УДК 004.932.2

А.В. Гороховатский, Е.О. Передрий

Харьковский национальный экономический университет, Харьков

АВТОМАТИЗАЦИЯ ВЫДЕЛЕНИЯ ЗАГОЛОВКА ПО ИЗОБРАЖЕНИЮ ДОКУМЕНТА

Статья посвящена разработке метода локализации заголовка на изображении документа. Предложено использование эмпирических признаков, которые выделяют заголовок среди основного текста, и в комбинации с применением проекционного преобразования дают возможность выделить область расположения заголовка. Экспериментальные исследования подтвердили эффективность предложенного метода и позволили определить области его применимости.

Ключевые слова: *локализация заголовка, документ, изображение, блок, проекционное преобразование.*

Введение

Задача локализации заголовка на изображении документа часто является главной для анализа структуры документа и его содержимого с последующим распознаванием [1 – 3], например, с целью автоматического индексирования документации.

Выделяют два направления методов анализа структуры документов: физический и логический [4], первый из которых решает задачу разбиения изображения на текстовые или графические блоки, второй – позволяет выделить структурные элементы документа, такие как текст, заголовок, аннотация, список литературы, фамилии авторов, номера страниц и т.д.

В данном исследовании мы фокусируемся лишь на задаче логического анализа определения заголовка по изображению документа.

Как логический, так и физический анализ документа может быть выполнен одним из двух наиболее распространенных подходов [4 – 6]: «снизу вверх» («bottom-up» – анализ мелких элементов изображения с последующим их слиянием, образуя при этом элементы более высокого логического уровня: блоки, параграфы, абзацы и т.п.), либо «сверху вниз» («top-down» – проекционное разбиение изображения на абзацы, блоки с их последующим анализом и разбиением) [4, 5, 7].

Статья посвящена разработке метода для автоматической локализации заголовка по изображению документа с использованием проекционного разбиения и системы эмпирических признаков.

Эмпирические признаки заголовка и предварительная обработка

Заголовок в большинстве документов характеризуется некими общими признаками, к которым можно отнести следующие:

– заголовок чаще может иметь больший размер шрифта, чем основной текст;

– заголовок может быть написан прописными символами;

– заголовок обычно располагается перед основным текстом (в верхней части документа);

– заголовок часто выделяется пустой строкой/строками;

– заголовок обычно занимает всю ширину документа и не разбивается на колонки, в отличие от основного текста;

– длина заголовка обычно составляет не более 3-5 строк;

– последняя строка заголовка часто содержит меньшее количество слов, чем предыдущие.

Приведенные признаки являются наиболее явными из существующих, характеризуются точными количественными критериями и могут быть относительно легко верифицируемы в процессе анализа изображения документа.

Традиционными методами предварительной обработки изображений являются эквализация гистограммы, улучшение контраста и яркости, удаление шумов, нормализация поворота и других искажений аффинной и проективной групп и т.д.

Необходимым этапом предварительной обработки изображений, содержащих текстовую информацию, чаще всего является бинаризация, т.е. сведение изображения к двум градациям яркости, которые соответствуют пикселям текста и пикселям фона.

В данной работе использована классическая однопороговая бинаризация по уровню δ единиц яркости, заданному пользователем (принято значение по умолчанию $\delta = 100$, эффективность которого подтверждена экспериментальным путем).

Горизонтальное проектирование и формирование блоков

Предварительная обработка изображения документа позволяет создать благоприятные условия для работы с текстом на изображении, поскольку

необходимо выделить заголовок, который, в свою очередь, является частью текста.

Наиболее известным методом для определения положения линий текста на изображении является использование горизонтального проектирования.

В качестве функционала проектирования выбран сумматор яркости пикселей вдоль горизонтальных линий сканирования, что соответствует выполнению преобразования Радона [8 – 10] при угле проектирования $\theta = 0°$. Анализируя значения полученной одномерной проекции, можно определить местоположение линий текста (рис. 1).

Каждый из блоков, содержащих строку текста, можно описать в виде триплета $B =< S, F, H >$, где S – координата начала блока (соответствует номеру горизонтальной линии проектирования на изображении), F – координата окончания блока, $H = F - S$ – высота блока (рис. 2).

