Классификация изображений таблиц на основе структурного подхода

Т.Н. Шолохова[1], Л.М. Местецкий[1]
tanja200596@mail.ru|mestlm@mail.ru
[1]МГУ им. М.В. Ломоносова, Москва, Россия

В работе предложен новый алгоритм преобразования документа с таблицами в электронный вид. Целью алгоритма является поиск таблиц на изображении документа, их классификация и определение их содержимого на основе обучающего набора таблиц.

Алгоритм основан на непрерывно-морфологическом подходе. Непрерывно-морфологический подход позволяет уйти от необходимости оперировать растровыми терминами и перейти к векторному описанию - линиями, формами и фигурами. В основе алгоритма лежит метод скелетизации.

Ключевые слова: *изображение таблицы, распознавание структуры, классификация, непрерывно-морфологический подход.*

Table images classification based on structure approach

tanja200596@mail.ru|mestlm@mail.ru
[1]Lomonosov Moscow State University, Moscow, Russia

In this work, we propose an algorithm of digital representation of printed table documents. The key steps of the algorithm are detection of tables on an image of a document and matching table structure to one of the predefined table templates.

The algorithm is based on the structure approach. This approach allows describing an image of a table in terms of its shape, morphological structure, and geometrical properties rather than in terms of rasterized image. The core routine of the proposed algorithm is skeleton method.

Keywords: *table image, structure recognition, classification, skeleton approach.*

1. Введение

Одной из важных функций систем электронного документооборота является перевод документов на бумажных носителях в электронную форму. Современные средства - фотокамеры, сканеры - обеспечивают ввод изображений документов в компьютер. Далее необходимо преобразовать изображение в форму электронного документа. Важной частью этого процесса является преобразование таблиц, так как таблицы наиболее часто используются для представления структурированных данных. Целью данной работы является получение эффективного решения описанной задачи.

Для задачи распознавания таблиц необходимо сгенерировать признаковое описание таблицы. Распознавание структуры таблиц на основе поиска горизонтальных и вертикальных линий описано в работах [3, 5]. В статье [6] предложен метод поиска горизонтальных и вертикальных линий на основе гистограмм проекций насыщенности пикселей, обобщение этого метода (с использованием алгоритма Хафа для обнаружения произвольных кривых на изображениях) описано в статьях [4, 7]. В рассмотренных статьях решается задача распознавания качественных отсканированных документов, поэтому получение качественного изображения для исходной, возможно искаженной, фотографии является отдельной важной частью этой работы.

Предложенный подход основан на анализе изображений таблиц с использованием непрерывных скелетов - внутреннего и внешнего [1, 2]. Внутренний скелет представляет собой множеств срединных осей бинарного изображения, в котором чёрные пиксели соответствуют линиям и текстам, а белые - бумажному фону. Внешний скелет - наоборот, есть множество срединных осей фоновой части изображения. На рис. 1 приведены примеры внутреннего (зелёные линии) и внешнего (красные линии) скелетов.

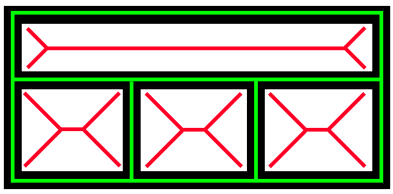
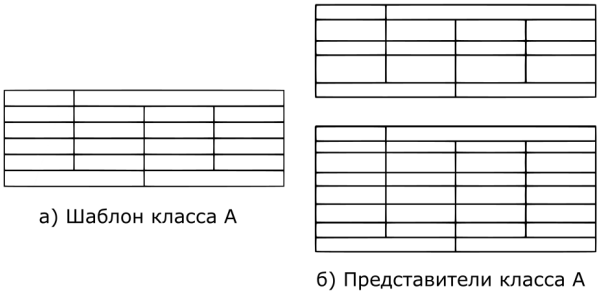


Рис. 1. Иллюстрация скелета

Основные идеи предложенного подхода:

1. Нахождение протяжённых вертикальных и горизонтальных отрезков на основе внутреннего скелета (замена алгоритма Хафа).
2. Отрисовка найденных отрезков на “чистой” странице в растровом формате (удаление лишних элементов - букв).
3. Нахождение ячеек на полученном изображении на основе внешнего скелета (компоненты связности и длинные срединные оси - ячейки).
4. Построение графа смежности ячеек. Граф смежности с информацией о вершинах ячейки - основа для оценки сходства структур таблиц.
5. Найденная структура используется для нахождения ячеек исходной таблицы. Найденные ячейки могут быть отправлены на распознавание в OCR.

