

Thresholding of Text Documents

Oliver A Nina William A Barrett

Thresholding or Binarization





- Simple method of image segmentation
- The image is separated in two parts:
 - object of interest– background







Thresholding

Important for the processing of scanned microfilms and OCR (Optical Character Recognition)

Miller, Riley is have tod	Miller, Riley
Noore, Elizabeth J.	Moore, Elizabeth J.
Moores Ina the bill west	Moore, Ira S.
Noore ; dimala the way to	Moore, Wm.
Mosier, Infant Son	Mosier, Infant Son
Mosier, neryli be a read	Mosier, Beryl
Mosier, James R. dich er	Mosier, James R.
Mosier, Jussie Baile you	Mosier, Jessie B.
Mosier, Rebecca Christia	Mosier, Rebeaca Christia
Mosier, Susan demetizing	Mosier, Susan
Mosier Thomas For off f	Mosier, Thomas F.
Mosier Harren Leo about	Mosier, Warren Leo

(Left) Original scanned record (Right) After Thresholding, Enhancement, and Antialiasing



The Problem

- Typical algorithms do a fairly well job on isolating the targeted object (text)
 - However, it is harder when the text looks similar to the background, such as with lighter pen strokes



In many cases important pixels from the image are removed.



Previous Work

- Thresholding algorithm classification
 - Global Thresholding
 1.1 Bi-modal
 1.2 Multi-modal
 1.3 Multi-spectral
 - 2. Adaptive or Local Thresholding2.1 Hierarchical data structures2.2 Small window



Thresholding Algorithms

- Examples of thresholding algorithms
 - Mean or Median value
 - Iterative Method
 - K-means
 - Otsu
 - Niblack
 - Yanowitzand Bruckstein



Related Work

- Another similar recursive approach
 - By Cheriet, Said, and Suen (June 1998)
 - Used for bank checks
 - They use a training set to learn the background (S=95%)
 - It only works if the targeted value is the darkest value in the image.





Our Approach "Rotsu"

1. Background Estimation



2. Background Subtraction (Hutchinson 2004)



3. Apply Otsu Iteratively in different parts of the histogram





1. Estimation of Background

- We apply a median filter with a kernel of radius ~21 or bigger to the image



2. Background subtraction

We subtract the original image from the background
We normalize the histogram in order to get rid of negative values and be able to see remaining pixels





3. The Otsu Algorithm



Goal: Minimize within variance class



3. The Otsu Algorithm



Goal: Minimize within variance class



3. The Otsu Algorithm

Optimal Threshold



Goal: Minimize within variance class





Mathematically

σ^2 Within(T) = nB(T) σ^2 B(T) + nO(T) σ^2 O(T)

$$\begin{array}{ll} T-1 \\ nB(T) = \sum p(i) & \sigma^2 B(T) = \mbox{the variance of the pixels in the background} \\ i=0 & (\mbox{below threshold}) \end{array}$$

$$N-1$$

nO(T) = $\Sigma p(i)$
i=T

 $\sigma^2 O(T)$ = the variance of the pixels in the foreground (above threshold)



Otsu

- Calculating within-class variance is too expensive
- Another way is to maximize between-class variance

$$\sigma^2 = \sigma^2 \text{ Within}(T) + \sigma^2 \text{ Between}(T)$$









Rotsu Recursive Otsu



The algorithm

threshold = Otsu(image)thresholdImage(image,thImg,threshold) While(threshold < 255) { // until no more to threshold excludePixels(image,thlmg,excludedImage) threshold = Otsu(excludedImage) thresholdImage(excludedImage,thImg,threshold) saveAndDisplayImage(newImg)



The algorithm





Results





Original with background substracted







First Set = S1







Second Set =S2







Third Set = S3







Fourth Set = S4







S1 + S2 + S3 + S4







Original with background substracted (K=41)







First Set =S1







Second Set = S2

Third Set = S3

S 1+ S2 + S3

•

solumente à un y le nombre Dionin furtin Dominge y ch. talvador y y Feedora à

Background Approximation

solumente à un y le nombre Dionis inos de este pueble. alador y y Feedora à bueble a quienes adde

First Threshold = T1

First Iteration soluminante à un y le mombre Dioniquation & continue y ct. alvador y y Jeodera e buckle a quince a de

solumente à un y le nombre Dionis omingo y ch. quotin Q einos de este pueblo. alador y y Seedora bueble a quimes

Remaining Pixels

Substracted Pixels allow an anonth come from a Ca north Bann MA heador on matte buchto a quiver

solumente à un y le nombre Dionis omingo y A quistin 0 einos de este pueblo aliador y y Feedera bueble a quienes adde

Second Threshold = T2

Substracted Pixels allogoagagggaggte a la monte Decore Barriso Degenoros y the and formalles. 20 marco Tallow of y Docla bacallo a gariones

solumente à un y le nombre Dionis Justin Domingo y A. ernos de este pueblo. alador y y Feedora à bueble a quienes and

T1 + T2

Rotsu ale and a second y le montre Dianie Surtin Domings y A cionos de vote foneblo. aliavor y y Deodora bueblo a quimes alle

