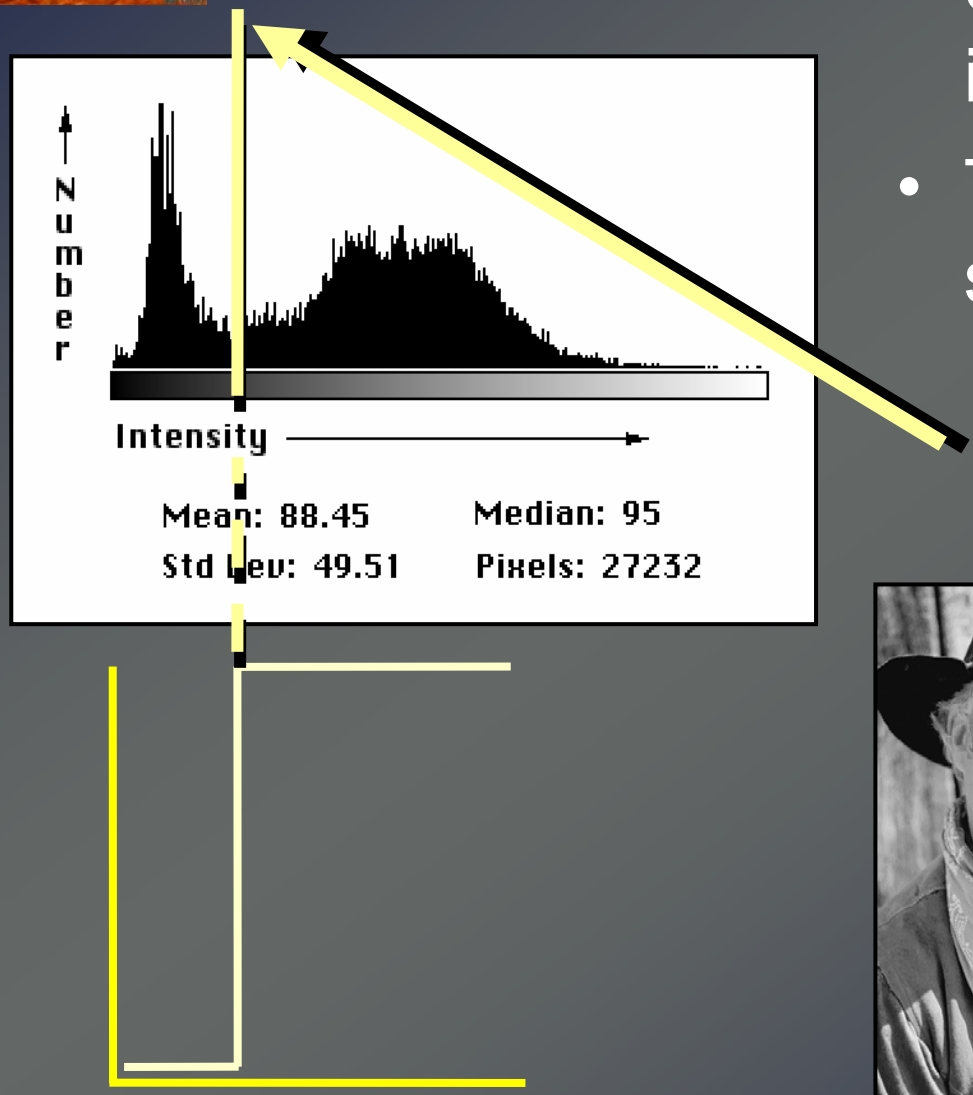
Thresholding of Text Documents

Oliver A Nina William A Barrett

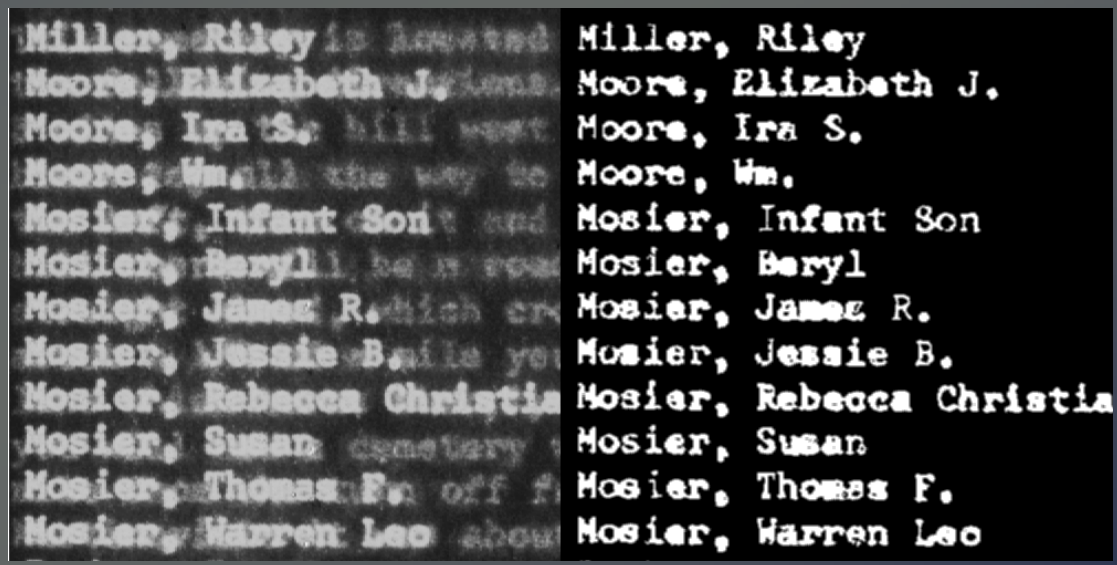
Thresholding or Binarization

- 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)



(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***

1. Global Thresholding
   - 1.1 Bi-modal
   - 1.2 Multi-modal
   - 1.3 Multi-spectral

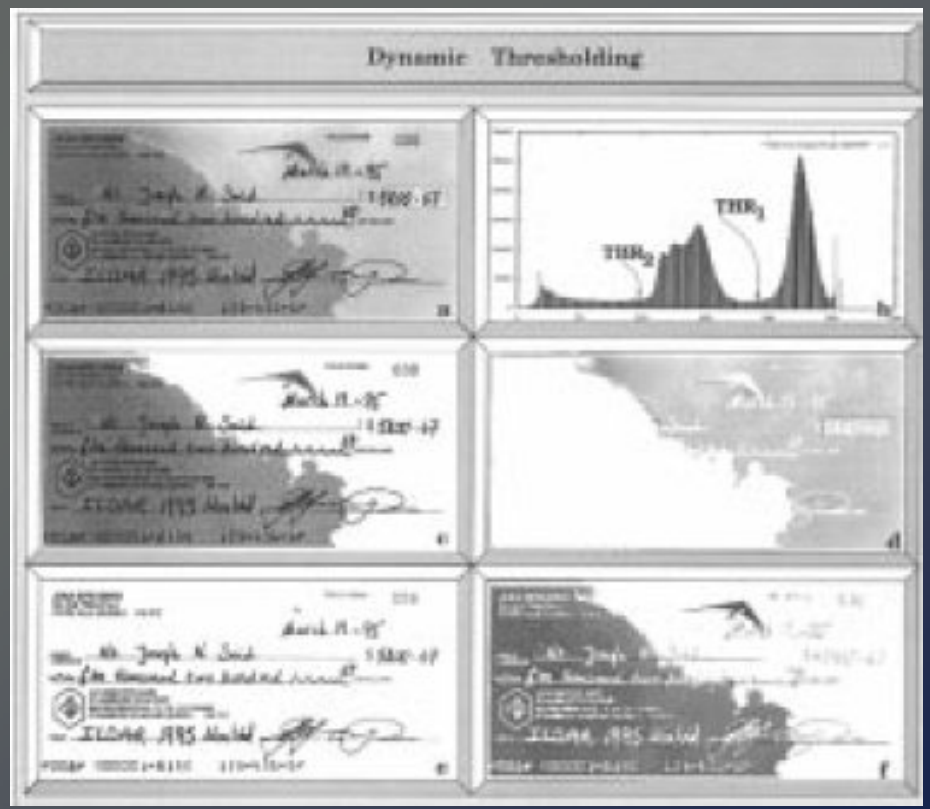
2. Adaptive or Local Thresholding
   - 2.1 Hierarchical data structures
   - 2.2 Small window