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18



а) Шаблон класса А

б) Представители класса А

Рис. 2

2. Постановка задачи

Процесс идентификации таблиц состоит в сравнении исследуемого изображения таблицы с описаниями, заданными базой эталонов. Эталоны описываются заранее в процессе разметки.

Размеченные данные представляют собой качественные изображения, каждое изображение содержит ровно одну таблицу, в дополнение к изображению хранится файл разметки этой таблицы, который содержит информацию о местоположениях и типах ячеек.

Внутри одного класса таблиц может изменятся: наполнение таблицы, количество строк в таблице, высота строк в таблице. Не может изменяться: количество столбцов, ширина столбцов.

Размеченные данные будем называть шаблонными данными или шаблонными таблицами. На рис. 2 изображен пример шаблонной таблицы (слева) и её возможных вариаций (справа).

В процессе тестирования необходимо по данному изображению документа (который может содержать несколько таблиц) определить положения всех таблиц, классы найденных таблиц, координаты и типы ячеек.

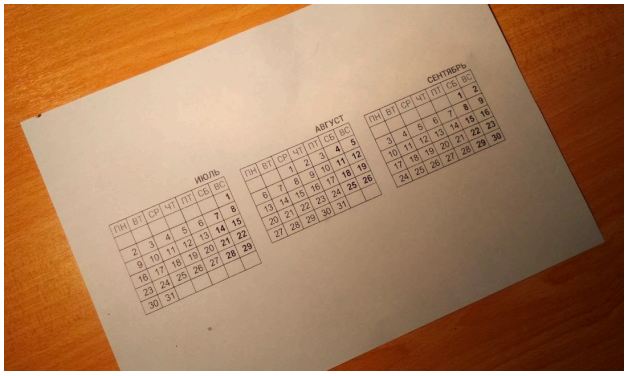


Рис. 3. Пример исходного изображения

На рис. 4 сверху изображена шаблонная таблица, ниже изображен бинаризованный и верно-ориентированный документ из рис. 3, содержащий три таблицы того же типа; на документе цветом отмечены найденные таблицы и ячейки, каждая ячейка соответствует некоторой ячейке шаблонного документа.

ЯНВАРЬ

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

а) Шаблон

ИЮЛЬ

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

АВГУСТ

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

СЕНТЯБРЬ

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

б) Результат работы

Рис. 4

3. Предобработка входного изображения

Первым шагом необходимо качественно бинаризовать исходную фотографию. Она может быть зашумлена и содержать искажения.

Для бинаризации из исходной фотографии вычитается фон, далее используется адаптивная бинаризация. Для получения фона изображения используется медианное сглаживание с большим радиусом окна (равным четверти высоты изображения).

Для получения верной ориентации используется следующий метод: на бинаризованном изображении строится внутренний скелет, каждое ребро которого имеет некоторую длину и некоторый наклон. На множестве полученных рёбер строится гистограмма углов наклона, каждое ребро имеет вес, равный его длине. На исходных изображениях преобладание по длине рёбер у линий-границ таблиц и максимум гистограммы соответствует преобладающему углу, следовательно при повороте изображения на преобладающий угол линии-границы таблиц станут параллельны осям координат.

На рис. 5 представлен пример предобработки документа, изображенного на рис. 3.

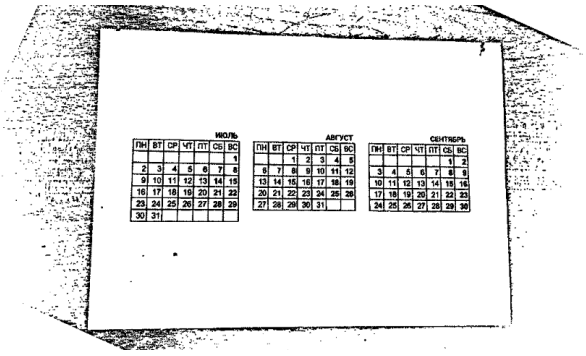


Рис. 5. Результат предобработки

4. Поиск линий и ячеек

Для поиска ячеек необходимо найти на изображении прямые линии, далее на основе линий найти связные ячейки таблицы.

Для поиска линий используется внутренний скелет изображения. Рёбра внутреннего скелета разделяются

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

на горизонтальные и вертикальные. Далее рёбра объединяются в линии в соответствии со следующим алгоритмом:

```
СОРТИРОВКА горизонтальных рёбер
           по координате левого конца
горизонтальные линии = {}
ЦИКЛ по горизонтальным рёбрам:
ЕСЛИ ребро по высоте соответствует одной из линий,
ТО добавить ребро к соответствующей линии
ИНАЧЕ создать новую линию
```

Аналогичный алгоритм используется для объединения вертикальных линий.

