DOI: 10.11992/tis.201709038
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.tp.20180417.1720.010.html

基于图表示和匹配的表单定位与提取

谭婷[1]，吕淑静[2]，吕岳[1,2]
（1. 华东师范大学 上海多维度信息处理重点实验室，上海 200062; 2. 中国邮政集团公司上海研究院 图像分析与智能系统联合实验室，上海 200062）

摘　要：为了实现对不同类型、分辨率和方向的快递表单上用户感兴趣区域信息的获取，本文提出了一种基于图表示和匹配的表单定位与提取方法。选择参考表单中已有的印刷图案或字符等关键区域作为基准位置，进行图的表示。基于图像分割得到的候选关键区域对待处理表单进行图表示。然后，根据图的属性计算待处理表单与参考表单的相似度。最后，将最大相似度对应的同构图作为参考表单图的最优匹配，并建立同构图与参考表单图位置映射，定位出表单。本文实验数据集来源于真实场景下采集的快递包裹表单图像。实验结果表明：本文算法在快递包裹表单图像上具有良好的性能，对旋转、光照变化、局部遮挡具有较好的鲁棒性。

关键词：图像分割；表单提取；表单定位；图表示；图匹配；同构图；快递包裹分拣
中图分类号：TP751.1　　**文献标志码**：A　　**文章编号**：1673−4785(2019)02−0231−08

中文引用格式：谭婷，吕淑静，吕岳. 基于图表示和匹配的表单定位与提取 [J]. 智能系统学报，2019，14(2)：231–238.
英文引用格式：TAN Ting, LYU Shujing, LYU Yue. Form location and extraction based on graph representation and matching[J]. CAAI transactions on intelligent systems, 2019, 14(2): 231–238.

Form location and extraction based on graph representation and matching

TAN Ting[1], LYU Shujing[2], LYU Yue[1,2]
(1. Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200062, China; 2. ECNU-SRI Joint Lab for Pattern Analysis and Intelligent System, Shanghai Research Institute of China Post, Shanghai 200062, China)

Abstract: To obtain information of a user's interested region on express package images of different types, resolutions, and directions, a form location and extraction method based on graph representation and matching is proposed in this paper. A reference form is needed in this method. First, key regions such as the existing printed patterns or characters in the reference form are chosen as nodes to build the reference graph. Second, graph representation is conducted on the form to be processed based on the candidate key region derived from image segmentation. Then, the similarity between the reference form and the candidate form is calculated according to attributes of the graph. Finally, the isomorphic graph with the maximum similarity is chosen as the optimal matching of the reference form and graph, and the position mapping of the isomorphic graph and the reference form and test image is established to locate the form. The experimental datasets in this paper originate from express package images collected in practical scenarios. Experimental results indicate that the proposed algorithm has good performance on express form images. Especially, good robustness is achieved for rotated, illuminated, and partially shaded images.

Keywords: image segmentation; form extraction; form location; graph representation; graph matching; isomorphic graph; express package sorting

表单作为重要的信息载体在实际生活和工作中有着广泛的运用，表单中某些特定字段、图案、符号等都有可能包含用户感兴趣的重要信息，如订货单中的订单号、发票的具体项目及金额、快递运单中的收货地址和手机号码等。人工录入的方式采集数据，费时费力，而且容易出错，因此利

收稿日期：2017−09−20.　**网络出版日期**：2018−04−18.
通信作者：谭婷. E-mail: tanting_hn@163.com.

用计算机对表单图像进行自动化信息提取有着强烈的应用需求，可以大幅度降低工作量，提升工作效率。

表单自动化处理的主要过程包括表单图像采集、表单定位、信息区域提取、识别等[1]。其中表单定位和提取是表单识别前重要的预处理过程，预先获取表单关键信息区域有利于更方便、准确地识别表单填写的内容信息。本文方法主要工作是对物流快递表单中与用户信息相关的文本区域进行定位和提取，如快递表单上收/寄件人姓名、电话号码、地址等信息。该处理过程得到文本图像块可用于后续识别工作的输入数据，建立字符图像数据库，图像特征学习的训练样本，具有广阔的应用前景。

表单提取过程中常见的方法是检测表单中的直线，将其作为表单提取的参考位置[2-3]。基于直线的检测法所处理的对象更倾向于类似于表格类结构化的表单，但对缺乏框线和非固定形式的非结构化表单的处理存在明显的不足。

另一类表单定位与提取的方法是采用对表单的布局或表单元素进行描述的方法，如建立搜索分类树[4]或设定提取信息的关联指令[5]。这种对表单的布局或表单元素进行描述的方法缺乏灵活性。

表单图像具有特定的布局方式，因此采用参考模板来提取表单也是一种重要的研究方法，如使用空白表单模板与待匹配表单基准点对齐[6-7]或使用傅里叶−梅林变换重定向表单[8]的方向。Cesarini [9]提出通过属性图结点的具体数值和图的模型特征实现刚性配准，建立待处理图和参考图的对应。