Thresholding Algorithms

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

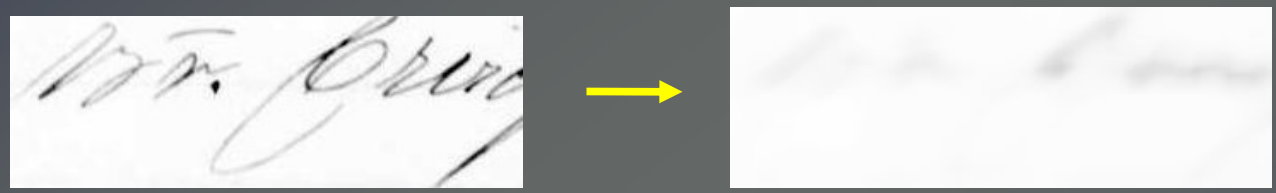
Related Work

- 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.



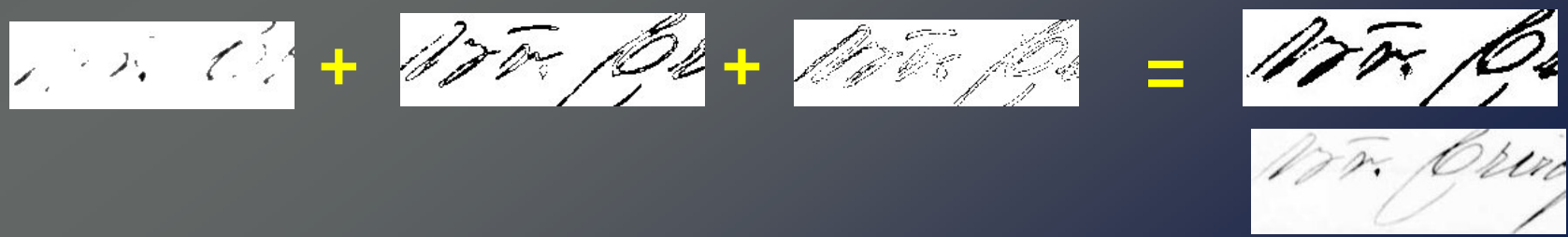
Our Approach
"Rotsu"

1. Background Estimation



2. Background Subtraction (Hutchinson 2004)

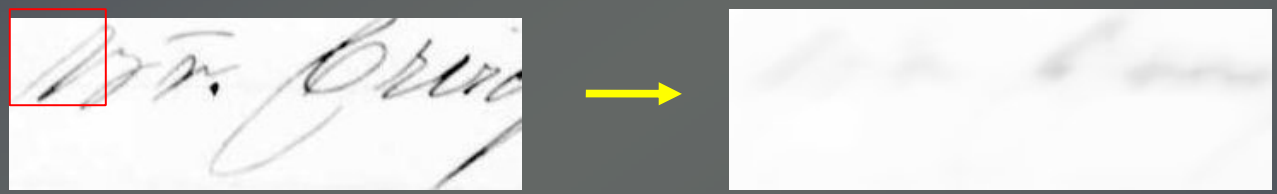
3. Apply Otsu Iteratively in different parts of the histogram