Для поиска ячеек на изображении остаются только линии, после этого выполняется построение внешнего скелета.

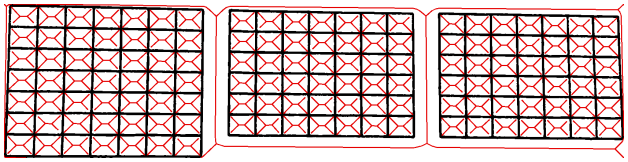


Рис. 6. Внешний скелет

Пример построенного внешнего скелета представлен на рис. 6. Все компоненты связности внешнего скелета, считаются найденными ячейками.

5. Описание структуры таблицы

После определения местоположений всех ячеек таблицы, необходимо классифицировать данную таблицу, определить к какой именно шаблонной таблице она относится. Для этого необходимо определить структуру таблицы.

Структурой таблицы, считается граф, вершинами которого являются ячейки, а рёбра соответствуют смежности ячеек. При этом определяются два типа рёбер: горизонтальная связь и вертикальная связь.

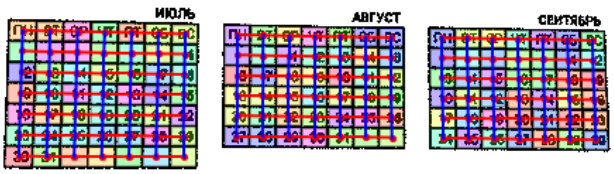


Рис. 7. Структура документа

На рис. 7 представлена полученная структура. Вершины, отмеченные красными точками, расположены в центрах найденных ячеек, горизонтальные рёбра изображены красным цветом, вертикальные — синим.

Дополнительными структурными признаками являются количество столбцов и их ширина. Для определения этих признаков все горизонтальные компоненты графа (все строки таблицы) преобразуются в численные вектора: $col_i = (w_1, w_2, \dots,)$, в которых w_k — соответствует длине очередного ребра (ширине очередного столбца). Далее, из всех векторов $col_1, col_2, \dots,$ находится вектор, который встречается в наборе наибольшее количество раз, назовём его повторяющимся вектором либо повторяющейся строкой. Повторяющийся вектор сохраняется и включается в описание таблицы. Количество столбцов определяется длиной такого вектора.

6. Классификация

Для определения к какому шаблонному типу принадлежит каждая таблица в документе для всех шаблонных таблиц подсчитываются структуры. Так как размерности признаковых пространств для разных таблиц могут быть разными (в качестве признака хранится ширина столбцов, разное количество столбцов означает разную размерность вектора признаков), используемый алгоритм классификации — метод ближайших соседей со специально введённой метрикой для таблиц.

Так, после вычисления структуры таблица T описывается следующим набором:

$$\begin{aligned} describe(T) = [& \\ comp_h(T) &= \{K_h^1, K_h^2, \dots, K^{|T_h|}\}, \\ comp_v(T) &= \{K_v^1, K_v^2, \dots, K^{|T_v|}\}, \\ cols(T) &= [w_1, w_2, \dots, w_{|T_{cols}|}] \quad] \end{aligned}$$

где K – компонента, которая описывается парой чисел: количество вершин и количество рёбер в ней;

$comp_h$ – множество горизонтальных компонент,

$comp_v$ – множество вертикальных компонент,

$cols$ – список размеров столбцов, упорядоченный в порядке столбцов и нормированный на ширину всей таблицы.

В метрику включены величины, описывающие близость таблиц. Например, мощность пересечения множеств горизонтальных компонент одной таблицы с множеством горизонтальных компонент другой таблицы:

$$\frac{|comp_h(T_1) \cap comp_h(T_2)|}{\min(|comp_h(T_1)|, |comp_h(T_2)|)} \cdot \gamma_1$$

То же для вертикальных компонент:

$$\frac{|comp_v(T_1) \cap comp_v(T_2)|}{\min(|comp_v(T_1)|, |comp_v(T_2)|)} \cdot \gamma_2$$

При равенстве количеств столбцов, евклидово расстояние между наборами размеров столбцов:

$$\sqrt{(cols(T_1) - cols(T_2))^2} \cdot \gamma_3$$

. Так как ширина столбцов нормирована на ширину всей таблицы, все вектора будут расположены внутри шара радиуса 1. Значит евклидово расстояние между любой парой векторов не будет превышать 2, то есть в случае разного числа столбцов можно использовать константу 2 в качестве большого расстояния между таблицами. γ_i — параметры классификации настраиваются по валидационной выборке.

