别是第 i 类表单第 k 个像素点的随机性权重、一致性权重和抖动性权重。

3 个权重相结合能够更加突出版面相似中文表单版面的区分信息, 同时减少用户填写部分和位置偏移对分类的影响, 达到更加准确的分类效果。

3 实验结果

如上文所述, 中文表单尤其是商业中文表单具有用户隐私信息, 目前尚无相关的中文表单数据库公开, 而且版面相似中文表单的数据库更加难以获得。因此本文实验在作者生成的中文表单数据库中进行。

本文生成了 4 种中文表单数据库: 用户填写信息位置变化较小的版面相似中文表单、用户填写信息位置变化较大的版面相似中文表单、版面不相似中文表单以及综合版面的中文表单(包括版面相似中文表单和版面不相似中文表单)。本文模拟真实中文表单的填写情况, 将各个中文表单库中的用户填写部分用打印机进行填写, 填写的内容、字体和相对表格的位置均不相同。

3.1 版面相似表单数据库上的对比实验

本文生成了两个版面相似中文表单的图像数据库, 其中分别为用户填写信息有较小位置偏移的图像库和有较大位置偏移的图像库。这两个图像库的用户填写部分在内容和字体上均不相同, 尽可能真实地模拟了实际中文表单的填写情况。两个中文表单图像库均有 200 张版面相似的中文表单, 包含 10 类表单, 每类 20 张。本文比较的方法是对用户填写信息变化鲁棒的基于 EMD 的最新表单分类算法[13]。另外, 实验也对比了直接与平均表单进行欧氏距离计算而不加入任何权重信息的类似均值分类器的方法(“平均比较”), 将其作为基准方法。

本文的对比实验结果见图 4, 分别是用户填写信息位置变化较小的图像库实验结果(图 4(a))和位置变化较大的图像库实验结果(图 4(b))。

从实验结果可以看出, 本文所提出的方法对用户所填信息的变化较为鲁棒, 并且在版面相似中文表单的分类中, 本文的分类方法远优于最新的

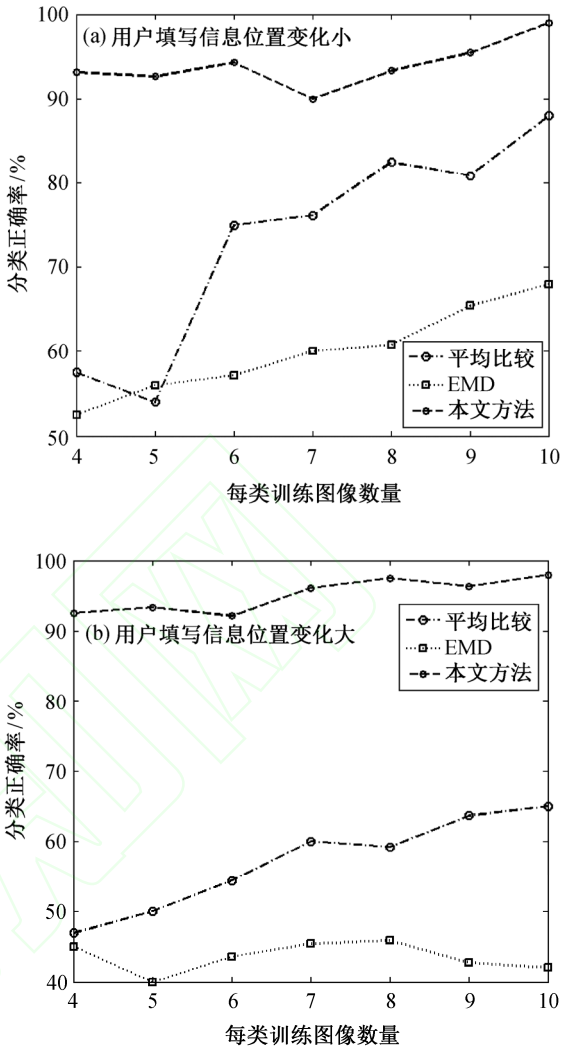


图 4 相似中文表单图像库的分类结果
Fig. 4 Experiment results on similar forms

EMD 方法[13]和基准方法。可见, EMD 算法由于没有利用相似版面区分信息, 并不适用于版面相似中文表单的分类问题, 而本文的方法则在版面相似中文表单分类上具有较大优势。

3.2 版面不相似表单数据库上的对比实验

本文也对比了具有不相似版面的中文表单分类情况。不相似版面的中文表单图像库包含 200 张中文表单, 也包含 10类表单, 每类 20 张表单。实验所选取的对比方法与 3.1 节实验相同。

实验结果如图 5, 从中可以看出, 在版面不相似中文表单的分类结果中, 各个方法差别较小, 准确率均在88%以上，其中 EMD 算法的结果比其在版面相似表单图像分类中的结果有大幅提高。在版面不相似的表单图像库中, 本文提出的方法同样取得最