Our Approach

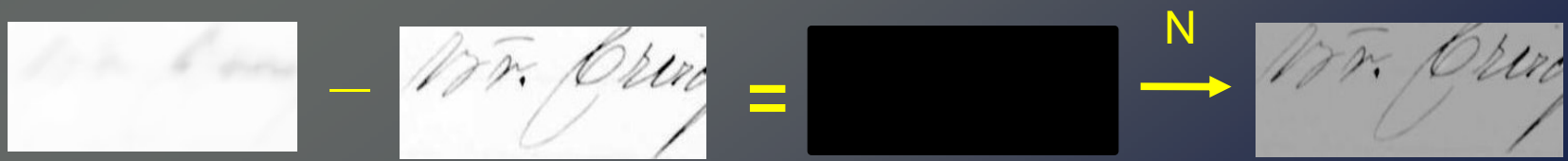
1. Estimation of Background

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



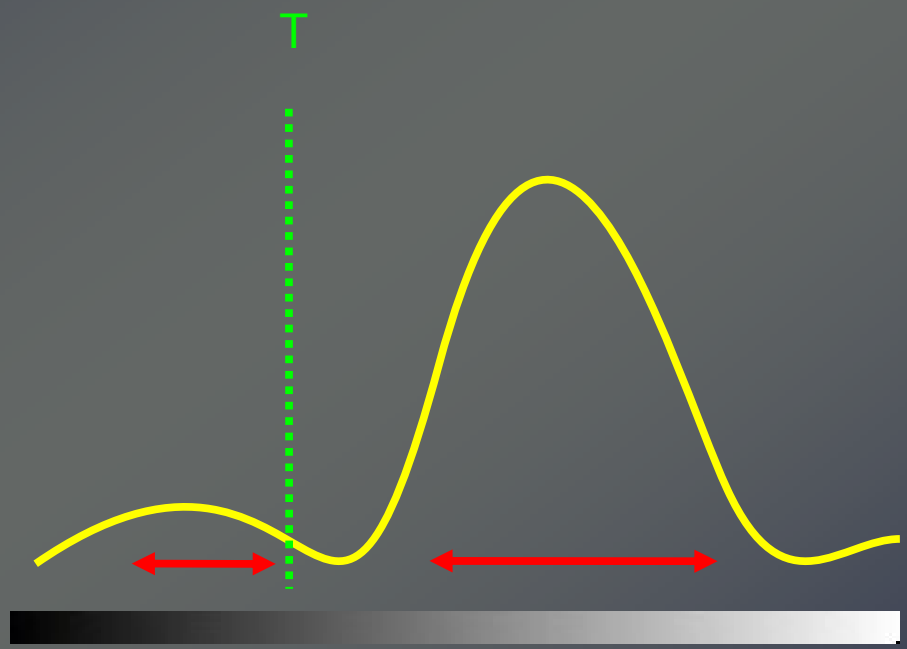
2. Background subtraction

- 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



Our Approach

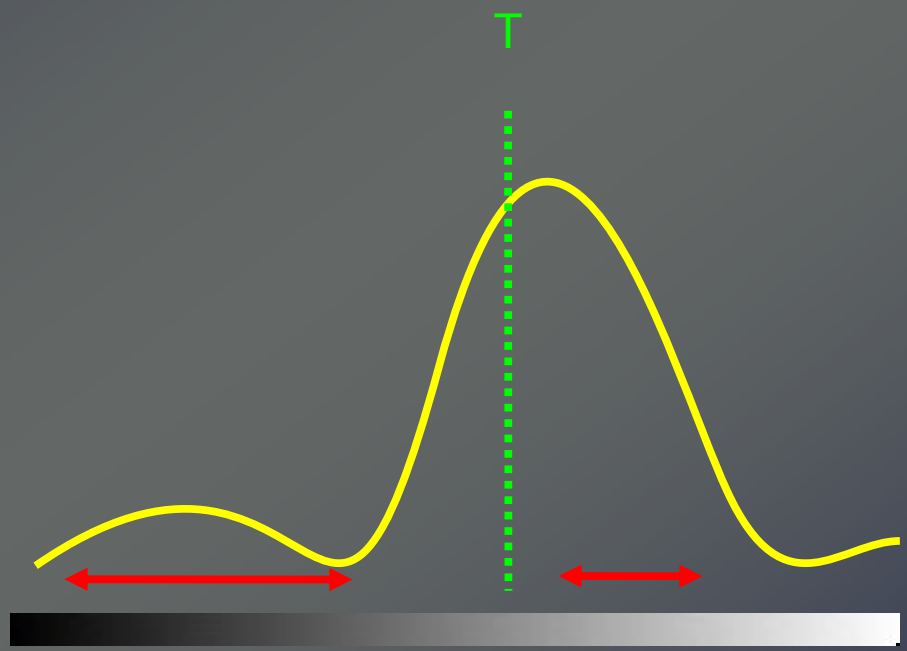
3. The Otsu Algorithm



Goal: Minimize within variance class

Our Approach

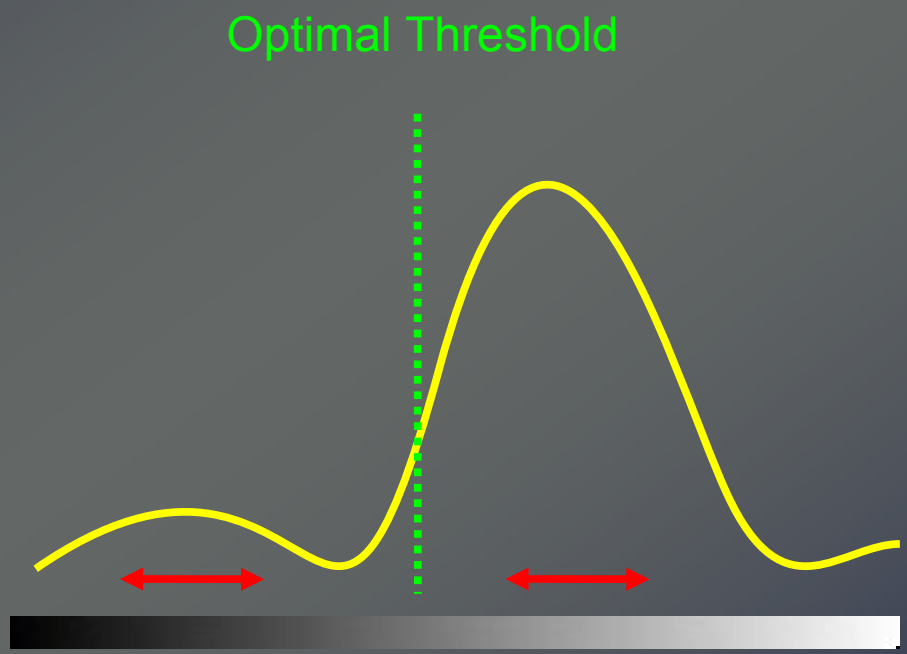
3. The Otsu Algorithm



Goal: Minimize within variance class

Our Approach

3. The Otsu Algorithm

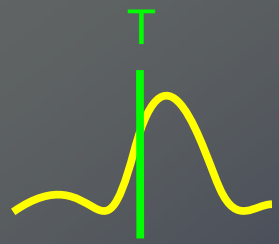


Goal: Minimize within variance class

Otsu

- **Mathematically**

$$\sigma^2 \text{ Within}(T) = nB(T)\sigma^2B(T) + nO(T)\sigma^2O(T)$$



$$nB(T) = \sum_{i=0}^{T-1} p(i)$$

$\sigma^2B(T)$ = the variance of the pixels in the background (below threshold)

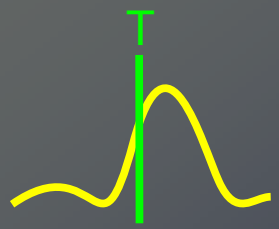
$$nO(T) = \sum_{i=T}^{N-1} p(i)$$

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

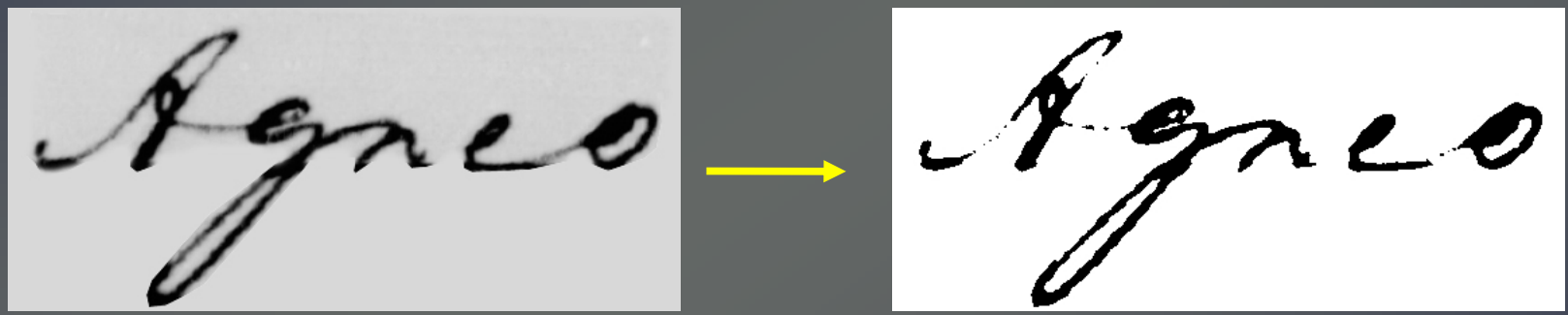
Otsu

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

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



Otsu



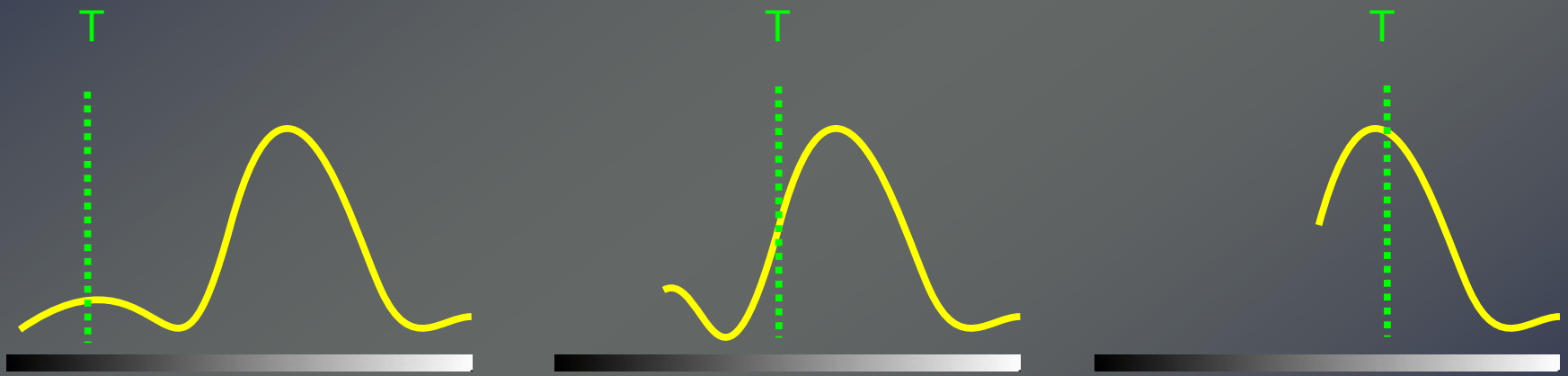
Rotsu

Recursive Otsu

The algorithm

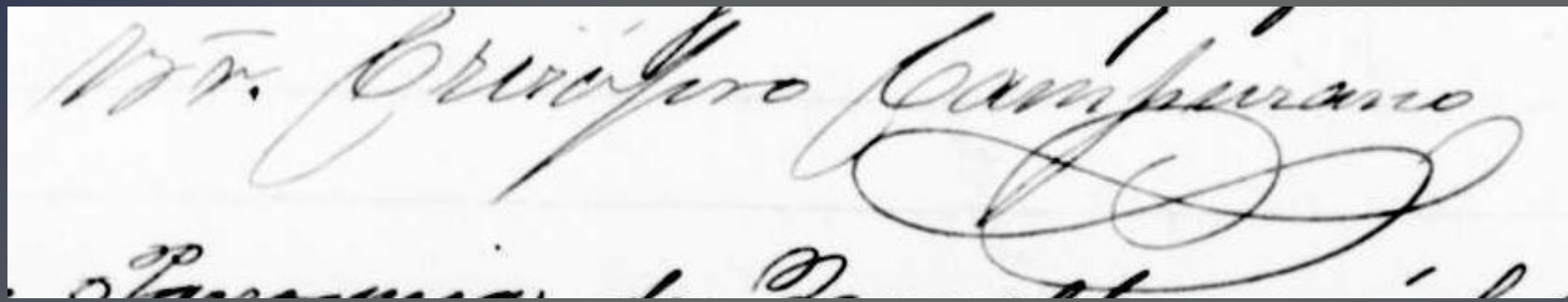
```
threshold = Otsu(image)
thresholdImage(image,thImg,threshold)

While(threshold < 255) {  // until no more to threshold

    excludePixels(image,thImg,excludedImage)

    threshold = Otsu(excludedImage)

    thresholdImage(excludedImage,thImg,threshold)

    saveAndDisplayImage(newImg)

}
```