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

7. Определение типов ячеек

При выделении ячеек таблиц на тестовых документах необходимо восстановить положения ячеек тестовых таблиц и попытаться соотнести их с ячейками шаблонных. Могут возникнуть ошибки двух типов: некоторая ячейка не распознана совсем, либо несколько ячеек объединились в одну.

В структуре тестовой таблицы подсчитывается количество строк и высота каждой строки. Считается, что таблицы могут иметь несколько строк-заголовков и несколько строк-окончаний, отличающихся от повторяющейся строки, именно количество повторяющейся строки может изменяться в тестовой таблице по сравнению с шаблонной. (Строками здесь называются горизонтальные компоненты).

Для восстановления положения ячеек предложен следующий подход: шаблонная таблица преобразуется в соответствии с подсчитанными в тестовой таблице параметрами, после этого в явном виде помещается на место тестовой таблицы. После этого все размеченные ячейки шаблонной таблицы переходят в ячейки тестовой таблицы.

8. Эксперимент

Эксперимент проводится на изображениях финансовых документов.

Эксперимент состоит из двух частей: оценка качества классификации таблиц и оценка качества метода определения положений ячеек.

Качество данного классификатора оценивается на выборке таблиц, часть из которых является качественными отсканированными документами, вторая часть является фотографиями с возможным шумом и искажениями. Метрика качества — точность (accuracy).

Количество эталонных размеченных таблиц равно 10. Для каждой шаблонной таблицы есть ≈ 5 качественных изображений с различными вариациями структуры. Итого ≈ 40 качественных изображений. Каждое качественное изображение было распечатано и сфотографировано ≈ 5 раз. Итого ≈ 200 зашумленных изображений.

Полученное качество классификации таблиц $\approx 93\%$, на отсканированных документах качество достигло 98%, на фотографиях 87%. Ошибки связаны с недостаточным качеством предобработки, либо с похожими таблицами в шаблонной выборке и невозможностью различия их исключительно по структуре.

Для оценки качества метода определения положений ячеек используются только те документы, для которых верно предсказаны типы всех таблиц. Ячейка считается распознанной верно, если отношение совместной площади к площади пересечения с найденной ячейкой $> 90\%$ и совпадает тип. Метрика качества — точность (accuracy).

Полученное качество определения ячеек 82%, на отсканированных документах качество достигло 96%, на фотографиях 69%. Ошибки связаны с недостаточным качеством предобработки, так как фотографии могут содержать проективные искажения, либо искажения второго порядка, в таком случае наиболее частая ошибка — это неверное определения числа строк в таблице.

Основной проблемой в задаче осталась проблема правильной предобработки изображений.

9. Заключение

В ходе работы предложены и реализованы следующие методы:

- метод описания структуры таблицы;
- метод классификации таблиц на основе структуры (качество 93%);
- метод определения положений ячеек таблиц (качество 82%);

10. Благодарности

Работа выполнена при поддержке РФФИ грант 17-01-00917.

11. Литература

[1] Местецкий Л.М. Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. Москва, Физматлит, 2009.

[2] Местецкий Л.М. Скелет многоугольной фигуры — представление плоским прямолинейным графом. Санкт-Петербург, ГРАФИКОН-2010.

[3] Fan K-C, Wang Y-K, Chang M-L. Form document identification using line structure based features. Guilin, Chinese control and decision conference, CCDC'09, 2001.

[4] Liolios N, Fakotakis N, Kokkinakis G. On the generalization of the form identification and skew detection problem. Pattern Recognit, 2002.

[5] Liu J, Jain AK Image-based form document retrieval. Pattern Recognit, 2000.

[6] Mandal S, Chowdhury S, Das A, Chanda B. A hierarchical method for automated identification and segmentation of forms. Guilin, Chinese control and decision conference, CCDC'09, 2005.

[7] Ohtera R, Horiuchi T. (2004) Faxed form identification using histogram of the Hough-space. Guilin, Chinese control and decision conference, CCDC'09, 2004.

Об авторах

Местецкий Леонид Моисеевич, д.т.н., профессор кафедры математический методов прогнозирования факультета вычислительной математики и кибернетики Московского государственного университета. e-mail: mestlm@mail.ru.

Шолохова Татьяна Николаевна, студентка магистратуры кафедры математический методов прогнозирования факультета вычислительной математики и кибернетики Московского государственного университета. e-mail: tanja200596@mail.ru.

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18