The Fourth SEEDI Conference Digitization of cultural and scientific heritage Belgrade, Serbia, June 12-15, 2008

A software tool for searching in binary text images

Nikolay Kirov Kirov

Computer Science Department, NBU and Institute of Mathematics and Informatics, BAS Sofia, Bulgaria Abstract: In this paper we present a software tool for searching word images in scanned text documents. We consider that the document pages are represented as files in tif, jpg, gif, png, bmp and other graphic file formats. Our experiments prove the efficiency of the proposed approach and show that such type of search could be successful. Examples of using various languages are presented. The software is user oriented and can be applied to any collection of scanned documents.

1/22

This research has been partially supported by a Marie Curie Fellowship of the European Community program "Knowledge Transfer for Digitalization of Cultural and Scientific Heritage in Bulgaria" under contract number MTKD-CT-2004-509754.

Introduction

Optical character recognition (OCR) is the usual way of conducting text retrieval from scanned document images.

It converts text images into a text file, recognizing every symbol and mapping it to a number, which is called code.

The most often used codes are ASCII (one byte code) or UTF-8 (two bytes code).

This technique is well developed and has high accuracy.

Searching words in a text file is relatively easy task.

But sometimes OCR is a very difficult process requiring dictionaries in the corresponding languages. Often human efforts are needed to correct OCR errors which is quite tedious work. There are some obstacles to successful OCR:

- The quality of page images.
- Language dependency (alphabet and coding, unknown language):
 - dictionaries;
 - old grammar, obsolete words and phrases and idioms;
 - old letters, outside of the coding tables;
 - multi-lingual documents;
- Errors in automatic OCR, human intervention needed.

Nicolas de Bonnefons, Ch. de Sergy, (1692), University of Gent, Digitized by Google (2007)

de la Terre, &c. Fort peu de sens commun.

Quant à la Terre, si vous la rencontrez bonne, ce vous sera un grand avantage,& une grande épargne ; mais rarement en pourrez-vous trouvez,où il n'y ait beaucoup à travailler, dautant que telle paroîtra passablement bonne au dessus, qui étant ouverte de la profondeur d'un fer de Béche seulement, se trouveraj Argileufe desfous ; ce fonds' est pire aux Arbres que le Tuf, ou la Roche, à cause qu'il s'y rencontre de petites veines où les Racines peuvent s'étendre & profonder, afin de tirer la fraîcheur de plus bas, & prendre quelque noutriture; mais l'Argileuse ou Terrre franche ou rouge, fait comme un plancher qui par sa dureté & densité, ne peut être percé par aucunes Racines, & qui dans les grandes ardeurs de l'Eté, em-

nȕl)uanrΓ¤, la l'erre,^! vouz la rencomrex t,onne,ce vouz lera un ^ranF¤ avanra^e,sc u/B»e ^ranF¤B« epar^ne ; maiz raremenr en pourrex-vouz trouvex, on il n'v air tieaucouri a travailler , claur'nc 'ue teile riarorrra rrallälllernenr cionne au clelluz,c^ am ouverre cle la rirokoncleur F¤'un iec Twe Lecrre ieulemenr , re rrouveral ^r?lleule cleclouz : ce fonclz eli riire aux ^r-PMrez c^ue le l"uf,on la l^t,cke,a caule c^u'il z'^ reuconrre cle r/erirez veiuez on lez ^.acinez peuvenr z'ccenf¤resc rirofoncler,arlB« cle rirer la ir,uclleurcle^luz baz,sc rirenF¤re ^ uelc^ue nourricure; malz l^r^ilcule ou ' I'errre francrre ou rou^e, fair comme un plancrrer c^ui riar ia F¤urere sc clenlrre , ne peur ecre rierce riar aucunez li.acinez, sc c^ui cl^nz lez ^r,inclez arcleurz cle l'^,te,emB» We suggest a different approach: instead of applying two steps – OCR and searching in text documents, we will directly search words in scanned text documents.

We can organize retrieval of words, similar to a given pattern word, searching in the binary text images.

The document pages can be represented as binary images in any graphic file format.

Input data of the software are collection of files representing text document.

The software

The software system supports three user windows:

- Main
- Parameters
- Found.