На этапе поиска линий возможно также игнорирование верхнего колонтитула документа (который чаще содержит техническую информацию о документе), если заранее известно о его наличии. Совокупность всех блоков B может содержать как блоки с текстом, так и блоки с другими элементами изображения. Для выделения текстовых блоков выполняется их отсеивание. В качестве признака, характерного для текстовых блоков, был выбран один из наиболее простых в реализации, а именно – количество разрывов $m \geq M$ на горизонтальной медианной линии сканирования $med = S + H/2$, которая проходит через середину блока (по умолчанию принято $M = 25$).

Размер шрифта основного текста можно грубо оценить как среднее значение H_{avg} всех высот текстовых блоков, что позволяет выделить потенциальные блоки, которые имеют большую величину H и, соответственно, могут быть частью заголовка. Назовем такие блоки кандидатами в строки заголовка. Максимальное количество N потенциальных блоков заголовка задано априори (принято по умолчанию $N = 30$), как и максимальное количество строк K, из которых может состоять заголовок ($K = 5$).

Анализ блоков на предмет соответствия заголовку

Определим заголовок документа как $k \leq K$ блоков-кандидатов, идущих подряд в верхней части изображения и имеющих высоту $H > H_{avg}$ при выполнении условия $|H - H_{max}| < \varepsilon$, где значение ε было определено экспериментально в виде $\varepsilon = 0.1 H_{max}$.

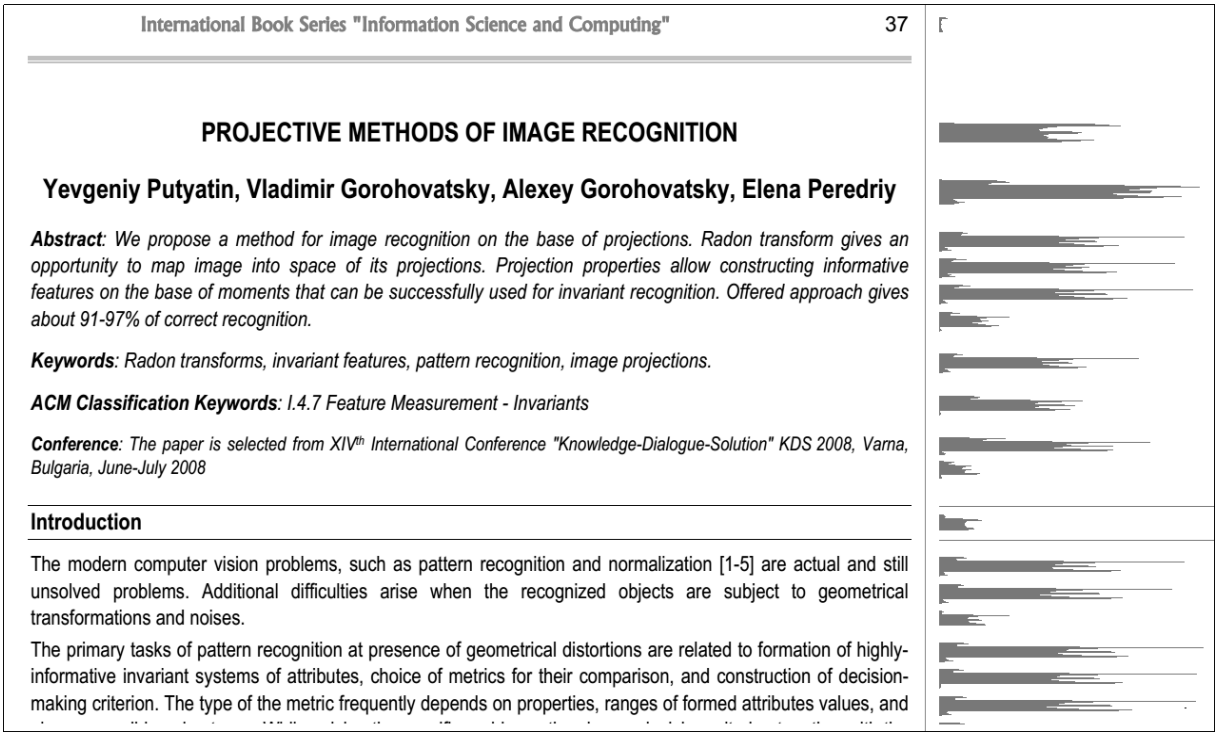

Рис. 1. Блоки горизонтального проектирования

Рис. 2. Параметры блока

Во время анализа блоков-кандидатов могут возникать неопределенные ситуации, связанные с особенностями реализации процедуры сопоставления параметров блоков между собой. Результатом неточного анализа может быть неправильное определение строк заголовка, т.е. $k \neq k^*$, где k^* – истинное априорное количество строк заголовка.

