ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ
УНИВЕРЗИТЕТ У БАЊОЈ ЛУЦИ

ПРОНАЛАЖЕЊЕ РИЈЕЧИ У СКЕНИРАНОМ ТЕКСТУ БЕЗ ОПТИЧКОГ ПРЕПОЗНАВАЊА ЗНАКОВА (*WORD SPOTTING*)

Студенти: Игор Јанковић,
Славко Марковић

Број индекса: 66/03
13/04

Предметни наставник: Проф. др Зденка Бабић
Предметни асистент: мр Владимир Рисојевић

Број бодова:

Бања Лука
Јун, 2009. године

ПРОНАЛАЖЕЊЕ РИЈЕЧИ У СКЕНИРАНОМ ТЕКСТУ БЕЗ ОПТИЧКОГ ПРПОЗНАВАЊА ЗНАКОВА (*WORD SPOTTING*)

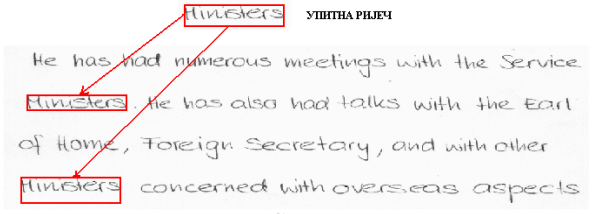
Игор Јанковић, Славко Марковић
E-mail: igorr.jan@gmail.com, slavkomarkovic@gmail.com

1. САЖЕТАК

Рад описује метод проналажења ријечи у скенираном тексту, како куцаном, тако и у рукопису. Алгоритам за проналажење ријечи, имплементиран у MATLAB-у, састоји се из два дијела. Први дио се односи на сегментацију текста на слике ријечи и њихово индексирање. Затим се од корисника захтијева да изабере ријеч која ће касније послужити као упит. Други дио се односи на креирање дескриптора за сваку слику ријечи, на основу којих вршимо поређење датих ријечи са упитом. Као дескрипторе користимо моменте. За поређење ријечи користимо Танимотов коефицијент, који представља проширење косинусне сличности. Као резултат извршавања добијамо листу ријечи које највише личе упитној ријечи.

2. УВОД

У последње вријеме, проналажење ријечи у старим рукописима поприма све већу пажњу. Велика количина информација се налази у разним рукописима и старим књигама, а савремено друштво захтијева брзо и лако приступање истим. Досадашњи резултати у овом пољу су далеко од идеалних, али много тога је учињено. Истраживања у овој области су углавном базирана на OCR рјешењима, који је погодан за штампане документе. Међутим, рјешења базирана на OCR-у зависе од писма које се користи у одговарајућем документу. Поред тога, два човјека која користе исто писмо различито га руком записују, тако да у овом случају OCR не даје жељене резултате, па се морамо окренутим ка другим методама проналажења ријечи.



Сл.1.

Као алтернатива рјешењима базираним на OCR-у јавља се *word spotting* [4]. *Word spotting* представља метод базиран на проналажењу сличности између ријечи. Пошто обично један човјек пише одговарајући документ, претпостављамо да је варијација између слика мала. Овом методом се проналази сличност између слика ријечи и упитне ријечи дефинисане од стране корисника (слика 1).

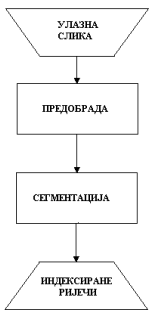
3. ПРЕТХОДНИ РАДОВИ

Уочавање ријечи у скенираном тексту последњих година поприма све већу пажњу. Рукописни документи су често лошег квалитета и могу бити деградирани различитим неповољним факторима, а за разлику од штампаних докумената, постоје варијације у начину на који су ријечи написане. Због тога, традиционалне *Optical Character Recognition (OCR)* технике, које обично препознају ријечи карактер по карактер, имају значајне пропусте када се примјењују на рукописе.

За проналажење сличности између слике ријечи која представља упит и слике ријечи у документу, користи се *Dynamic Time Warping (DTW)* [1], који се примјењује у обради говорног сигнала. Приступ је обећавајући када су у питању енглески писани документи, док за *Hindi* документе *DTW* не постиже задовољавајуће резултате. Такође, приступи засновани на *DTW*–у су спори. Рјешења заснована на филтрима [12], показују добре резултате за униформне величине и типове фонтова, али нису у могућности да прате варијације фонтова и транслацију.

4. ДИЈАГРАМ ТОКА

Алгпоритам се може подијелити у двије логичке цјелине, представљене на следећим дијаграмима тока. Први дијаграм приказује поступак сегментације и индексирања ријечи, док се други односи на креирање упита и издвајање ријечи.