以往的模板匹配方法依赖于对基准点的严格要求和预先约定，而基于非层次有向关系属性图[9]方法在寻找对应区域位置时，难以避免预先识别关键字。本文将模板匹配和图匹配的方法相结合，提出一种基于图表示和匹配的表单定位与提取方法。

图匹配方法在计算机视觉领域有着广泛的应用，如特征点对应[10-11]、形状匹配[12]、目标检测和识别[13-15]、视频分析[16]，图像的视觉特征在图匹配过程中考虑两图之间最小结构失真以实现对应。

本文方法在处理多个类别的表单图像时，需要预先选取对应类别表单图像中已有的图案区域设计匹配待处理表单的参考表单模板图，该过程避免了对字符的识别，简化了分类提取的过程。另外，图匹配方法适用于混杂场景下目标检测和异常点判别，结合这一优势，在定位表单时采用图匹配的方法对定位的正确性进行验证。

1　表单图表示

1.1　参考表单的图表示

1.1.1　参考表单关键区域的选取

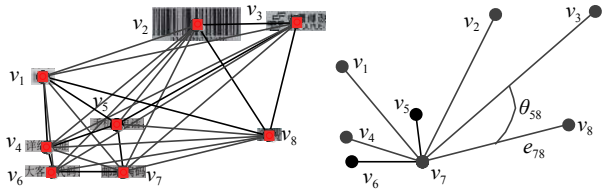
本文在建立参考表单图表示时，由用户手动选择能反映表单特征的关键区域，比如具有可区分特征的表单公司标志、特定图案、字符块等。由于表单图像上字符较多，背景较为复杂，后续图匹配过程需要足够多的关键区域实现配准，同时匹配计算量适度，建议选取 5~8 个图案完整、清晰、大小适中的图像块作为关键区域，图 1 给出一个从邮政快递包裹面单上选取关键区域的例子。

图 1　参考表单关键区域选取样例

Fig. 1　An example for key area selection of reference form

1.1.2　关键区域的图表示

以关键区域为图结点，建立如图 2(a) 所示参考表单的全连接无向图表示。将该无向图定义为 $q=(V,E;o,\varphi)$，其中 V 为图结点，对应表单的关键区域。E 为图的边，对应结点间的相互连接关系。ω 表示每个结点 v 的结点属性，φ 表示每个结点 v 在图 q 中的结构属性。图 2(b) 为图 2(a) 中结点 v_7 的结构属性表示。



(a) 参考表单图表示　　(b) v_7 的结构属性

图 2　参考表单图样例

Fig. 2　An example for graph of reference form

1) 结点属性 $\boldsymbol{o}$。SIFT 对图像局部特征的描述具有良好旋转和尺度不变性，对光照有较强的鲁棒性。采用 SIFT 来描述图的结点属性表示为

$$\boldsymbol{o}_i=\{\boldsymbol{f}_{i1},\boldsymbol{f}_{i2},\cdots,\boldsymbol{f}_{im}\}\quad m=1,2,\cdots,M \tag{1}$$

式中：$\boldsymbol{f}_{ij}$ 为 128 维的 SIFT 特征向量，表示 v_i 中第 j 个特征点；M 为正整数，表示结点 v_i 的特征维度。

2) 结构属性φ。φ表示结点 v_i 结构属性, 它包括两个子属性: 结点权重属性ω和夹角属性θ, 结点 v_i 的结构属性表示为$\varphi_i=\{\omega_i,\theta_i\}$。

权重属性ω_i。该属性表示以结点 v_i 为固定端点, v_i 与其所有邻接点 v_j 连接边 e_{ij} 的长度的向量集合。该属性表示如下:

$$\omega_i=\{e_{i1},e_{i2},\cdots,e_{ij}\}\quad i,j=1,2,\cdots,N,i\neq j \tag{2}$$

如 v_7 射线簇属性为$\{e_{71}, e_{72}, e_{73}, e_{74}, e_{75}, e_{76}, e_{78}\}$。

夹角属性θ_i。$\alpha(e_{ij},e_{ik})$ 表示图中以结点 v_i 为顶点, e_{ij} 和 e_{ik} 分别为与邻接点 v_j 和 v_k 连接边缘所组成的夹角, 结点 v_i 所具有的夹角属性表示为以 v_i 为顶点的夹角向量集合θ_i, 表示如下:

$$\theta_i=\{\alpha(e_{ij},e_{ik})\}\quad i,j,k=1,2,\cdots,N,i\neq j\neq k \tag{3}$$

根据上述描述, $q=(V,E;o,\varphi)$即为参考表单关键区域的图表示。

1.2　待处理表单的图表示

1.2.1　待处理表单候选关键区域的选取

本文采用选择性搜索方法[17]将待处理表单分割得到许多图像小块, 这些图像块中包含与参考表单关键区域对应的区域或部分的区域。如图 3 所示, 该算法使得灰度相似且位置相近的像素合并, 然后根据图像块的大小、灰度梯度实现图像块粗略过滤, 选择图案、字符相对较集中的区域作为待处理表单图像的候选关键区域。