Main window displays one page of the document – the current page. The current directory containing document image files and the current file name are given on the top of Main window. It is possible to go forward and backward through the document pages.

Searching in binary text images (version 0.2, 04.2008)	? ×
Parameters Segmentation Find Directory Book2b File p001.tif Next Pressure	revious
Но и едните и другите със своето жайсторство са	_
пренасали душите на своите слушатели в друг мир. С живит	•
хороводни нелодии то са карали непринудено хората да игр	6.#7
хора и ръченици, карали са със своите гласови възможност	H
да забранят делинчните трудности, като същевреженно са л	од
държали будно националното съзнание посредством българси	X78
налеви и словесна съдържавие.	-

Parameters				<u>?</u> ×
-Clean				
		Cle	an 3	÷
-Segmentation				
 always 		0	optional	
Min row height 60	÷	Margin	5	÷
Row white 10	÷	Row space	20	÷
Min length word 80	÷	Shrink white	1	-
Recognition				
Adjustment				
Method SHD	•	PHD: f=	0.90	÷
Diff. length 20	÷	a =	0.10	÷
Max dist 25	÷	$\beta =$	0.01	÷
Words found 20	÷	M-HD: τ =	4	÷
		ок	Cano	:el

Find	<u>? ×</u>
Go to page Book2b/p001.tif Save YES 2	8 0.43
Пазарджик Пазарджик Пазарджик Пазарджик Пазарджик Пазарджик Пазарджик Пазарджик Пазарджик Пазарджик Пазарджик	
Пазаріжни Киевската	-

Segmentation

We use horizontal projection for line extraction. If the lines are horizontal straight lines or have small slopes, the histogram has near zero values between lines.

To segment the words in a line, we use vertical projection – a histogram obtained by counting the number of black pixels in each vertical scan at a given horizontal position. If the words are well separated, the histogram should have areas of zero values between words.

As a result of the segmentation, every word is associated with a word image – minimal rectangular frame that contains the corresponding word. So we consider any word image as a rectangle, which consists of white and black pixels. The black pixels form a set, which is used in calculating word similarities.

на пред близки негови клиенти, които много обичали да го слушат. Тремолирането на дясната му ръка е било ненадминато

For segmentation step we use a number of parameters, which are important for successful word separation:

- Minimal row height: This paramter helps us to avoid creating (due to noise) rows with small height;
- Margin: This paramter allows us to process only a part of the page.
- Row white: When the value at a point in row histogram is less than the value of this parameter, we suppose that this point belongs to the white space between the words.
- Row space: Minimum value of the white space between words.
- Minimum word length: The system does not segment words with length less than the value of this parameter. Shrink white: This parameter concerns additional step conducted when we have already separated words, and words are framed. At this step we try to shrink the height of rectangles if the points of histograms have values less than this parameter.

Searching

After segmentation of a page, we must choose a pattern word image – this is a word, which we want to find in the document pages.

This step starts when the user push the button Find in Main window. It activates the process of inspection all pages to measure similarities of segmented words and the pattern word.

We can see a part of the retrieval data in Find window. Pushing GoTo button, the page containing the marked word downloads in Main window.

The program code is written in C++ with help of Qt - a cross-platform application development framework (open soutce from Trolltech).

Experiments

Bulgarian typewritten document (about 1940), 335 pages, tif (2400×3200) , 1 BPP

I РАЗДЕЛ

ПЕВЦИ И МУЗИКАНТИ ПРЕДИ И СЛЕД ОСВОБОЖДЕНИЕТО

Историята на музиката в гр.Пазарджик започва от незапомнени времена и достига до наши дни, като резултат от дейността на знайни и незнайни труженици, които, кой повече, кой по-малко, е внесъл в нейната съкровищница своя дял.

От материалите, с които разполага Окръжния исторически музей – Пазарджик, респективно сведенията, които е събрал БОРИС ХАДЖИ РАШКОВ от гр.Пазарджик,относно певци и музиканти преди и след Освобождението се установява, че битовите нужди, свързани с годежи, сватби, занаятчийско-еснафски сбирки, хора, вечеринки и пр. са били задоволявани от музиканти – професионалисти и любители.