Сл. 2.1

Сл.2.2

5. ИМПЛЕМЕНТАЦИЈА

5.1. Моменти

Геометријски моменти реда $(p+q)$ за континуалну функцију $f(x,y)$ се дефинишу као:

$$M_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x,y)dxdy \qquad (1)$$

гдје су $p,q = 0,1,2,...\infty$. Наведена дефиниција представља пројекцију функције $f(x,y)$ на $x^p y^q$. Сет геометријских момената сачињен од свих M_{pq}, за које вриједи $p+q <= n$, садржи $\frac{1}{2}(n+1)(n+2)$ елемената [2]. У нашем случају, гдје функција $f(x,y)$ има само двије могуће вриједности 0 и 1, једначина (1) постаје:

$$M_{pq} = \sum_X \sum_Y x^p y^q f(x,y) \ , \qquad (2)$$

гдје X и Y представљају x, y координате слике.

Геометријски моменти дефинисани са (1), нису инваријантни на транслацију, ротацију и промјену скале. Међутим, централни моменти дефинисани са:

$$M_{pq} = \sum_X \sum_Y (x-\overline{x})^p (y-\overline{y})^q f(x,y) \ , \qquad (3)$$

гдје су:

$$\tilde{x} = \frac{M_{10}}{M_{00}}, \tilde{y} = \frac{M_{01}}{M_{00}} \ , \qquad (4)$$

координате центара гравитације слике, су инваријантни на транслацију координата. Централни моменти се могу представити као линеарна комбинација M_{jk} и момената нижег реда. Варијансе момената се дефинишу као:

$$\sigma_x = \sqrt{\frac{\overline{M}_{20}}{M_{00}}}, \sigma_y = \sqrt{\frac{\overline{M}_{02}}{M_{00}}} \ . \qquad (5)$$

Коришћењем нормализованих координата:

$$x^* = \frac{x-\overline{x}}{\sigma_x}, y^* = \frac{y-\overline{y}}{\sigma_y} \ , \qquad (6)$$

добијамо једначину момената:

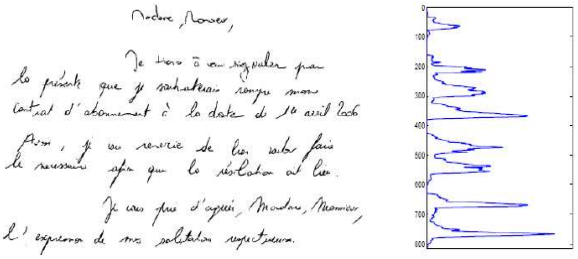
$$m_{pq} = \frac{\sum_X \sum_Y (x^*)^p (y^*)^q f(x,y)}{M_{00}} \ , \qquad (7)$$

која је инваријантна на транслације слике и промјену скале.

5.2. Сегмантација текста

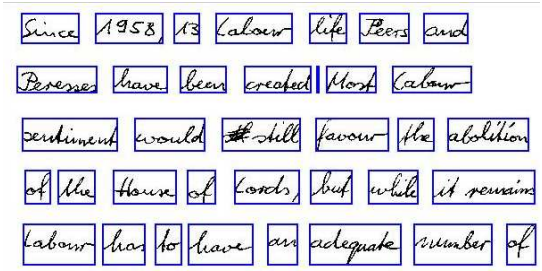
Сегментација линија текста врши се коришћењем хоризонталног профила, а затим слиједи одређивање вертиклног профила, којим се издвајају појединачне ријечи слике. Хоризонтални профил се одређује сумирањем црних пиксела по врстама матрице која репрезентује дату слику. С обзиром да црни пиксели имају вриједност 0, а бијели 1, слику је претходно потребно „инвертовати“. Издвајање линија текста врши се поређењем добијене суме, која представља хоризонталну пројекцију, са прагом одлучивања, за сваку врсту матрице. Треба напоменути да се праг одлучивања мијења у зависности од врсте текста, као и величина и типова фонтова. Уколико је вриједност суме већа од задатог прага, поставља се доња граница. Затим, слиједи испитивање да ли је дата вриједност мања од прага. Када се то деси одређена је горња граница линије текста.

Поступак се наставља док се не дође до краја врсте. На сљедећој слици, приказан је рукопис за који је одређен хоризонтални профил.