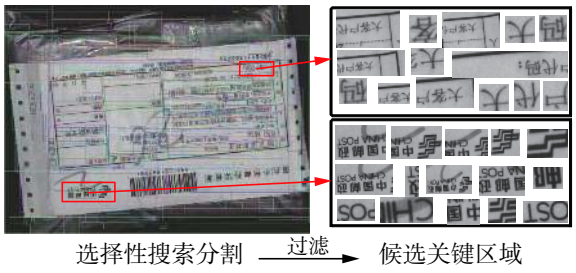


图 3　待处理表单候选关键区域选取样例

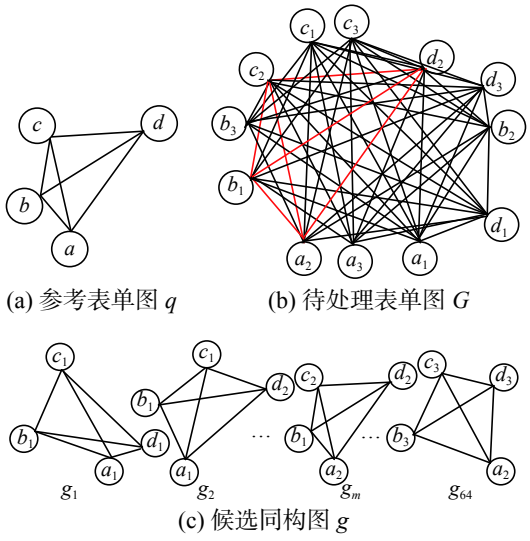
Fig. 3　An example for candidate key area selection of test form

1.2.2　候选结点筛选

为提高匹配参考表单图的效率, 比较候选关键区域与关键区域的结点属性 ω 相似度, 筛选出相似度最高的前 3 个图像块作为图匹配的候选结点, 建立关键区域与候选关键区域的对应关系, 去除大量相似度过小的候选关键区域, 降低匹配复杂度。

1.2.3　候选关键区域图表示

对候选结点参照参考表单图建立的过程, 建立如图 4(b) 所示待处理表单的全连接图 G。与参考表单图全连接不同的是, 对应同一关键区域的 3 个候选关键区域间不连接。随后, 对图 G 中标签互异的候选子图 g 进行结点和结构属性描述。



(a) 参考表单图 q　　(b) 待处理表单图 G

(c) 候选同构图 g

图 4　候选同构图

Fig. 4　Candidate isomorphic graph

2　表单图匹配

由于图像分割策略局限性, 分割的候选关键区域可能出现欠分割和过分割的问题。另外, 对应于关键区域的位置出现局部遮挡, 容易得到错误的候选关键区域。为此, 通过对参考表单与待处理表单进行图匹配, 验证和确认关键区域是否对应准确。

2.1　候选同构图

给定 $G=(V, E)$ 和 $G_1=(V_1, E_1)$ 是两个图, 假设存在双射$\varsigma: V\to V_1$使得对所有 $x,y\in$ V 均有 $xy\in E$ 等价于$\varsigma(x)\varsigma(y)\in E_1$, 则称 G 和 G_1 是同构的。假设参考表单图表示为$q=(V^q,E^q;o^q,\varphi^q)$, 待处理表单图表示为$G=(V^G,E^G;o^G,\varphi^G)$, 图 g 与图 q 中对应的候选结点赋予与 q 中相同的标签, 如图 4(a) 中的$\{a\}$对应于图 4(b) 中的$\{a_1,a_2,a_3\}$。图 G 中结点标签互异的图 g 与图 q, 恰好满足ς这一映射关系, 故称图 g 为图 q 的同构图。因此, 图像匹配过程为从图 G 中寻找一个与图 q 最相似的同构图 g, 或寻找与子图 qs 最相似的同构子图 gs, 是一个图匹配的问题。通过度量同构图 g 与图 q 的相似性, 找到相似差异最小的同构图 g_m 或同构子图 gs_m, 作为与图 q 最佳匹配图。如图 4 所示, 按照同构映射ς的定义, 在图 4(b) 所示图 G 中, 图 4(a) 所示图 q:$\{a,b,c,d\}$对应的候选同构图有 g_1:$\{a_1,b_1,c_1,d_1\}$, g_2:$\{a_1, b_1, a, d_2\}$,$\cdots$,g_{64}: $\{a_3,b_3,c_3,d_3\}$。图匹配目的即为在图 4(b) 所示的图 G 中找到最佳匹配的候选同构图 g_m:$\{a_2,b_1,c_2,d_2\}$。ς表