The algorithm

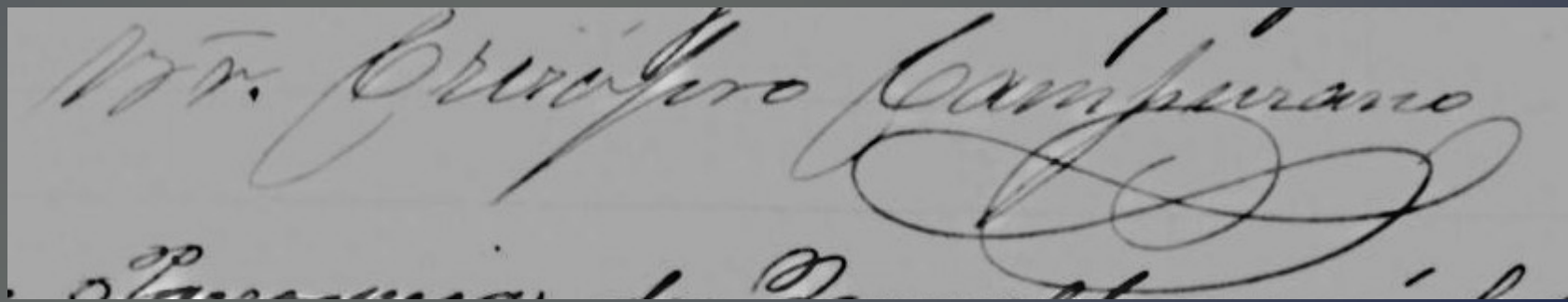


Results

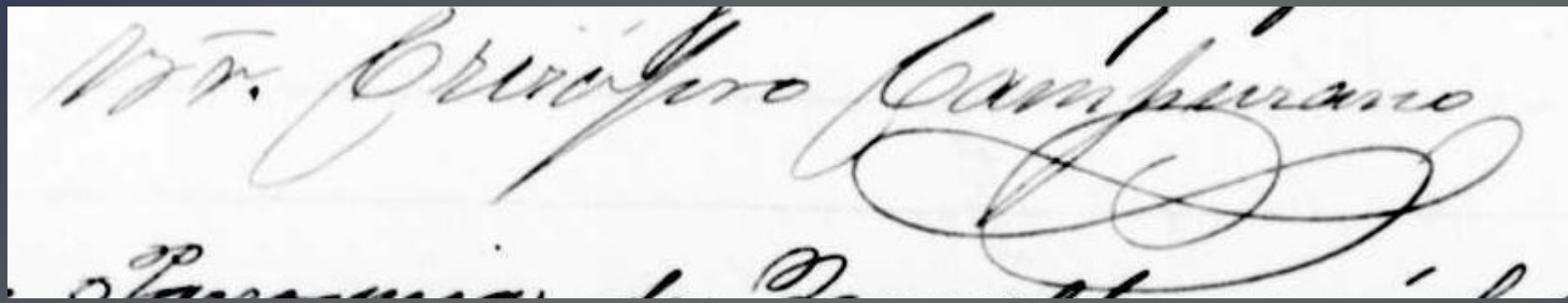
Original Image



Original with background substracted

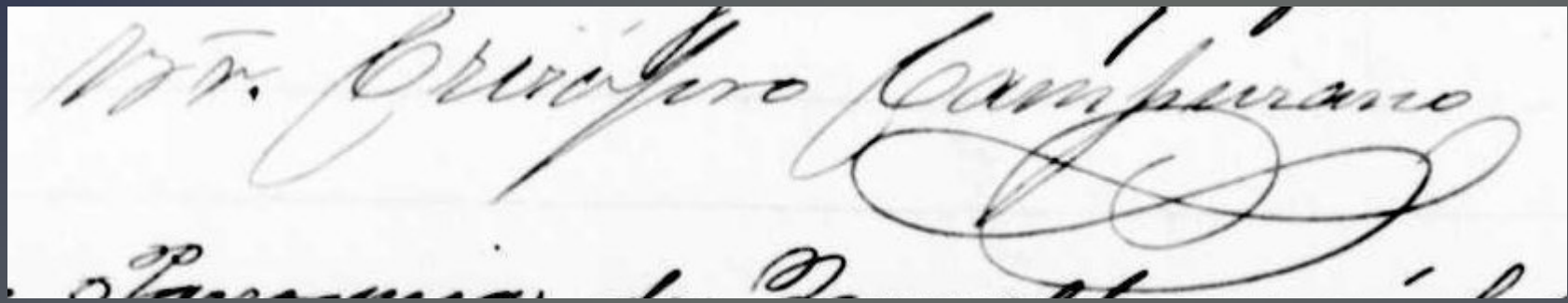


Original Image



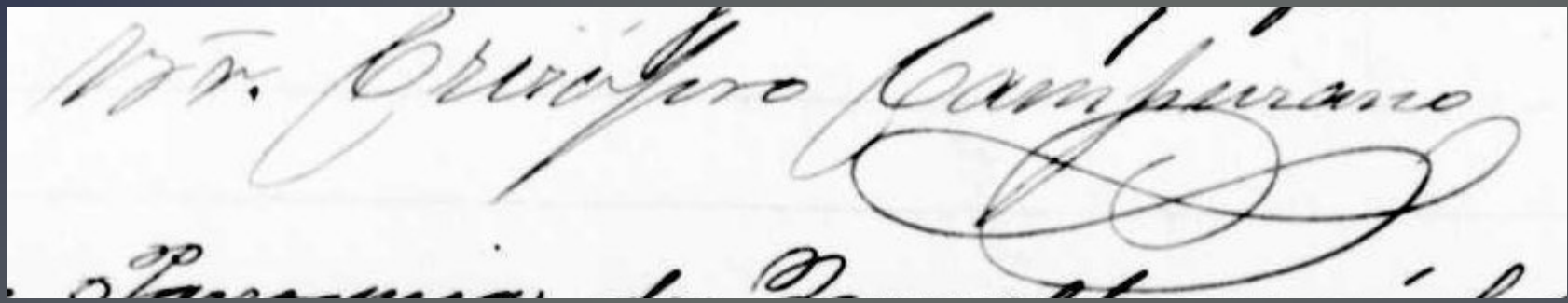
First Set = S1

Original Image



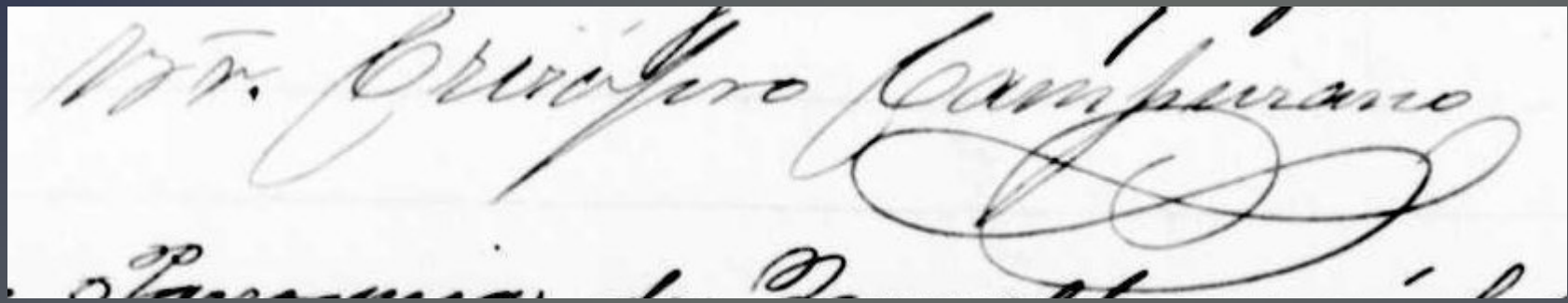
Second Set =S2

Original Image