Сл. 3

Слично, као и у претходном случају, за одређивање вертикалног профила врши се сумирање црних пиксела слике, али по колонама. Ова техника је позната под називом *break cost* (*Tsuimoto*, 1991.). Колоне у којима „цијена коштања“ има минималну вриједност су кандидати за мјеста на којима ће дата слика бити сегментирана. Уколико се ради о штампаном документу, карактери ће бити раздвојени, па је у овом случају потребно одредити размаке између појединих карактера. Издвајањем максималне вриједности размака, одређује се оптимална вриједност прага и у обзир узимају само они размаци који најмање одступају од оптималне вриједности. На тај начин се врши сегментација ријечи у документу. Примјер рукописа у којем је извршена сегментација, приказан је на слици 4.



Сл. 4

5.3. Издвајање ријечи

Користећи једначину (7) рачунамо моменте до седмог реда, јер за моменте већег реда прецизност опада [1]. За сваку ријеч добијамо вектор који се састоји од 36 вриједности момената. Ове вриједности представљају дескрипторе на основу којих поредимо слике ријечи.

Корисник креира упит, десним кликом на ријеч која је претходно индексирана. Затим програм врши поређење упита са осталим индексираним сликама ријечи. За одређивање сличности користили смо Еуклидову дистанцу и косинусну метрику сличности. С обзиром да је косинусна сличност давала боље резултате, одлучили смо се за ову метрику. Поређење се врши користећи Танимотов коефицијент:

$$T(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{a^2 + b^2 - \vec{a}\cdot\vec{b}},$$

који представља проширење косинусне сличности.

Вриједности Танимотовог коефицијента су у интервалу од -1 до 1. Издвајање ријечи се остварује поређењем вриједности Танимотовог коефицијента са претходно усвојеним прагом одлучивања.

6. РЕЗУЛТАТИ

Већ смо напоменули да су моменти инваријантни на транслацију, ротацију и линеарну промјену скале. Овдје ћемо демонстрирати иваријантност на транслацију и типове фонтова.

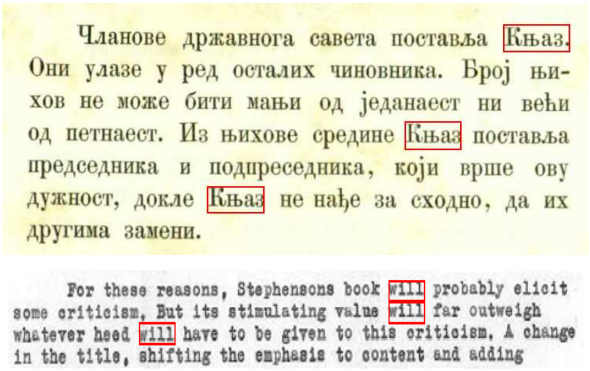
а) б) в)

Сл. 5

Парови слика ријечи	а)	б)	в)
а)	1	1	0.9628
б)	1	1	0.9628
в)	0.9628	0.9628	1

У претходној табели дати су вриједности Танимотовог коефицијента између парова слика ријечи за ћирилично писмо.

Сада ћемо на неколико примјера демонстрирати рад програма. Прво дајемо резултате за штампане скениране текстове.

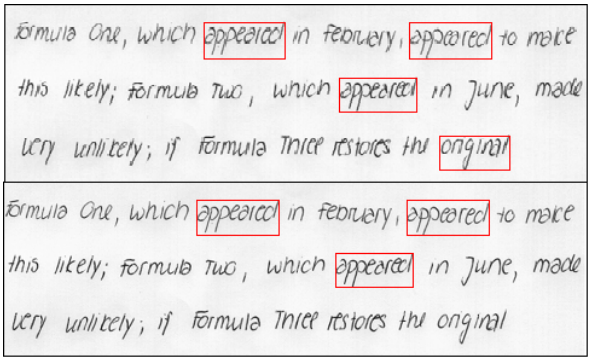


Чланове државнога савета поставља Књаз. Они улазе у ред осталих чиновника. Број њихов не може бити мањи од једанаест ни већи од петнаест. Из њихове средине Књаз поставља председника и подпреседника, који врше ову дужност, докле Књаз не нађе за сходно, да их другима замени.

For these reasons, Stephensons book will probably elicit some criticism. But its stimulating value will far outweigh whatever heed will have to be given to this criticism. A change in the title, shifting the emphasis to content and adding

Сл. 6 *Скенирани штампани документи*

На претходним сликама видимо да програм даје очекиване резултате на постављени упит. Сада дајемо резултате за ручно писане документе.