示筛选与对应关键区域最相似的前 3 个候选关键区域，这些候选区域中，可能包含了纹理相似，但在表单上位置不同的图案区域，从而导致候选图中对应的结点出现较大幅度的位置偏差，如图 4(b) 中 b_2 和 d_2。因此，需要进一步度量图结构的相似度，寻找图 G 中与参考表单图 q 最相似的同构图，如 g_m，或去除误匹配结点的最相似的同构子图，如果 d_2 为误匹配结点，则目标匹配为同构子图 gs_m:$\{a_2,b_1,c_2\}$。

2.2　距离度量

将表单进行图表示和属性定义，然后通过度量 G 中同构图 g 和 q 间的属性差异，衡量两图间的距离，距离越小则表示子图 g 和 q 的结构越相似，根据属性的差异，确定最相似的同构图 g_m 或同构子图 gs_m。可以从以下几个方面度量图的差异。

1) 结点相似度

对 g 和 q 中结点 V^g 和 V 的 SIFT 特征点采取最近邻匹配进而得到匹配特征点对的 F-Score 值 $F_1(\boldsymbol{o}_i^g,\boldsymbol{o}_i)$，则图结点间的相似距离定义为

$$d_o(i)=1-F_1(\boldsymbol{o}_i^g,\boldsymbol{o}_i) \tag{4}$$

式中：$\boldsymbol{o}_i^g$、$\boldsymbol{o}_i$ 分别为 g 和 q 结点的纹理，$d_o(i)\in[0,1]$。

2) 结构相似度

夹角相似度。图中 v_i^g 和 v_i 的夹角相似度用其向量余弦相似度来表示：

$$d_\theta(i)=1-\cos(\boldsymbol{\theta}_i^g,\boldsymbol{\theta}_i) \tag{5}$$

式中 $d_\theta(i)\in[0,1]$。

权重相似距离，即以连接结点 v_i^g 和 v_i 所有边缘长度比，作为 g 和 q 中结点 v_i^g 和 v_i 权重相似距离；如果两个图结点处于完全对应的位置，则该结点相连接的对应边具有相同的相似比，这里将两图中对应边的长度比设为 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$，其中 $x_i=|e_{ij}^q|/|e_{ij}^g|$，$|\cdot|$ 表示边 e_{ij} 的长度，i 表示当前结点标号，j 表示 i 的邻接点，n 表示与结点 i 连接的结点数量。通过 $\boldsymbol{X}$ 的离散程度来计算图中对应结点所有连接边缘的整体相似距离：

$$\mathrm{sim}(\boldsymbol{\omega}_i^g,\boldsymbol{\omega}_i^q)=\exp^{-\frac{\mathrm{DX}}{\mathrm{EX}^2}} \tag{6}$$

式中：EX、DX 分别为变量 X 的均值和方差。则该图的权重相似距离为

$$d_\omega(i)=1-\mathrm{sim}(\boldsymbol{\omega}_i^g,\boldsymbol{\omega}_i^q) \tag{7}$$

式中 $d_\omega(i)\in[0,1]$。

同构图 g 和图 q 对应结点的相似度定义为

$$D(q(i),g(i))=d_o(i)+d_\theta(i)+d_\omega(i) \tag{8}$$

在进行参考表单 q 与 g 的图匹配时，考虑到 g 与 q 对应结点缺失或选择错误的情况，若 g 与 q 对应结点纹理极为相似，但实际位置并不匹配，较高的纹理相似度会对图的相似度有一定程度的干扰；同样的，当夹角、射线簇边缘相似度过高，同样会影响该结点的整体相似度的评判。故需对当前匹配的 g 中的结点剪枝，对结点 v_i 中 $d_o(i)$、$d_\theta(i)$、$d_\omega(i)$ 设置一定的阈值，不符合条件的 v_i 设值为离群点，同时将离群点纳入相似度量的整体评价中，即对 g 和 q 的子图进行匹配，寻找一个与 q 最相似的同构子图 gs_m。该离群点相似度度量如下：

离群点相似度，用经剪枝过后的离群点数量 (outlier Number, ON) 表示：

$$d_{\mathrm{Num}}=\frac{\mathrm{ON}}{N^q} \tag{9}$$

式中：$d_{\mathrm{Num}}\in[0,1]$，N^q 表示图 q 中结点 V 的数量，图 g 和图 q 的相似距离表示为

$$d(q,g)=\frac{3}{N^q-\mathrm{ON}}\sum_i^N c_i[d_o(i)+d_\theta(i)+d_\omega(i)]+d_{\mathrm{Num}} \tag{10}$$