Професионалисти били онези музиканти-инструменталисти или певци, като най-често инструменталиста е бил и певец, които са свирили и пеели срещу възнаграждение, а любителионези, които със своето пеене и свирене са радвали душите и сърцата на хората по сборове, хорища и др., без да получават възнаграждение.

Но и едните и другите със своето майсторство са пренасяли душите на своите слушатели в друг мир. С живите хороводни мелодии те са карали непринудено хората да играят хора и ръченици, карали са със своите гласови възможности да забравят делничните трудности, като същевременно са поддържали будно националното съзнание посредством българските напеви и словесно съдържание.



Bulgarian book, Christomatia (1884), 1000 pages, tif (2300×3800), 8 BPP

Повече-то отъ ранни-тѣ му стихотворения ск любовни пѣсни, по подражание на гръцки-тѣ, и не прѣдставлявать лигературна стойность; стихотворения-та му въ "Смѣсна Китка" при всичко, че повечето ск слаби подражания на руски-тѣ, нъ свидѣтелствовать вече за поетическо-то дарование на г. Славейкова; най-добри-тѣ му стихотворения ск обнародвани-тѣ по-послѣ въ "Читалпште," отъ конто "Не пѣй ми се," "Жестокость-та ми се сломи" и "Тогасъ понѣ" дкхать съ истински лиризмъ и заслужено привлѣкоха внимание-то на читателе-тѣ. Славейковъ, който е вѣшть въ българский езикъ, пръвъ доказа гквкость-та му въ поезия-та. Като се числи между първи-тѣ борци по черковний въпросъ, той захвашта въ скшто-то врѣме почтенно мѣсто въ редъ-тъ на малко-то ни добри литератори.

Велико влияние е упражнила възъ пробуждане-то духъ-тъ камъ свобода-та на независимость-та у българский народъ доста общирна-та литературна д'ятелность на Георгий Сава Раковски (род. въ Котель 1818, умр. въ Букурешиь 1868 г.). Въ личность-та и въ дела-та на Раковски се отрази най-нагледно тогавашно-то състояние на умове-тѣ, нужди-тѣ, стремления-та и идеали-тѣ на народъ-тъ ни. Тако-речи едничькъ дѣець по онова врѣме, той писува, работи всичко. Той искаше да обгърне въ своя-та широка дѣягелность вслчкитв нужди на народъ-ть ни, да удовлетвори всички-тв национални купнѣяния, да осжштестви най-смятни-тѣ и въжделени мечти. Той възсъздаде съ фанатический въсторгъ минжло-то и приготви бидиште-то. Б'єше въ скшто-то врёме поеть, историкъ, етнографъ, нублицисть, агитаторъ и хайдутинъ. Нито на единъ български дъятель животъ-тъ не е билъ напълненъ съ толкова неутолима и разнообразна деятелность и напьстренъ съ толкова беди, приключения и странности. Той се бѣше училъ въ Атина, Парижь, Цариградъ и въ Руспя. Знаеше руский, сръбский, румхнский, турский, гръцкий, староелинский, французский, арабский и дори отъ чясти санскрит-



Old Greek text (approx. in the third century BC), 50 pages, jpg (1077×1416) , 8 BPP

τοιαύτα κατά σχήμα πάντη είσιν άτρεπτα, οία μ Σωμα δε σύνθετον έν τη φύσει έδεν τοιθτον άπ βλίψει, τομή όπωσθν ύπείκει, η λίθακες οι ζεβρ τητι διαφέρων άδάμας, μηδενός όλως έξαιρυμένυ, μοιρον, ώς δέδεικται. Ων δε σωμάτων τα μέρη κει, η βάζα διαχωρίζεοθαι πέφυκεν άπαλά ταυτ μέλι, άργιλος, κτ: ° δσω δε ήττονι η άτονωτέρο πλοκής άποζείχει, τοτήτω η άπαλώτερα, ές άκρ έδενι γάρ έζιν έντυχειν, ου τα μέρη μή όπωσθν