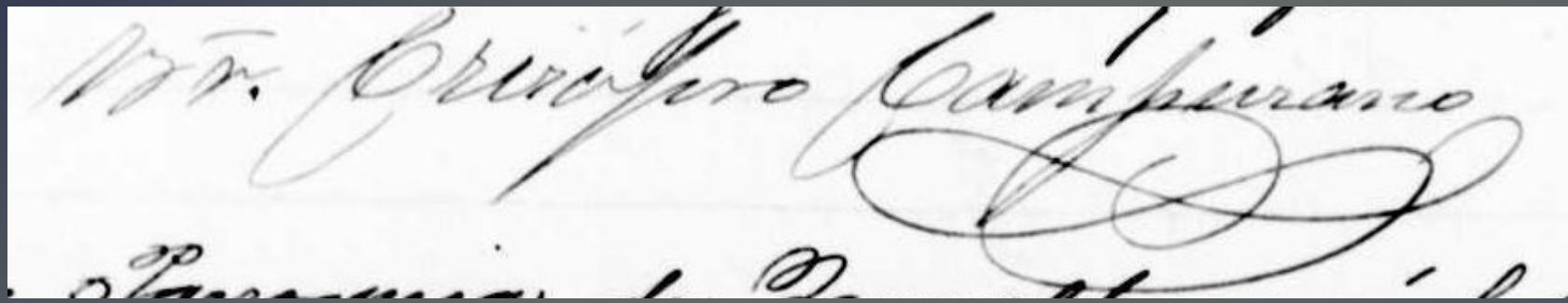
Third Set = S3

Original Image



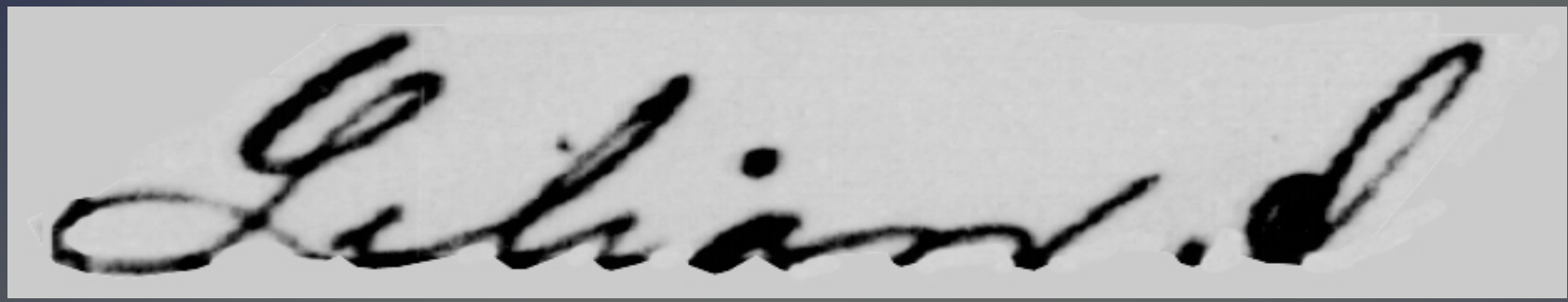
Fourth Set = S4

Original Image

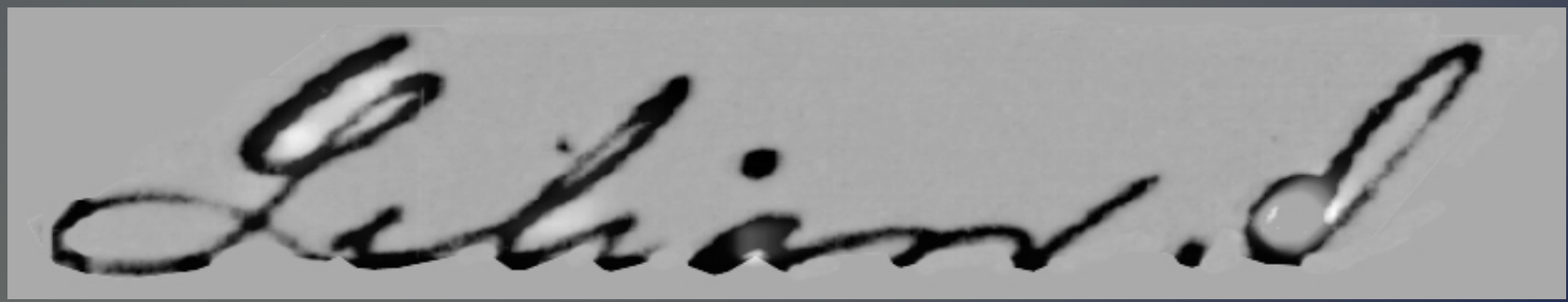


S1 + S2 + S3 + S4

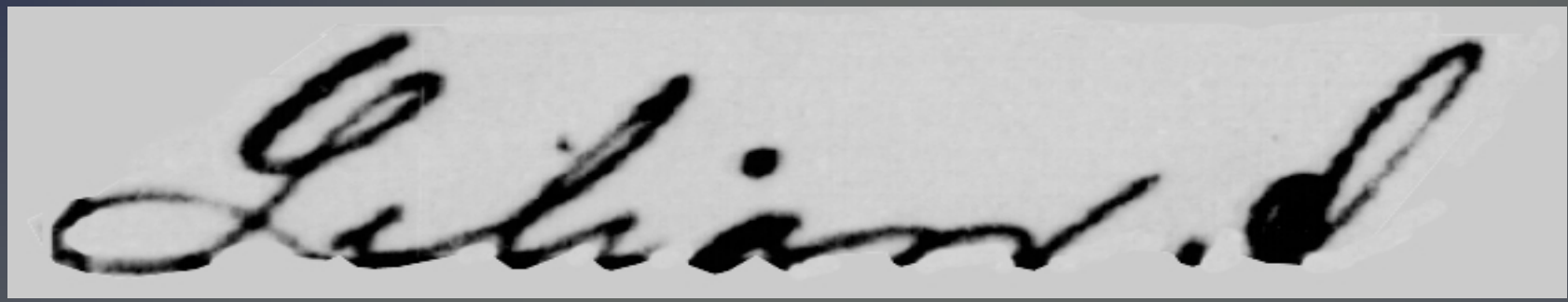
Original Image



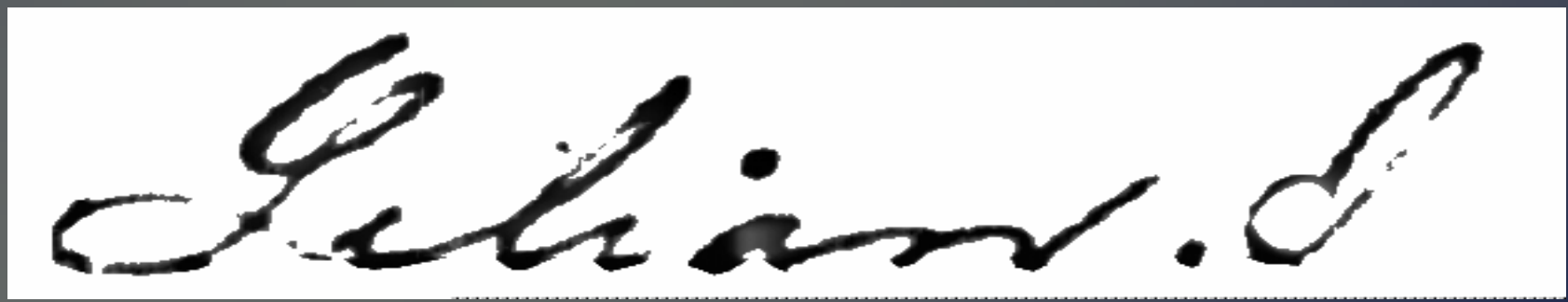
Original with background substracted (K=41)