式中：$c_i\in\{0,1\}$，0 表示离群点，1 表示符合阈值要求的结点，$d(q,g)$ 值越小则两图的相似度越大；故在 G 中，将与 q 相似距离最小的 g_m 或 gs_m 作为 G 与 q 的最终匹配结果：

$$D(q,G)=\mathrm{argmind}(q,g) \tag{11}$$

通过图相似性度量，得到与参考表单图最佳匹配的同构图 g_m 或同构子图 gs_m，图 5 给出了一个待处理热敏表单最佳匹配结果。

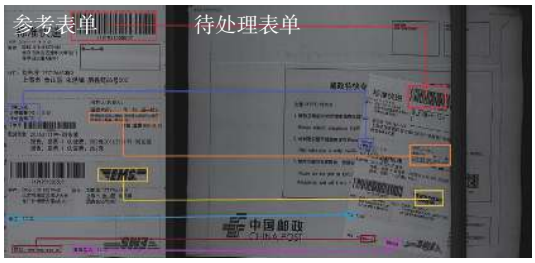


图 5　热敏表单图匹配结果

Fig. 5　Graph matching result for free form

2.3　待处理表单定位

如图 5 所示，参考表单与待处理表单的关键区域仅实现了部分对应，且匹配出的图像块不完整或轮廓不吻合，这是由于图像分割算法对复杂的字符图案分割不准确所致，这将直接导致表单提取的位置不准确。因此，本文在提取后处理过程中对匹配关键区域的位置进行修正，即迭代建立参考表单与待处理表单的位置映射函数，以此提高表单提取的准确性。通过映射函数，实现待处理表单上任意感兴趣区域的定位，从而完成表单信息的提取。

3 实验及分析

3.1 数据集

对快递包裹分拣机中采集的两类快递表单图像, 建立多联表单 (table like form, TF) 和热敏表单 (free form, FF) 两类实验数据集, TF 和 FF 共计 1 477 幅灰度快递表单图像。这些表单图像的分辨率偏转角度不同, 且未进行归一化处理。其中 TF 为表格类图像, 该类表单由制表单位统一印刷, 表单内容依据表格线布局, 包括中国邮政国内快递小包邮件详情单 (C-XB) 和 EMS 国内标准快递 (EMS-MULT), 这些图像的字符和图案较为清晰, 其中有部分图像具有褶皱、模糊、扭曲或缺损、遮挡或字迹重叠等问题。FF 数据集为非表格类表单图像, 该类表单常见于物流集散点、商家网点自行打印, 包括 EMS 标准快递 (EMS-FLAT) 和韵达快递表单 (YUNDA), 除存在上述 TF 数据集中特点以外, 该数据集中表单印刷墨迹清晰度不一。另外, 为验证算法在光照、尺度、旋转变换等情况下具有良好的鲁棒性, 本实验将 TF、FF 数据集记为 *o-i*, 对 *o-i* 进行了旋转、缩放、亮度调节等扩展数据集。旋转扩展是对 *o-i* 分别旋转 45°、90°、135°、180°, 新增 *r*-1、*r*-2、*r*-3、*r*-4 扩展数据集。缩放扩展是对 *o-i* 缩扩放至原表单图像的 75%、50%、125%、150%, 新增 *s*-1、*s*-2、*e*-1、*e*-2 扩展数据集。亮度调节扩展是对 *o-i* 的亮度提高至原来的 125%、150% 和降低至原来的 75%、50%, 新增 *b*-1、*b*-2、*d*-1、*d*-2 扩展数据集。经过数据集扩充, 本文实验的表单图像共计19 201 幅。

3.2 评价标准

本文通过表单图匹配的置信度和表单相关信息的提取结果准确率来分析算法的性能。

首先, 采用表单图匹配的置信度来衡量根据图匹配所建立的参考表单图像与待处理图像的映射是否可靠, 该置信度由重叠率 (average overlap, AO) 和平均准确率 (mean average precision, MAP) 来评定。如果映射的置信度高, 那么表单信息提取的准确性也会提高。重叠率定义为映射过程中关键区域重合度比例的均值:

$$\mathrm{AO}=\frac{1}{n_l}\sum_{i=1}^{n_l}\mathrm{overlap}(\mathrm{vgt}_i^q,v_i^q) \tag{12}$$

式中: n_l 为关键区域的数量, vgt_i^q为参考表单关键区域的位置, v_i^q为待处理表单图像上关键区域的定位结果, overlap(·)表示区域的重叠率。

MAP 是当重叠率 AO 高于某一阈值 T 时, 则待处理表单的匹配位置为准确位置, 故 MAP 表示为

$$\mathrm{MAP}=\frac{\mathrm{num}(\mathrm{AO}\geqslant T)}{I} \tag{13}$$

式中: num(AO≥*T*) 表示阈值为 T 时准确定位的图像数量, I 为测试图像的数量。

此外, 采用标注工具 LableImg 标记待处理表单中提取区域真值, 计算真值与检测目标交叠率 (intersection-over-union, IOU), 准确表示为

$$\mathrm{IOU}=\frac{\mathrm{DetectionResult}\cap\mathrm{GroundTruth}}{\mathrm{DetectionResult}\cup\mathrm{GroundTruth}} \tag{14}$$