Γ΄. Το σκληρον σώμα ύπο κύφης τε ή άσξεν πεφυκός, εύ β αυζον άκύει τοι αῦτα χάλυψ τὰ κεράμεια σκεύη Τέτων τὰ σερεὰ μέρη ἐχ ἐτω c ἐλήλοις, διο ή βặσα τῆς ἀμοιβαίας παφῆς ἀφίσατι Δ΄. Το ἐκ πολών οίονει λεπίδων πάνυ λεπτῶν ἀ σάμενον, εὕσχισον σώμα προσείρηται τέτε τὰ τ μόρια σεβρότερον προσκεκόληται ἀλήλοις, ἡ λεπὶς Γ ἔρα τὰ τοι αῦτα σώματα ἐις λεπίδας ἀναλυόμενα· καλέμενοι λίβοι οἱ ἐξ Γβηρίας, ή Καππαδοκίας,

Searching in binary text images (version 0.2, 1	04.2008)		' ×
Parameters Segmentation Find	Directory greek	File 070.tif Next Previous	
10 11 1 10 L	when uses and so	1185 A OI CHEMIGUI ONWATONISS,	
V@71V XVEU UN	Find	^{II} × aXiss σωματis é	
bynn Bilgcugle	Go to page greek/070.tif Save	VES 1 0 VAI 9 B TUVANTION	
ε άπο των με	TWILLTO	🗎 тё Врахитате и	
σαιρέσεως υλ	<i>COLATIN</i>	0 TŨV Ê L Q U C N T	
Autoreouters -	d BILLETE MUNAUM	win & august 94	
and to Denne	<i>C</i> WLATA		
TOIL OF OWILIC	OLTROY	IDETA BION YOUTO	_
to symmetry sin	παρΕεωι	TLOV; B ELOVINU O	
όλου πόσμου ι	SUYALLY Contract	is y is ra asiax	
74 81 a 22 4 2 w	TOINUTA TOINUTA	λέγμε! (2)	
B'. $\Omega^{\gamma}_{y\tau}$	UTAVTA	R. SUGROAWS R. H	
DE ELENWORZON	δέπατε	à Kateitai, 2 àv	
Ra de Romina	Reuteoa	ANTEN ATOTICEST	
la oc tolauta	TAUTAC		
σεμια δυναμις	σωματι Company	reautal, y orwar	
Cupata Tay	Sumina	a a kiel 3 2 a v T i 1	
·····	TWLATA		
(1) П UXVI	TEQUEEN	- has a why ois . Itanon	
ar a Andar y	yeion then the sal sorge	elas mus ra profia reindae. that g	
Tegi Xerv, to de	iRelwess (Ages. iv 1	roig megi mois). (2) ("qu rapa 1	-
l	•	<u> </u>	

կառակունեան ոգին ամեն աստիճանի մարդոց սրը, որին մեջ տարածուեցաւ, և սորվեցուց անոնց մեկ զմեկ ատել ու մեկզմեկե գարչիլ. մինչև անգամ մեկ խանունի մեջ գործող արչեստաւորը սորվեցաւ նղովը տալ իր բովի խանունին մեջ բանող դրա. ցիին, ան պատճառով որ անիկայ _ոգւցն Արբոյ բղիսումը իրեն չամաձայն չդաւանիր. ո՛չ մեկը և ո՛չ մեկայը չասկընալով նե ի՞նչ կ՝ըսեն, կամ ի՞նչ բան կ'ուզեն չաստատել:

Ուստի այսպիսի անտեղի վիճարանուԹիւնները պատճառ տուին որ երրոր մարդիկ Հոգւոյն Որդց կրայզը խորճին, գրեԹե հիայն աս մեկ նիւԹիս ուղզեն միտբերնին, այսինըն Թէ՝ Հոգւցն բղխու մը միայն Հօրմեն է, կամ Հօրմեն ու Որդիեն։ Նմենն ալ կը դաւանին Թե Հոգին Սուրը՝ Նր րորգուԹեան մեկ անձն է. բայց ո՞վ կրնայ ըսել Թէ անիկայ ինչ ներդործուԹիւն կ՛ընէ մարդուս Հոգւցն փրկուԹեանը մարդս երկինըը բարձրացընե լու Համար: Նյա աս մեծ և ամենա Հարկաւոր նպատանիս Համար է որ ՆրրորդուԹեան մարդաւ պետուԹիւնը յայոնուած է։ Հայրը խրկեց Որդին աշխարճը փրկելու, հնչու որ Նստուած անանկ սիրեց աշխարճը մինչև որ իր միածին Որ-



Text in Spainish (1901), 30 + 57 pages, gif (1400×2500), 4 BPP Alonso, Rogelio M., Cartilla histyrico-descriptiva del tйrmino municipal de Macuriges. Habana: Impr. La Propagandista, 1901, HOLLIS Catalog, Harvard University

No señor; todas las fincas azucareras tienen sus chuchos que conectan con las líneas del ferrocarril y hay adomás caminos reales, trasversales y vecinales, estos en estado natural. (1)

Qué entiende V. por caminos reales, trasversales y vecinales?