Рассмотрим несколько ключевых ситуаций, которые могут возникать при различных количествах строк в заголовке k^*.

1. При $k^* = 1$ заголовок состоит из одной строки, остальные строки, как правило, отличаются другими характеристиками. Детектирование заголовка не вызывает сложностей, за исключением случая, описанного ниже в п. 3.

2. При $k^* > 2$ имеем несколько строк, обладающих признаками заголовка. В этом случае предложено проводить дополнительный анализ расстояния между строками заголовка, которые обычно отличаются от расстояния после заголовка. Каждое из расстояний должно соответствовать предыдущему, т.е. $l_{1,2} \approx l_{2,3} \approx ... \approx l_{k-1,k}$, где $l_{i,j}$ – расстояние между i и j линиями заголовка соответственно.

3. Случай $k^* = 2$ является достаточно распространенным, поскольку во многих документах заголовок состоит именно из двух строк. Затруднение также представляет тот факт, что часто вторая строка заголовка содержит меньшее количество слов, что делает невозможным сравнение по значениям величин проекций. В этом случае мы также не можем использовать сравнение с предыдущим межстрочным расстоянием ввиду его отсутствия.

Аналогичная проблема также возникает при $k^* = 1$ и $H_1 \approx H_2$, при этом строка, идущая за заголовком, может быть ложно отнесена к части заголовка. На рис. 3 показан пример ложной локализации заголовка, состоящего из единственной строки.

UDC 004.932.2

IMAGE INVARIANT RECOGNITON METHODS IN THE SPACE OF PROJECTIONS

A.V. Gorokhovatskyi

Image recognition methods based on invariant features of projections are investigated. Influence of geometrical distortions on Radon space is shown, systems of invariant features for recognition based on Radon transform and classical moment invariants are suggested. Computer modeling results on the real images are shown.

Keywords: Radon transform, projection, invariant features, moment invariants, recognition.

INTRODUCTION

Artificial intelligence applications in the tasks of the automatic image processing, analysis and pattern recognition attract greater attention of researchers. It can be explained by the development of computer technologies, which enables the decision of complex practical problems, related to the image analysis and interpretation [1]. Solution of similar tasks often becomes difficult because of external influences, different background and local covering by noise components. One of methods, which can avoid these problems is Radon

$$\mathrm{Rec}: I \to B_E,\ \mathrm{Rad}: B \to I,\ B_{input}: B_E \to B,$$

where Rec is a recognition operator, Rad is Radon transform operator, $I = \{I_1, I_2, ..., I_k\}$ – is the system of invariant features, $B_E = \{B_1, B_2, ..., B_n\}$ – is the set of etalons, $B = \langle a_1, a_2, \varphi, \alpha, \beta \rangle$ – is an input image under the action of distortions or transforms.

The result of the most widespread geometrical transforms influence (movements, rotating and scaling)

Рис. 3. Пример неудачной локализации строки заголовка при $k^* = 1,\ k = 3$

Рассмотрим постфактум сравнение межстрочного расстояния $l_{1,2}$ с величиной $H_1 + H_2$ при достижении условия $k = 2$ в процессе анализа блоков-кандидатов. Выполнение условия $l_{1,2} > H_1 + H_2$ говорит о том, что величина межстрочного расстояния между первой и второй строками является значительной. Это дает возможность отбросить ложные идентификации строк, имеющих высоту, близкую к высоте строк, входящих в заголовок.

На рис. 4 показан результат обработки изображения, представленного на рис. 3, с предложенной модификацией. Как можно увидеть, использование рассмотренного условия позволяет отделить ложные определения строк текста.

Результаты экспериментов и применение метода

Для оценки качества предложенного метода реализовано автоматическое определение координат заголовка (с параметрами, выставленными по умолчанию и приведенными выше) с его последующим распознаванием на основе Tesseract OCR[11], поскольку сравнения априорного количества строк k^* с полученным результатом k недостаточно ввиду

возможных ошибок локализации.

На тестовом наборе изображений статей количеством 100 штук реализация предложенного метода дала возможность правильно детектировать и распознать заголовок на 78 изображениях. 12 изображений тестового набора были локализованы частично, т.е. либо была корректно найдена часть заголовка, либо весь заголовок с лишними строками до или после него.