其中, IOU DetectionResult 和 GroundTruth 表示信息提取区域检测位置和工具标注区域真值位置。

3.3 实验结果及分析

通过实验对 TF、FF 数据集分别计算了阈值 T 为 0.5、0.6、0.7、0.8、0.9 时图像的平均准确率和图像的平均重叠率 (mean average overlap, MAO)。当 T=0.8 时, 表示验证过程中参考表单和待处理表单中关键区域相互映射的重叠区域高于 80%, 实验表明: 此时用于定位的映射关系相对准确, 能实现大部分图像的准确定位和提取。因此本文实验将该阈值对应的 MAP 作为算法准确定位的置信度。

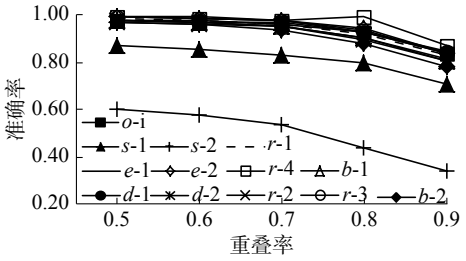
表 1 是 TF、FF 数据集的平均准确率和重叠率的实验统计情况, 其中 MAO 反映了样本中通过映射关键区域的整体重合情况。数据显示: TF、FF 中原图像数据集和扩展数据集的 MAO 主要分别在 90% 以上和 80% 以上, 说明根据图匹配建立的关键区域映射关系, 能较好的实现待处理表单与参考表单上关键区域的位置对应, 因此可以通过这种映射进行表单的提取。TF、FF 数据集的 MAP 大部分在 87%~98% 和 75%~86%, 这表明本文算法对多联表单和热敏表单具有良好的定位准确率。图 6 中, 当 T=0.9 时, FF 数据集的 MAP 相对 TF 数据集低约 20%~30%, 波动幅度较大, 原因有以下两点: 1)TF 数据集中, 关键区域均为表单出厂印制图案和字符, 同类表单的差异较小, FF 数据集表单印制要求不统一, 故而差异较大; 2)FF 数据集为非表格类表单, 其内容的自由度较大, 选取关键区域的难度较大, 可参照的关键区域少, 因此建立表单映射时严格匹配的特征点对较少, 因此对阈值高的 AO 的 MAP 值相对较低。图 6, TF 中原图像数据集在进行旋转、亮度调节变换后, 平均准确率的变化趋于重合, FF 数据集的平均准确率仅有小幅度范围内的波动。因此, 该表单提取算法对旋转和亮度变化的图像具有良好的稳定性。另外, 图 6 中图像缩至 75%, T=0.8 时, TF、FF 数据集的分别为 79.83%、70.11%, 与原图像数据集 *o-i* 相比 MAP 分别下降了 48.89%、19.54%, TF 数据集 *s*-2 与原图数据集 *o-i* 和其他

扩展数据集偏离幅度较大，FF 数据集也有明显的降低。出现这种变化的原因有：图像缩小比率过大时，表单图像上关键区域块纹理信息损失较多，这将导致图匹配时可参考的正确位置少，同时过度缩小的图像使得关键区域中对应的特征点位置出现偏差，建立表单的映射关系缺乏准确的参考点，则重合度偏差大，准确率下降，定位不准确。总体来说，算法对旋转、亮度调节、放大变换、小幅度缩小变换的表单图像的提取能保持良好的稳定性。

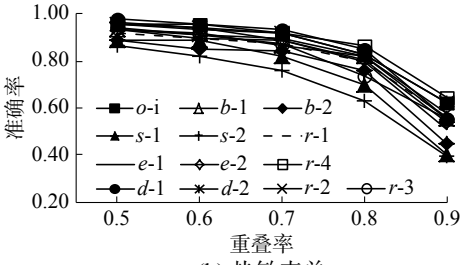
表 1　多联表单和热敏表单的平均重叠率和平均准确率

Table 1　Mean average overlap (MAO) and mean Average Precision (*MAP*) of TF and FF datasets

数据集		*o-i*	*b*-1	*b*-2	*d*-1	*d*-2	*e*-1	*e*-2	*s*-1	*s*-2	*r*-1	*r*-2	*r*-3	*r*-4
多联表单(TF)	MAP	0.927	0.905	0.876	0.926	0.904	0.939	0.945	0.798	0.438	0.922	0.898	0.929	0.989
	MAO	0.924	0.916	0.900	0.925	0.908	0.937	0.938	0.828	0.567	0.922	0.912	0.931	0.937
热敏表单(FF)	MAP	0.828	0.816	0.759	0.851	0.828	0.793	0.805	0.701	0.632	0.793	0.816	0.736	0.862
	MAO	0.855	0.850	0.815	0.878	0.855	0.832	0.847	0.801	0.772	0.846	0.849	0.833	0.870