Caminos reales, son los caminos abiertos por el gobierno Español desde los tiempos primeros de la colonización de la Isla de Cuba y tienen de ancho 24 varas; caminos trasversales son los que solo tienen de ancho 12 varas y vecinales los pasos permitidos por los propietarios de fincas, para acortar distancias de un lugar á otro y salvar lo mal que pudieran estar los caminos por el fango, las piedras ó la yerba.

Cuántos ingenios para la fabricación de azúcar tiene en la actualidad el Término todo?

Los siguientes: «Santa Filomena» en el barrio de Navajas propiedad del Sr. Leandro Soler, «Elizalde» del Sr. Alberto Broch en el Ciego y «Santa Catalina» del señor Enrique Heedigg en el mismo barrio; «Carmen» del Sr. Alexander en Navajas, «Socorro» del Sr. Pedro Arenal en Tramojos y «Dolores» del Sr. Francisco Rosell en Platanal, todos centrales y con magnificos aparatos.



Handwritten document in Russian (1840), 44 pages, jpg (700×900), 24 ВРР Дом живоначальной Троицы, Свято-Троицкая Сергиева Лавра, Собрание славянских рукописей, 43: Житие схимонаха Феодора

Уже нога его недвиганаль съ обра бытозней; жады амертный очедотника его члены. Осодоръ раствориив свои новора, имаролив соственного тала, сограваль очтенновающее тало буховна го друга, покрывание горанушии побланияти его уды, освященные чистотого давства, иобиль_ нышкогнения Поэтественных внагодати. На рукахь Деодора спонгания великий Никонай, и мощей его не дерэнцио прихоснутая тигоние. Descopto necosesanto er Manny to go 1801 2000 въпродоножении сего времани увидтоло оно кончину, высскаго митения Никоная, увидовить кон чину изнашенитаго Пансия. Пресиникъ сего

Searching in binary text images (v	Find	<u>? x</u>	? ×
Parameters Segmentation Fi	Go to page Feodor_ru/25-16.tif Save	?? 1 0	Previous
nougues cana	Dendosio.	-	ueem_
an sunte no	Viciorito. Deodoor		16 codon
or Hiraceyxin	mulb		З
Decesto n	Heredri		нори
mours Mana	Devcont		coue v
Monaimsiperi	Dendonto		rnuau
Comouxo omy	Sauresree.		Cinne-
Ho-Spereckoro	1 Jeadons		nal_
www.a.nis. Co mus	adumant		3a
sucmo, u nocais	гаятые Нежиди		e Greenade
naugruor chan	Нальза		pece.
маетия въ пу	Combaulb		vy,om
myda depemts ar	Joenoat		ar, u
A	Heogona.	•	v

Text in French (1692), 388 pages, jpg (2048×3550), 8 BPP

Nicolas de Bonnefons, Ch. de Sergy, (1692), University of Gent, Digitized by Google (2007)

Quant à la Terre, si vous la rencontrez bonne, ce vous sera un grand avantage,& une grande épargne ; mais rarement en pourrez-vous trouvez,où il n'y ait beaucoup à travailler, dautant que telle paroîtra passablement bonne au dessus, qui étant ouverte de la profondeur d'un fer de Béche seulement, se trouveral Argileuse dessous ; ce fonds est pire aux Arbres que le Tuf, ou la Roche, à cause qu'il s'y rencontre de petites veines où les Racines peuvent s'étendre & profonder, afin de tirer la fraîcheur de plus bas,& prendre quelqué noutriture; mais l'Argileuse ou Terrre franche ou rouge, fait comme un plancher qui par sa dureté & densité, ne peut être percé par aucunes Racines, & qui dans les grandes ardeurs de l'Eté, em-