Наиболее распространенными ошибками являются: локализация в качестве заголовка других строк документа, имеющих похожий размер; ложные пропуски коротких строк заголовка (вследствие нарушения условия $m \geq M$).

UDC 004.932.2

IMAGE INVARIANT RECOGNITON METHODS IN THE SPACE OF PROJECTIONS

A.V. Gorokhovatskyi

Image recognition methods based on invariant features of projections are investigated. Influence of geometrical distortions on Radon space is shown, systems of invariant features for recognition based on Radon transform and classical moment invariants are suggested. Computer modeling results on the real images are shown.

Keywords: Radon transform, projection, invariant features, moment invariants, recognition.

INTRODUCTION

Artificial intelligence applications in the tasks of the automatic image processing, analysis and pattern recognition attract greater attention of researchers. It can be explained by the development of computer technologies, which enables the decision of complex practical problems, related to the image analysis and interpretation [1]. Solution of similar tasks often becomes difficult because of external influences, different background and local covering by noise components. One of methods, which can avoid these problems is Radon transform (RT), that presents an image as a set of projections [1-3]. The advantages of this method are the high level of Radon transform features informing, good noise immunity (because of integral properties), as well as

$$\mathrm{Rec}: I \to B_E,\ \mathrm{Rad}: B \to I,\ B_{input}: B_E \to B,$$

where Rec is a recognition operator, Rad is Radon transform operator, $I = \{I_1, I_2, \ldots, I_k\}$ – is the system of invariant features, $B_E = \{B_1, B_2, \ldots, B_n\}$ – is the set of etalons, $B = \langle a_1, a_2, \varphi, \alpha, \beta \rangle$ – is an input image under the action of distortions or transforms.

The result of the most widespread geometrical transforms influence (movements, rotating and scaling) on RT $R(p,\theta)$ can be estimated as [4]:

$$R(p,\theta,a_1,a_2) = \int\int B(x',y')\delta(p - x'\cos\theta - y'\sin\theta + a_1\cos\theta + a_2\sin\theta)dx'dy'$$

Рис. 4. Пример определения строки заголовка при $k^* = k = 1$ и модификацией $l_{1,2} > H_1 + H_2$

На последних 10 тестовых изображениях предложенный метод локализовал другие элементы документа вместо заголовка. Как правило, это вызвано сложной структурой верхней части документа, наличием изображений, логотипов, технической информации.

Реализация рассмотренного метода дает возможность оценить области его применимости. К тем свойствам изображения документа, которые позволяют в большинстве случаев удачно локализовать заголовок на изображении, можно отнести следующие:

– размер шрифта заголовка значительно превосходит размер основного текста;

– наличие пустых строк до строк заголовка и после него;

– начало документа с заголовка.

Следующие признаки затрудняют корректную локализацию заголовка:

– наличие изображений и других объектов в верхнем колонтитуле документа;

– подобие размера шрифта заголовка к размеру шрифта других элементов документа в верхней его части;

– надстрочные и подстрочные символы в строках верхней части документа.

Выводы

Предложенный в исследовании метод дает возможность локализовать заголовок документа по его изображению на основе проекционного анализа, а также предопределенных признаков, которые идентифицируют заголовок в документе. Преимуществом данного метода является низкая вычислительная сложность, реализация метода позволяет проводить автоматическую локализацию заголовков в режиме реального времени.

Использование параметров по умолчанию позволило распознать около 78% тестовых изображений, экспериментальные исследования показали, что этот показатель может быть улучшен за счет изменения параметров по умолчанию (таких, как δ,

ε, K, M, N). К преимуществам данного метода также можно отнести его простую масштабируемость и адаптацию к обработке изображения документов различного типа.

В качестве перспективных улучшений предложенного метода можно рассматривать использование и анализ дополнительной информации о заголовке, например, длину текстового блока на изображении.

Список литературы

1. Breuel T. Layout analysis based on text line segment hypotheses [Электронный ресурс] / T. Breuel // 3rd Int. Workshop on Document Layout Interpretation and its Applications (DLIA) DLIA'03. – 2003. – Режим доступа к ресурсу: http://www.science.uva.nl/events/dlia2003/program/ 23-26-breuel.pdf

2. Fu Chang. Chinese Document Layout Analysis Using An Adaptive Regrouping Strategy [Электронный ресурс] / Fu Chang, Shih-Yu Chu, Chi-Yen Chen // Pattern Recognition, 38(2). – 2005.– P. 261-271.– Режим доступа к ресурсу: http://www.iis.sinica.edu.tw/papers/fchang/1566-F.pdf. – Название с экрана.