(a) 多联表单



(b) 热敏表单

图 6　多联表单和热敏表单平均准确率

Fig. 6　Mean average precision (MAP) of TF and FF

本文实验通过计算提取结果与 LableImg 工具标记真值交叠率来评估定位的准确性。常见目标检测系统中常将 0.8 交叠率值作为正确检测阈值，本文在评估提取区域的准确率和平均交叠率时，这两组值变化趋势与映射置信度变化大致相似。因此，仅在表 2 中列出两类图像评估结果的平均情况。对比表 1 和表 2，说明图匹配结果越准确映射变换置信度越高，定位和提取的准确率越高。当 IOU 阈值为 0.8 时，多联表单和热敏表单提取准确率分别为 97.41% 和 83.93%，说明本文算法对这两类表单具有良好的定位与提取效果。

通过图匹配结果对待处理表单的候选关键位置进行修正，使参考表单到待处理表单的位置映射关系更加准确。通过对上述图匹配和映射后置信度的评估，验证了算法能对表单图像进行良好的定位。据此，图 7~10 所示为表单图像中用户感兴趣关键区域的定位与提取结果，其中图 7 和图 8 为 TF 类表单图像，图 9 和图 10 为 FF 类表单图像。图 7~10 中 (b) 图的提取结果自上往下分别表示提取的收货人地址、姓名、手机号。上述 4 组表单图像具有不同分辨率、亮度、方向偏转、面单褶皱和形变的差异，定位结果说明本文算法能适应不同图像质量差异和不同类别的图像。由于保证了准确定位的置信度，分割得到的表单区域的字符较为完整、清晰、准确。此外，对表单分割得到的图像块进行简单的字符连通域合并，得到图 7 中 4 组表单相关信息的提取结果。

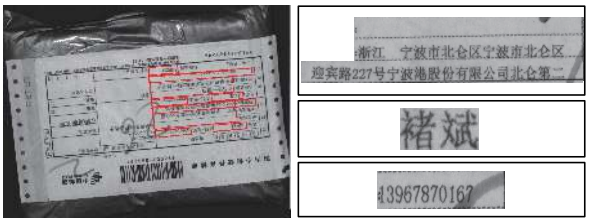
表 2　多联表单和热敏表单的提取准确率

Table 2　Extraction precision of TF and FF datasets

数据类别	准确率		平均交叠率
	IOU≥0.8	IOU≥0.9	
多联表单(TF)	0.974 1	0.864 5	0.934 8
热敏表单(FF)	0.839 3	0.667 6	0.816 6

本文方法与文献 [10, 13-14] 中方法类似，均为采用模板匹配的方法解决表单填写内容提取的问题，该方法的关键问题是实现参考表单和待处理图像配准。文献 [10, 13] 中采用傅里叶−梅林算法以表单局部区域或全局图像为配准目标，能实现不同方向的表单矫正，但该方法难以适应参考表单和待处理表单不同尺度的情况，不能准确找到表单图案的对应位置。此外文献 [13] 提取文本字符时的像素投票策略对图像噪声较为敏感，处理分拣机中现实采集到的污损和局部遮挡难以达到理想的提取效果。文献 [14] 中预先设定表单配准起始和终止参考点，作为表单方向校准的基准

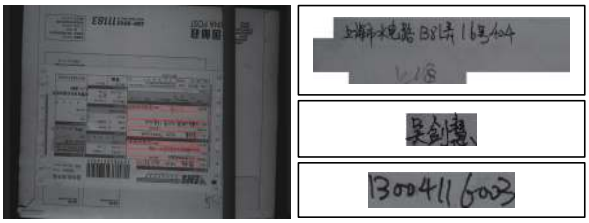
点, 该方法更适用于具有相同分辨率、亮度和对比度的扫描图像, 另外, 当基准点出现异物遮挡或缺损的情况难以灵活处理。本文方法采用表单图匹配的方法以解决上述处理过程中存在的不足, 根据不同表单已有的图案选取多个参考关键区域构建图, 采用图匹配的配准方式以解决单一参考基准点鲁棒性差的问题。此外图匹配配准方式能更好的适应不同尺度、方向、分辨率、光照条件的图像, 以及基准位置局部遮挡的问题。



(a) 定位结果　　(b) 提取结果

图 7　C-XB 表单定位和提取结果

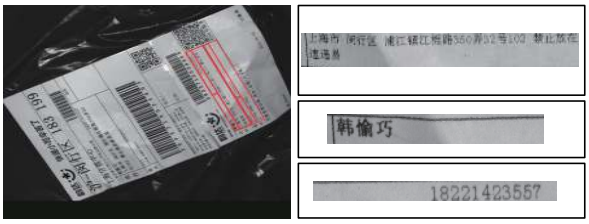
Fig. 7　Results for C-XB form Location and extraction