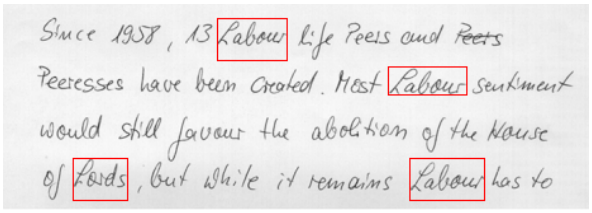
formula One, which appeared in February, appeared to make this likely; formula two, which appeared in June, made very unlikely; if formula three restores the original

formula One, which appeared in February, appeared to make this likely; formula two, which appeared in June, made very unlikely; if formula three restores the original

Сл. 7 *Скенирани рукопис*

На слици 7 је дат примјер за који је потребно извршити корекцију прага одлучивања. Повећањем прага одлучивања одбацујемо ријеч која не одговара упиту.

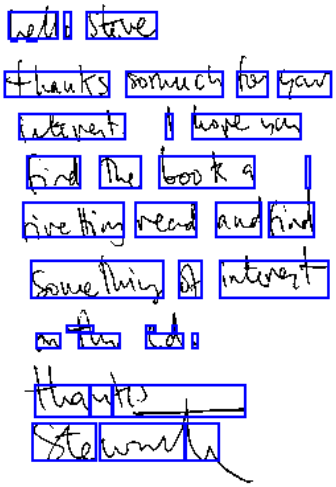
Међутим, постоје случајеви када корекција прага одлучивања не даје очекиване резултате.



Сл. 8 *Случај када повећање прага одлучивања не даје резултате*

У примјеру на слици 8, Танимотов коефицијент има већу вриједност за ријеч *Lords* него за једну од ријечи *Labour*. Повећањем прага одлучивања, у овом случају, избацићемо и ријеч која одговара упиту.

Постоје рукописи у којима је велика варијација између ријечи, тако да наш алгоритам не показује добре резултате у погледу сегментације, а самим тим није могуће креирати упит. Овај случај је илустрован на следећој слици.



Сл. 9

7. ЗАКЉУЧЦИ И БУДУЋИ РАД

Овај рад показује да су резултати за штампане скениране текстове и рукописе у којима постоји мања варијација између ријечи задовољавајући. До одступања долази у случајевима када су документи оштећени или у случају лоших рукописа.

Недостатак рада представља ручно подешавање прагова одлучивања за различите врсте докумената и величине фонтова, као и недовољна предобрада документа за сегментацију. Рад се може побољшати аутоматизацијом одређивања прагова одлучивања, како за сегментацију, тако и за рачунање сличности између ријечи. Поред тога, потребно је имплементирати алгоритам за квалитетнију предобраду, као што су *pruning* техника, која се заснива на одбацивању кандидата на основу величине оквира, и *Hugh*-ова трансформација за косо скениране текстове.

7. ЛИТЕРАТУРА

[1] A. Bhardwaj, D. Jose, V. Govindaraju, *Script Indenpended Word Spotting in Multilingual Documents*, CUBS, New York, USA, 2008.

[2] R. C. Papademetriou, *Reconstructing with Moments*, School of System Engeenering, Portsmouth, London

[3] S. Kane, A. Lehman, E. Partridge, *Indexing George Washington's handwritten manuscripts*, Center for Intelligent Information Retrieval, Computer Science Department, University of Massachusetts, Amherst, MA 01003

[4] R. Manmatha, W. B. Croft, *Word spotting: Indexing Handwritten Archives*, University of Massachusetts, Amherst, USA, 1999

[5] A. Andreev, N. Kirov, *Word image matching in Bulgarian historical documents*, [SEEDI Communications 1], Review of the National Center for Digitization, 8 (2006), 29–35.

[6] Дигитална Народна библиотека Србије, http://www.digital.nbs.bg.ac.yu/

[7] IAM Database for Off-line Cursive Handwritten Text, http://www.iam.unibe.ch/~zimmerma/iamdb/iamdb.html

[8] IAM Database for Off-line Cursive Handwritten Text, http://www.iam.unibe.ch/~zimmerma/iamdb/iamdb.html

[9] Javier Montenegro Joo, *Boundary geometric moments and its application to automatic quality control in the Industry*, Instituto de Física de Sao Carlos (IFSC), Dpto. de Física e Informática, Universidade de Sao Paulo (USP), Brazil, 2006

[10] *Moments and Moment Invariants in Image Analysis*, Jan Flusser, Barbara Zitová and Tomáš Suk, Institute of Information Theory and Automation, Prague, Czech Republic

[11] T. M. Rath, S. Kane, A. Lehman, E. Partridge and R. Manmatha, *Indexing for a Digital Library of George Washington's Manuscripts: A Study of Word Matching Techniques*, Center for Intelligent Information Retrieval Computer Science Department University of Massachusetts Amherst, MA 01003

[12] Huaigu Cao, Venu Govindaraju, *Template-FreeWord Spotting in Low QualityManuscripts*, the Sixth International Conference on Advances in Pattern Recognition(ICAPR), Calcutta, India, 2007.