The second		TA
TIS in the Parish of Hope as	nder Dia smort in the in the year One shousand	
Nume Abois.	When burled. Age. By whom the Dermony was performed.	650
lite Dals Luminster.	Fill 12 37 lotoyale.	R
ian Stafford Hope The itsue	april 12 94 Webyalt,	11
ward Gough . Woodmands	a april 10 76 lody att	K
imes Smith . Hill Hole	april 2 100 W hudrus	
homes quatter Hewston.	may 12 67 Mattakens.	A J
James Goodwin Quinen The setter	at June 15 64 M. Quarens	
John millicher The Pager A	Here your 10 swanterus	
Elizabet Bathan The Vallet	2 June 2 60 hologat	PAR
1 1/2 183		
1 1 1 4 1 4 1	A BAR AND	The second

1 1 1 1		1-1-1-	121
121	Ro G	and Participa	
US in the Parigh of Hope any of Hertford chuhdred and sightly.	under Diamin th	e year One thousand	
Name. Abole	. When barled. Ap	te. By whom the Ceremony was performed.	Arba
to Dalo Lonins.	ler. Fill 12t 3	7 Worker.	2
an Stafford Hope (ne its	me april 12 0	14 Webyatt.	11
ward Gough . Woods	antra april 15	76 lordigate	1
mas Smith . Hill I	Hole april 27	00 1 Madrus	
lomes Grafter News.	Ta . may 1th 0	57 Maldrews.	
meelovanin Diane The u	in Mark Die 15th 6	4 f.W. andrews	
John millich + The The	Hill Jene 13th ,	10 1 Willintreus.	art.
Elizabeth Baihan The U.	Aletto June 29 6	o tologate	17 N
i dia St	17	R. M	
1 IV	C. C. P.		

Background Subtracted

Page 6 By when When barled. 15toyas Febs L. Woodmanton april 10th 76 Gong april 27 100 1 W Hudrews Hole Hill Smith Formes Grafter Newton . May 1th 67 Maliakeus The lite host Die 15th 64 M. and res The Pigli milliches Küntrus. 11-11 ligo Elizabet Bathan The Vallets June 29 60 to loyatt

Phylos G	
ElS in the Parish of the sector Diamore in the sector of the sector in the sector of t	
Name Abels. When barnet, Age, By whom the Correctly was performed.	
the Sale Leminster. Fill 123 37 Boyale.	
ian Stafford The other april 12 94 Webyatt.	
wind Gough . Woodmanter april 10- 76 lor dry att	
unice Smith . Hill Hole april 24 100 11 Audrus	
homes Grafter Newton . Bray 1th 67 Milladrews.	
Smelfordin Quinne The cile	5.
John millich a The Tiger time gene 19th 10 160 atrens	
Elizabert Bailer The Valletto June 29 60 bo logat	E S

By who 15toyas Feb bodinanti april 10 april 27 100 1 W Rudrus Smit Grafter Newton may 1th 67 Malidkens. Park Due 15 64 M. andrew maple milliches Areus las Elizaberte Bathan The Valletto June 25 60 bo loga

lotora boodmantin april 10 123 100 1 W hudrus Newton may 1th Malakews. Pask June 15 64 1. Andrew tenaptie millichas Areus Elizabeth Batha up Aile June 29 60 W 1040

Page 6 then bas Fell lotoral Luminster bodinantin april 10 Gong april 27 100 1 W Kudrus Smith & Graiton Newston may 1th 67 Maladrews. unapla Park Duc 15 64 L. Andrew The Pratie he millichas 1 Wintrus. Hill ages Elizabath Bathan The Valletts up the June 29 60 to loyal

Original Image

Page G 1:15 in the Parch of the seater. Diagnoster asy of there and the in the year of the year 10,100 in the year One shousand By when the Carr Abole. West band. 1 Apr. N the Sal Fill 12 - 1 37 Lunister. 15toyan Hope .. april 12 94 Autopatt & Gough . Birdmanter Coul 10 76 Werky att mis Smith Hill Hole Gent 2 100 11 Queleur Grafter Newton . may 12 67 Kaldrens The will - here 15 - 64 f. H. andrew John millich of The Poplar time 14. U. trus June 19 - 10 Elizabert Barka The Valletts uneng 60 Wine 1 -1

S1 + S2 + S3

leven, George G. Bijurate, David Mo. Daviels Thomas IT. Reed, John Griffiths, John Gills, Richand Galmer. Edward J. Edwards, and toreph W. Morgan There, They had come there acconting to my directions. " The General Council of all the Monistry in The British Mission commine at 10, a. m. and These between has came to Birningham for this purpose, Knowing ho, Daister Smith, at king also awark of the esconfitant charges generally made for lodgings in all strange places, and The low circumstances of my bethien, I have to ho. a siste Smith some three weeks before hand, requesting them to indeavour to obtain lodgings for the bethren lat as low a charge as possible. They manager to find them lodgings at a decent public-house at 4d. each per night a I managed tot get bo, as writes Imith's condent to let them eat Their own for at their house in The morning and afternoon , by which the bethren were enabled to have sufficient of head a butter and tea a coffee at less Than 3de per meal, lesides the comfort of having their forth at a place where they could feel free a unfettered. Then I was at The Council that was held here at the commencement of 1859, I had to hay 1/ each for my lodgings, and row heary fince for foot. This was not ver exorbitant, get it was the times the sum the helpsen paid This time, at yet not so comfortable.