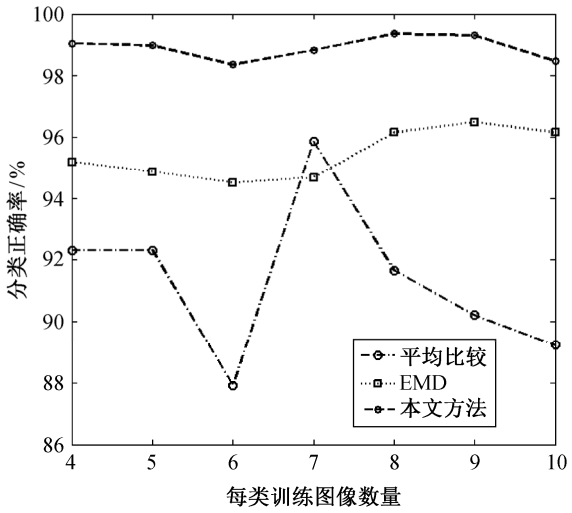


图 5　不相似中文表单图像库分类结果

Fig. 5　Experiment results and comparison of different classification methods in dissimilar data sets

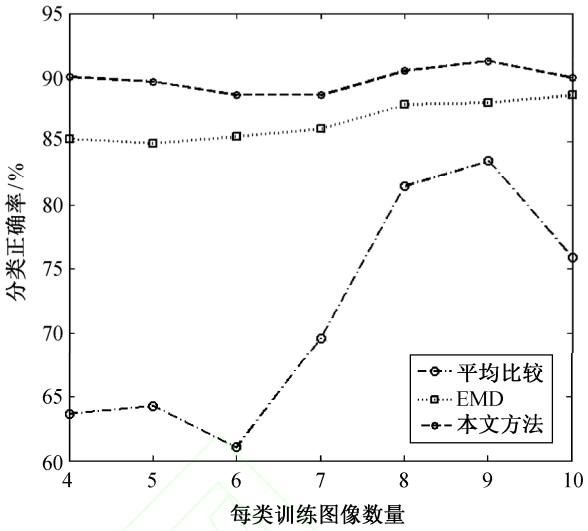


图 6　混合中文表单图像库分类结果

Fig. 6　Experiment results and comparison of different classification methods in mixed data sets

好结果。另外，从实验结果中可以看出，“平均比较”方法的结果对训练样本的数量比较敏感，随着样本数量的变化，结果抖动比较严重，而其他两种方法包括本文方法的性能比较稳定。

3.3　综合表单数据库上的对比实验

最后，本文生成了一个较大的具有 440 张中文表单的混合版面图像库。该图像库包含 22 类中文表单，每类 20 张，其中既有版面相似中文表单，又有版面不相似的中文表单。其中版面相似表单有 180 张，版面不相似表单有 260 张。对比方法与上述实验相同，分类结果见图 6。在混合版面的中文表单分类中，本文方法同样取得最佳效果，而基于EMD的算法[13]也表现出稳定有效的分类结果。与基于EMD算法相比，本文方法更加稳定有效，受到训练样本的影响也较小。版面相似表单和版面不相似表单的比例为9: 13，有大量的不相似表单存在，本文方法仍然有稳定有效的分类结果。可以看出，本文所提出的方法拓展性较佳，可以用于混合版面中文表单的分类，且更加切合实际应用场景。

4　结束语

本文提出了一种简单有效的中文表单分类方法，在距离度量的基础上引入了 3 种权重，分别控制版面相似中文表单分类中用户填写信息的随机性、版面相似内容的一致性和中文表单预处理中可能出现的抖动性，这 3 种权重结合起来进行版面相似中文表单分类。在 4 个不同表单图像库上的实验结果表明，无论在版面相似中文表单上，还是在版面不相似和混合版面的中文表单上，本文方法均取得较好的分类效果，而且性能稳定。未来我们会将本文方法扩展到日文、英文等外文表单上，同时进一步测试提高本文方法的性能。

参考文献

[1] Poon B, Saami R, Amin M A, et al. Dimensionality reduction and feature selection methods for script identification on document images. Information Technology in Industry, 2014, 2(1): 1–5

[2] Alaei A, Delalandre M. A Complete logo detection/recognition system for document images // 2014 11th IAPR International Workshop on IEEE, 2014: 324–328

[3] Saund E. Scientific challenges underlying production document processing // IS\&T/SPIE Electronic Imaging, International Society for Optics and Photonics. 2011: 787402–787402

[4] Sarkar P. Image classification: classifying distributions of visual features // ICPR 2006. 18th International Conference. 2006: 472–475

[5] Shimotsuji S, Asano M. Form identification based on cell structure // ICPR. 1996: 793–797

[6] Byun Y, Lee Y. Form classification using dp matching // Proceedings of the 2000 ACM symposium on

Applied computing-Volume 1. 2000: 1–4

[7] Ting A, Leung M K. Business form classification using strings // Proceedings of the 13th International Conference. 1996: 690–694

[8] Duygulu P, Atalay V. A hierarchical representation of form documents for identification and retrieval // International Journal on Document Analysis and Recognition. 2002: 17–27

[9] Bagdanov A D, Worring M. Fine-grained document genre classification using first order random graphs // Sixth International Conference on Document Analysis and Recognition. 2001: 79–83

[10] 殷绪成, 江世盛, 韩智, 等. 层次型金融票据图像分类方法. 中文信息学报, 2006, 19(6): 70–77

[11] Arlandis J, Perez-Cortes J C, Ungria E. Identification of very similar filled-in forms with a reject option // ICDAR. 2009: 246–250

[12] Chen N, Blostein D. A survey of document image classification: problem statement, classifier architecture and performance evaluation

[13] Bukhari S S, Ebbecke M, Gillmann M. Business forms classification using earth mover's distance // 2014 11th IAPR International Workshop on Document Analysis Systems (DAS). 2014: 11–15