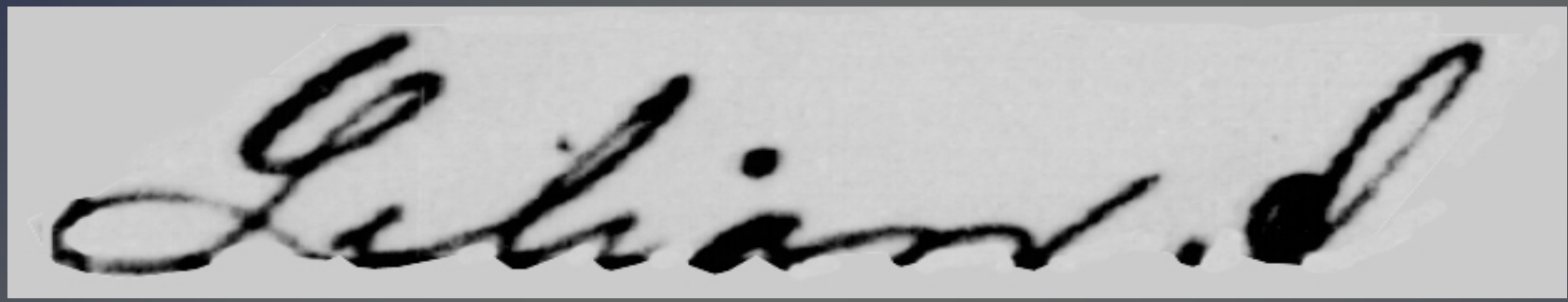
Original Image



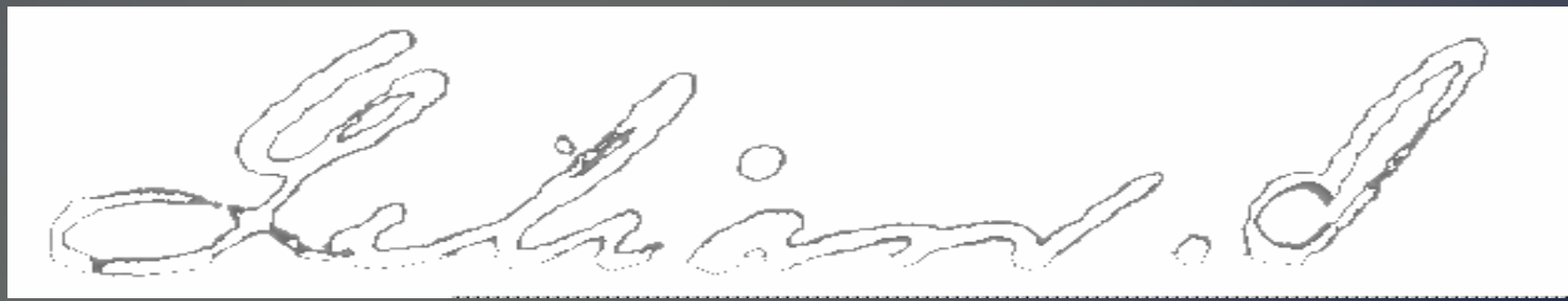
First Set =S1



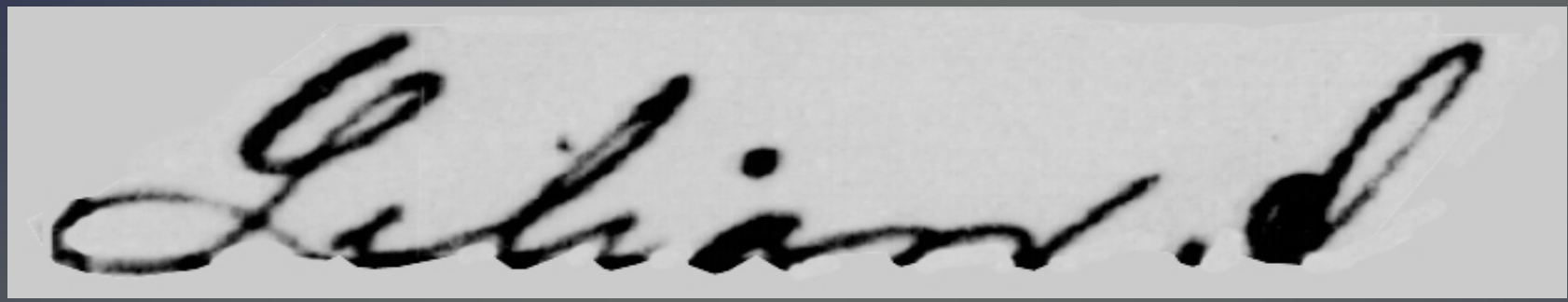
Original Image



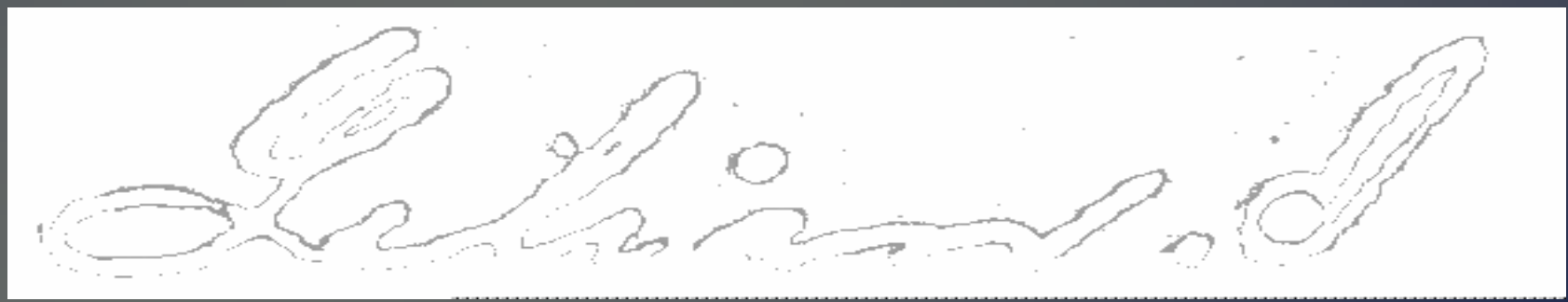
Second Set = S2



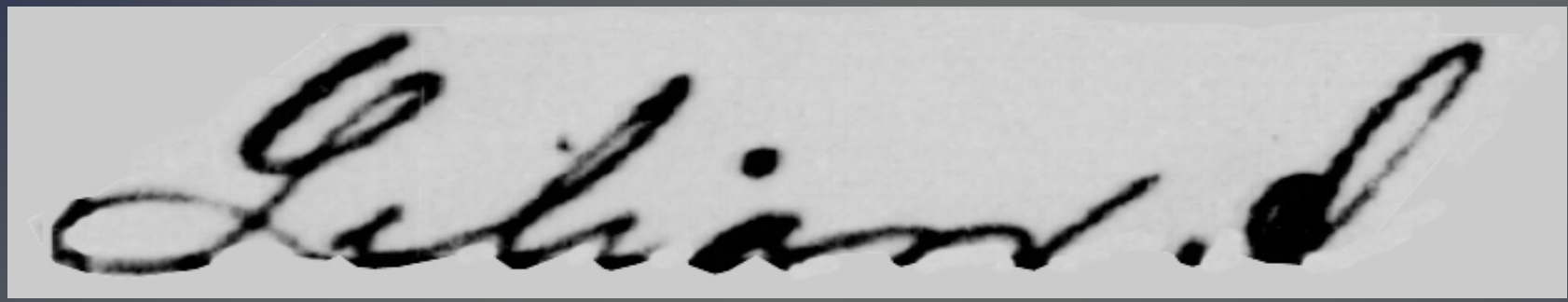
Original Image



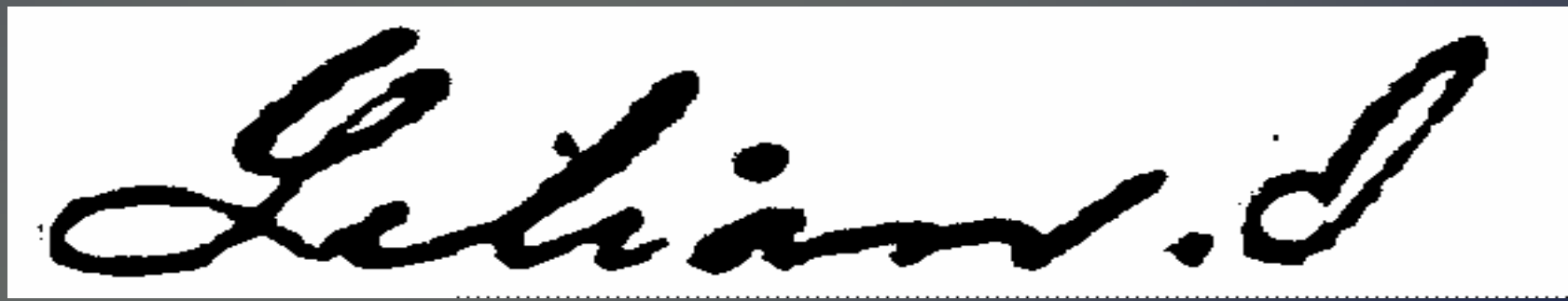