After I had a little repart, I repaire to Farmoto Chapel, Hockly; where the Council was king bety, The minutes of which I shall have after haster. This of tension is meeting communed at 4 at terminales at

. Correct, Projunto, David Mo. Davie Jo Inemail Hi Reed, John Giffiths, John Gills, Rich-1) (Talmer, Colward J. Edwards, and with IT. Morgan There, They had come There are conting to my directions. " The General Council of all the Ministry in The British Mission commine at 10, a. m. at These wetheren had came to Birmingham for this purpose, Knowing ho, Dister Smith, as king also aware of the exconfitant charges generally made for lodgings in all strange placed, and the low circumstances of my bethicn, I have to bo. a siste Smith dome three luceks before hand, requesting them to induriour to obtain lodgings for the between at as low a harge as possible. They manager to fins them lodgings at a decent public bouse at 4d leach per night a I managed tot get too a order Inuth a condent to let them cat Their own food at their house in The morningand afternoon , I by which the bethren were enabled to have Sufficient of bread as butter and tea as coffee at less Than 3ds per meal, lesides the confact of having Their forth at a place where the could fel free an unfettere? . " Then I was at The Council that was held here at the commencenext-of 1859, I had to hay 1/- each for my lodgings, and row heary fince for floop. This was not very exeorbitant, yet it was three times the sum the heltreen hair This time, and yet not so comfortable,

After I had a little reparts I repaired to Tarmoto bhipels, Hockley; where the Council was being bets, the minutes of which I ohall hereafter has to. This of temen is meeting communed at 4 as terminates at

Jeremy, George G. thomas IT. Reed, I and Galmer, Ed Joseph W. Moorgan

Jeremy, George G. Bijurate, David Mo. Davies Thomas IT. Reed, John Griffiths, John Gills, Richand Palmer. Edward J. Edwards, and Joseph W. Morgan There, They had come there acconting to my directions. " The General Council of all the Monistry in The British Mission commine at 10, a. m. and These between has came to Birmingham for this purpose, Knowing ho, Daister Smith, at king also awardes of the esconfitant charges generally made for lodgings in all strange places, and The low circumstances of my bethien, I herote to ho. a siste Smith some three weeks before hand, requesting them to endeavour to obtain lodgings for the bethren at as low charge as possible. They manager to find them to at a decent public-house at 4d. leach per n I manager tot get bo, a winter Smith's condent to eat their own for at their house in The ma and afternoon , by which the bethren were enal have sufficient of bread as butter and ted coffee at less Than 3ds he meal, besides comfort of having their fout at a place where could feel free a unfettered. Then I was at The Council that was held here at the commence ment of 1859, I had to hay 1/ each for my lodgings, and row heary fince for floot. This was not say exorbitant, get it was three times the sum the helpsen paid This Time, at yet not so comfortable. After I had a little repast, I repaired to Farm the Chapel, Hochley where the Council was being betw, The minutes of which I shall hereafter haster. This attenion is meeting communicat at 4 and terminates at

Original Image

may - george ge thomas We Read, Jo and Galmer, Ed charth W. Morgan

Saranges George G. Binister, Sarrid Mo. Jonets Thomas IT Read, John Griffithes, John fitte, Rich-and Balmer, Edward T. Edwards, et ... chough IT. Morgan There, " they bud court There are To it the Alonisto in The Bitter Mispin fime let "it as in ... tall there better her earne & Birning have for this for for Kinging for Dester Smill for where for bidginged on all strange placed, on the low Smith ane there frether before hand, sequesting then to dearant to selection bedriver of the better lat as how a light public - Anti at 4d peak permisto as the town and a to lat the It' and ford'at their hauter in The openhicks-Hornes by chief the between men enabled the afficient of land and latter and the and a Int of a barrie that fat at a prove others they the Connell that will have been at the consigned out of 1859, die to have If each the and bediego arad Anos have for this This find har if care orbitants i wat when the timen the sum the helther Hand This fling and yet not so comfortables is illing bad in little reports I repaired to Forspice Chefelia Hackets where the Campella - weiving beton the devalues a listich & shall have after heats Frick

Final Composite

Conclusion

- Although Rotsu is still a work in progress, it definitely shows promising results
 - Rotsu allows us to save softer strokes that would be lost with conventional methods otherwise.
 - -Relatively easy to implement.
 - Opens up the door to new ideas on how to improve thresholding.

Further Work

- Determine a better background estimate.
 - Automate the selection of kernel size for the median filter
 - Improve the criteria with which we decide to get rid of background pixels
 - Investigate to see if the combination of Rotsu with other techniques would be better

Questions?