3. Breuel T. High Performance Document Layout Analysis [Электронный ресурс] / T. Breuel // Symposium on Document Image Understanding (SDIUT), 2003 April 9-11, Greenbelt, MD.– Режим доступа к ресурсу: http://iupr1.cs.uni-kl.de/~shared/publications/2003-breuel-highperformancedo cum.pdf.– Название с экрана.

4. Smith R. Hybrid Page Layout Analysis via Tab-Stop Detection [Электронный ресурс] / R. Smith // 10th International Conference on Document Analysis and Recognition (ICDAR), 2009, 26-29 July.– P. 214-245.– Режим доступа к ресурсу: http://dejanseo.com.au/research/ google/35094.pdf. – Название с экрана.

5. Nagy G. Hierarchical representation of optically scanned documents / G. Nagy, S. Seth // International Conference on Pattern Recognition (ICPR), Montreal, 1984.– P. 347-349.

6. Faure C. Extracting the tables of contents from the images of documents [Электронный ресурс] / C. Faure // Режим доступа к ресурсу: http://perso.telecom-paristech.fr/~cfaure/articles/riao00.pdf. – Название с экрана.

7. Namboodiri A. Document Structure and Layout Analysis [Электронный ресурс] / Anoop M. Namboodiri, Anil K. Jain // Digital Document Processing. – 2007. – P. 29-48. – Режим доступа к ресурсу: http://pdf.aminer.org/000/348/003/document_page_segmentation_and_layout_analysis_using_soft_ordering.pdf.

8. Geometric Layout Analysis Techniques for Document Image Understanding: a Review [Электронный ресурс] / R. Cattoni, T. Coianiz, S. Messelodi, C.M. Modena // ITC-irst Technical Report TR#9703-09. – 1998. – Режим доступа: http://www.ee.bgu.ac.il/~dinstein/stip2002/ Seminar_papers/ David_Cahana_Geometric%20Layout%20Analysis%20Techniques %20-%20 a%20Review.pdf. – Название с экрана.

9. Learning Logic Programs for Layout Analysis Correction [Электронный ресурс] / M. Berardi, M. Ceci, F. Esposito, D. Malerba // In proceeding of: Machine Learning, Proceedings of the Twentieth International Conference (ICML), August 21-24, 2003, Washington, DC, USA. – P. 27-34. – Режим доступа к ресурсу: http://www.di.uniba.it/~malerba/publications/icml03.pdf

10. Gorokhovatskyi A.V. Image invariant recognition methods in the space of projections / A.V. Gorokhovatskyi // Прикладна радіоелектроніка. – 2010. – Том 9, № 4. – С. 574–576.

11. Tesseract-OCR [Электронный ресурс].– Режим доступа к ресурсу: http://code.google.com/p/tesseract-ocr/.– Название с экрана.

Поступила в редколлегию 30.11.2012

Рецензент: д-р техн. наук, проф. Е.П. Путятин, Харьковский национальный университет радиоэлектроники, Харьков.

АВТОМАТИЗАЦІЯ ВИДІЛЕННЯ ЗАГОЛОВКУ ЗА ЗОБРАЖЕННЯМ ДОКУМЕНТУ

О.В. Гороховатський, О.О. Передрій

Статтю присвячено розробці методу локалізації заголовку на зображенні документа. Запропоновано використання емпіричних ознак, які відрізняють заголовок від основного тексту, і, в комбінації із застосуванням проекційного перетворення дають можливість виділити область місця розташування заголовка. Експериментальні дослідження підтвердили ефективність методу, який було запропоновано, та дозволили знайти умови його застосування.

Ключові слова: *локалізація заголовка, документ, зображення, блок, проекційне перетворення.*

AN AUTOMATIZATION OF TITLE DETECTION ON IMAGE OF A DOCUMENT

O.V. Gorokhovatskyi, O.O Peredrii

Paper is devoted to the construction of the method of localization of the title of a document on the image. The using of empirical features, which distinguish title from main text, in combination with projection transform allows an opportunity of detection region with title on image. Experimental investigations confirmed an efficiency of the suggested method, and showed the conditions of applicability.

Keywords: *title localization, document, image, block, projection transform.*