Third Set = S3



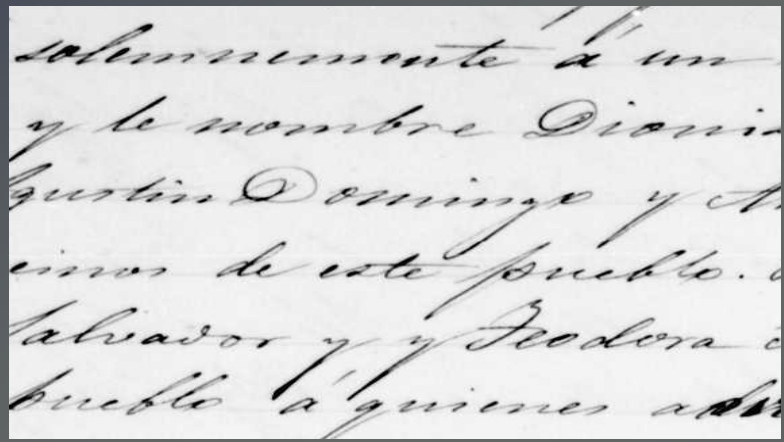
Original Image



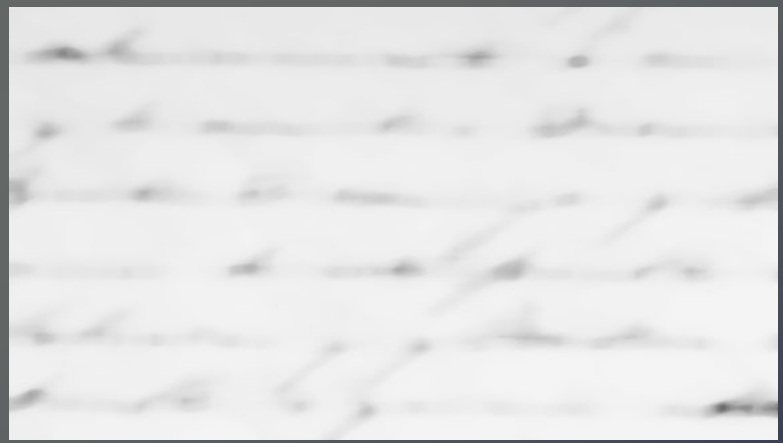
S 1+ S2 + S3



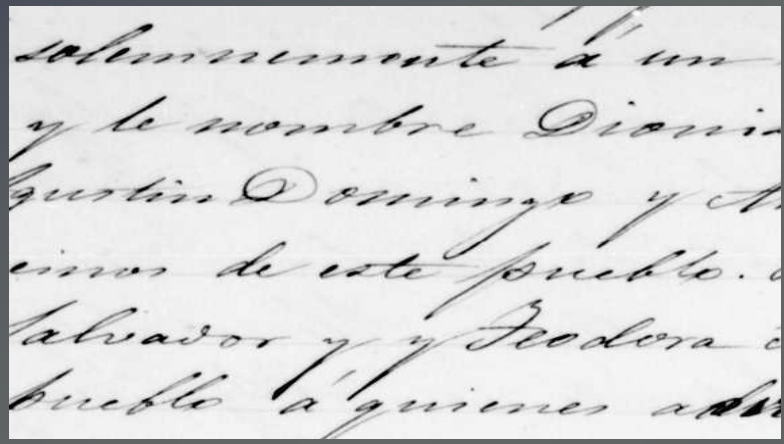
Original Image



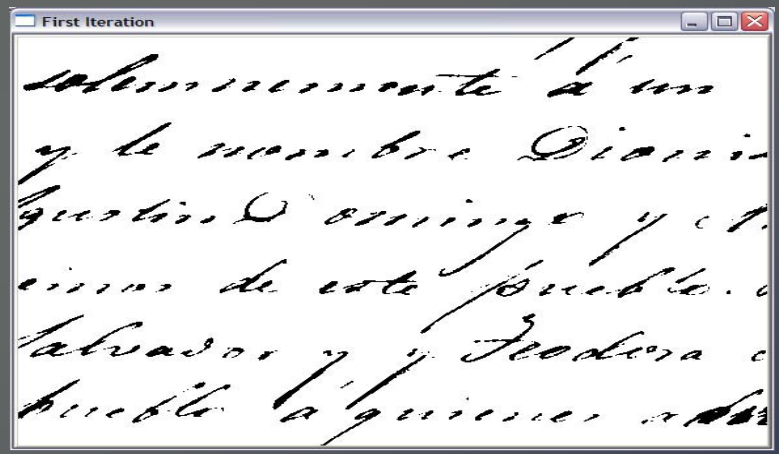
Background Approximation



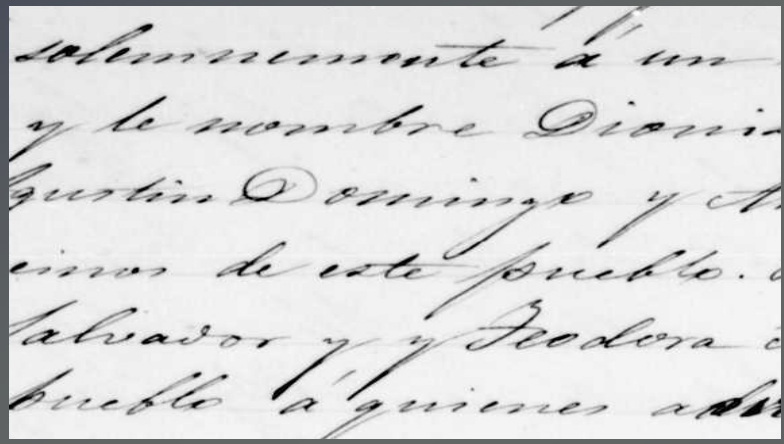
Original Image



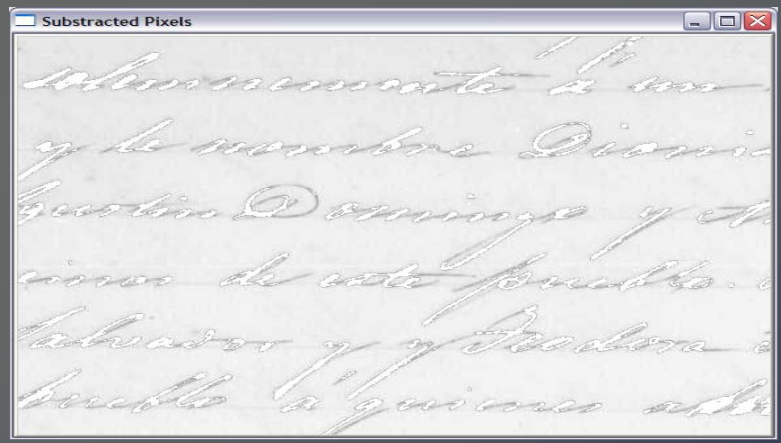