5earching in bi	inary text images (version 0.2, 04.2008	3)	<u>? ×</u>
Parameters	Segmentation Find	Directory french File out177.tiff	Next Previous
E	0 · /		
C	ltime avoir	aucune chaleur) &	coup e
r	olus à angle	s droits l'horison.	au'ils
1	0	Find ?	⊻
I	ie font pas	Go to page french/out177.tiff Save YES 1 0	
C	u'en biailaí	cette	ent le
	- Addahillame	cette	aible
	remeenneme	certe	ODIC.
1	e vous dira	Cette	Fraité
J			10 fr
1	tes moyens	entre	TC True
	con cette	cette	
1	Conditions	autre	a ara
	Concinons	autre	^{c.} P ¹⁰⁴ -
	fit de tout c	mere	us re-
1	-	autre	· · · ·
		roux	•
		ICHC	
		-	-

Slavonic manuscript, (1574), 747 pages, jpg (1249×1890), 24 ВРР. Дигитална Народна библиотека Србије, Нирилски рукописи, Збирка словенских рукописа Јернеја Копитара, Зборник "Златоуст"

COMPLITTH HMEYO . HINO LANEMA ATT HOLLECE Боудемь пасе, пь паба выстрышающато Мрытвые. наке штолнисые сымрытан ни ZEABH MA HHZEABHLE. MAN KE OV TOBAY WMA нако пещеп збавны. VIIIO ОУ БО РЪЦИМИ BACMOCHILINCE MICO WAO FOLIN CBOH . DMT стоже блодыть неповалован А авшому тпоко имаши Еже пеприель Есн. Аще АН Прієль Есн, уппосе увалиши нако пепріє МЬ . ПЕПТЫ БА ПОЗНАЛЬ ЕСН ПРАВДОЮ, ПЬ BETTERS ENTOCITO TTOZAA . BIELLE DETE HAYERE HOZNANH ENBLUEWEA приельеся добродателию . пытебь TT PHELLACTTRIEMA TOIETTA . L'ONERO DE ALLE HITOCITH THOV HM AL HITOCITH A HO FALYA WYA . HEBLI MENE H'EPACITE . Mb ь нора вась . ньли ани почьптень Есн BENE MYAPSIMBOVEMH ; HMAM BOBHNY

Find	<u>?</u> ×	? X
Go to page Zlatoust/KOP-005-012.tif Save YES 7	1.38 005-012.tif	Next Previous
Фарнсен	- OY A	hmerin
DACHCEA	_ To' 6	SAKE EN
нлални		1/2.1
DADHEEH	Ρ,•	12 DAO
CELLING OFH	V 5 0	Fb C_
HTTBIHAAE	TIAY	6 L TT PT1
фарнсен	1.	
ELCARCICA	- ,061A	выце
Анинсе . Единь	Блонс	EH AA
ттарь . фарнеен а	PUCK HAI	н прав
Groncen noteks	H REZ	This
	·····>	17.
SENO . HERE MHO	KAE I	τρτεζορ
TA/TEH, HEIKEAH	нмтн	IEME
TTTNU AAL LLAAAG	HIEM	A CHITT
אידות נעוצו אוא איז איז איז איז איז איז איז איז איז אי	h , (*Øl	roche

Conclusion

We show the possibilities of preliminary version of the program.

Experiments with 7 different languages from various epochs give us the certainty in practical benefit of our approach.

It is quite universal and does not required any specific features of the concrete language.

Word searching can be applied to any collection of scanned documents, immediately after the graphic files have been created by the scanner device.

In spite of relatively low efficiency of the Hausdorff type methods (the searching process takes a lot of time) high level personal computers will be able to solve the problem in reasonable time.

We think also that the accuracy of retrieval is quite sufficient for practice.

There are some directions to improve the software.

- Increasing the efficiency and speeding up the search.
- Searching with a part of word as a pattern.
- Character segmentation of a page (or pages) and composing pattern word from well separated letters.
- Feedback making second search for the same word with a different pattern word.
 - The user can choose this word among correct words found in the first search.
 - Produce a new pattern as an average of all or part of the correct words.
- Automatic or semi automatic choice of parameters based on image information.

Thank you for the attention.