(a) 定位结果　　(b) 提取结果

图 8　EMS-MULT 表单定位和提取结果

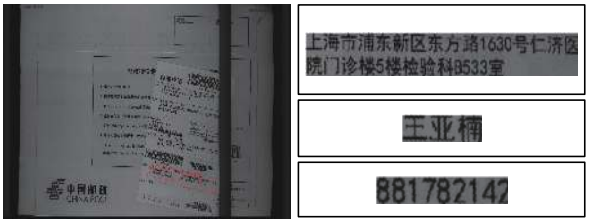
Fig. 8　Results for EMS-MULT form Location and extraction



(a) 定位结果　　(b) 提取结果

图 9　YUNDA 表单定位和提取结果

Fig. 9　Results for YUNDA form Location and extraction



(a) 定位结果　　(b) 提取结果

图 10　EMS-FLAT 表单定位和提取结果

Fig. 10　Results for EMS-FLAT form Location and extraction

4　结束语

本文提出了一种基于图表示和匹配的表单定位与提取方法, 实验表明: 本文方法适用于局部遮挡和不同类别、分辨率、方向、旋转、光照条件下的表单图像的处理, 是一种通用的表单图像准确定位和相关区域的提取方法。虽然本文方法实现了大部分表单图像相关信息的准确定位和提取, 但在缩小和单面形变幅度较大的图像上表现效果不佳, 下一步将考虑采用不同方法建立表单关键区域的映射, 以适应缩小比例大和较大范围形变图像的处理, 同时, 采用更为准确的后处理方法, 去除无关的空白区域, 使表单相关信息的提取精确到完整的字符串。

参考文献:

[1] SHARMA D V, LEHAL G S. Form field frame boundary removal for form processing system in Gurmukhi script[C]//Proceedings of the 10th International Conference on Document Analysis and Recognition. Barcelona, Spain, 2009: 256–260.

[2] CHEN J L, LEE H J. An efficient algorithm for form structure extraction using strip projection[J]. Pattern recognition, 1998, 31(9): 1353–1368.

[3] LIU Wenyin, DORI D. From raster to vectors: extracting visual information from line drawings[J]. Pattern analysis and applications, 1999, 2(1): 10–21.

[4] WATANABE T, LUO Qin, SUGIE N, et al. Layout recognition of multi-kinds of table-form documents[J]. IEEE transactions on pattern analysis and machine intelligence, 1995, 17(4): 432–445.

[5] LAM S W, SRIHARI S N. Multi-domain document layout understanding[C]//Proceedings of International Conference on Document Analysis and Recognition. 1991: 112–120.

[6] SACHDEVA R, SHARMA D V. Data extraction from hand-filled form using form template[J]. International journal on recent and innovation trends in computing and communication, 2015, 3(8): 5311–5317.

[7] NING L W, SIAH Y K, KHALID M, et al. Design of an automated data entry system for hand-filled forms[C]//Proceedings of 2000 TENCON. Kuala Lumpur, Malaysia, 2000: 162–166.

[8] BENSEFIA A. Extraction of Arabic handwriting fields by forms matching[J]. Journal of signal and information processing, 2015, 6(1): 53424.

[9] CESARINI F, GORI M, MARINAI S, et al. INFORMys: a flexible invoice-like form-reader system[J]. IEEE transac-

tions on pattern analysis and machine intelligence, 1998, 20(7): 730–745.

[10] CHO M, SUN Jian, DUCHENNE O, et al. Finding matches in a haystack: a max-pooling strategy for graph matching in the presence of outliers[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014: 2091–2098.

[11] SUH Yumin, ADAMCZEWSKI K, LEE K M. Subgraph matching using compactness prior for robust feature correspondence[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 2015: 5070–5078.

[12] SHARMA A, HORAUD R, CECH J, et al. Topologically-robust 3D shape matching based on diffusion geometry and seed growing[C]//Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA, 2011: 2481–2488.

[13] DUCHENNE O, JOULIN A, PONCE J. A graph-matching kernel for object categorization[C]//Proceedings of 2011 IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 1792–1799.

[14] ZHANG Quanshi, SONG Xuan, SHAO Xiaowei, et al. Attributed graph mining and matching: an attempt to define and extract soft attributed patterns[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014: 1394–1401.

[15] ZHANG Quanshi, SONG Xuan, SHAO Xiaowei, et al. Object discovery: soft attributed graph mining[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(3): 532–545.

[16] LEORDEANU M, SUKTHANKAR R, Hebert M, et al. Unsupervised learning for graph matching[J]. International journal of computer vision, 2012, 96(1): 28–45.

[17] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154–171.

作者简介：

谭婷，女，1992 年生，硕士研究生，主要研究方向为图像处理与计算机视觉。

吕淑静，女，1977 年生，高级工程师，博士，主要研究方向为图像处理与计算机视觉。

吕岳，男，1968 年生，教授，博士生导师，主要研究方向为模式识别、图像处理、生物特征识别、信息检索、数据挖掘、自然语言处理、机器视觉系统。9 次获得省部级科学技术奖，其中一等奖 4 次。授权发明专利 13 项。发表学术论文 100 余篇。