First Threshold = T1



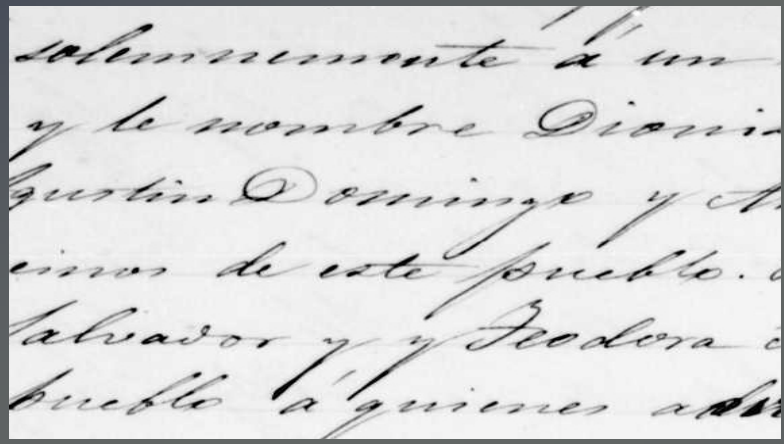
Original Image



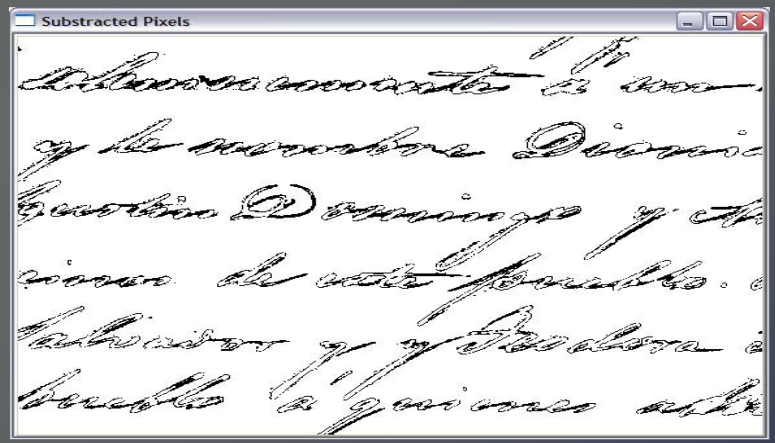
Remaining Pixels



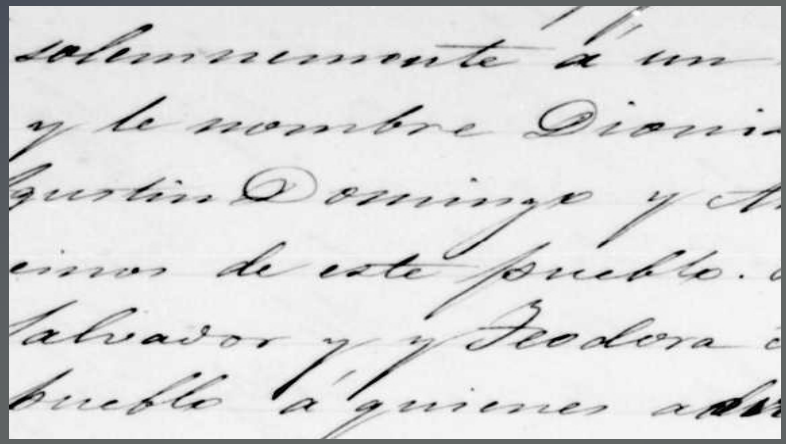
Original Image



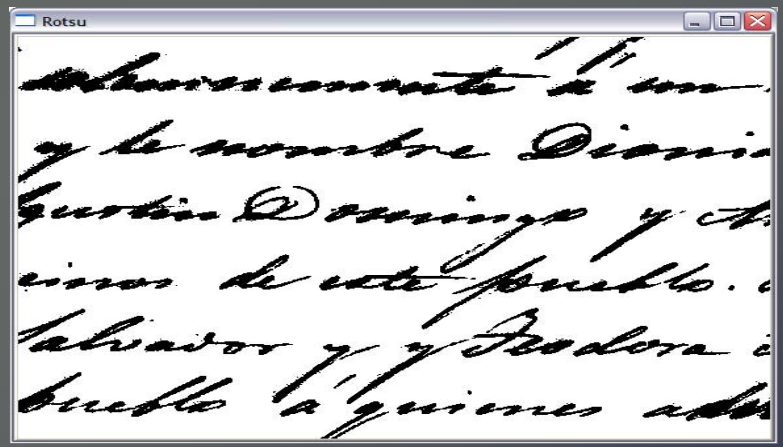
Second Threshold = T2

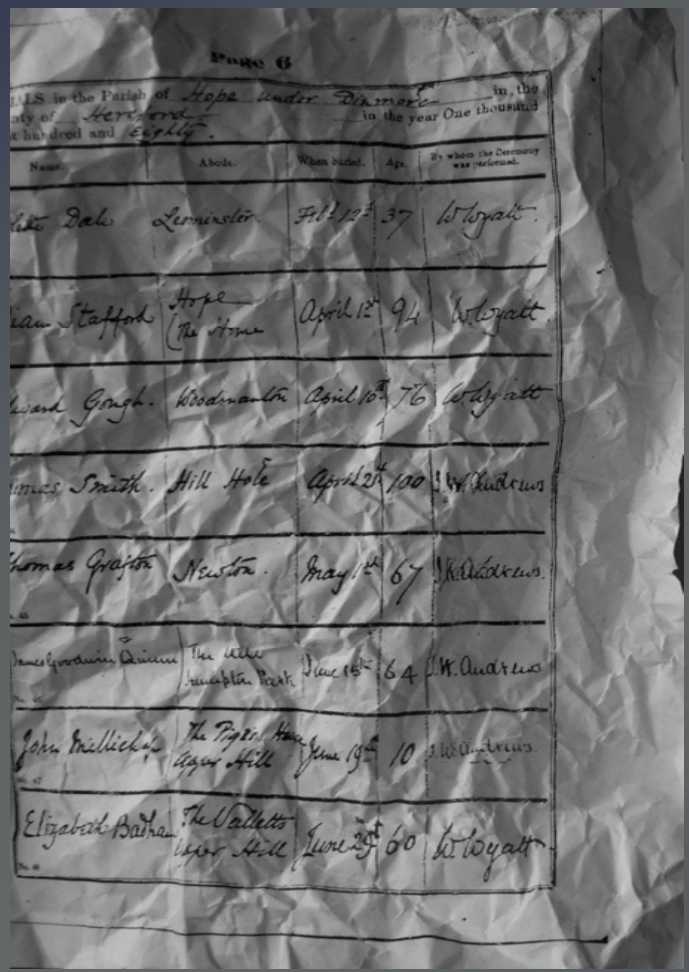


Original Image



T1 + T2

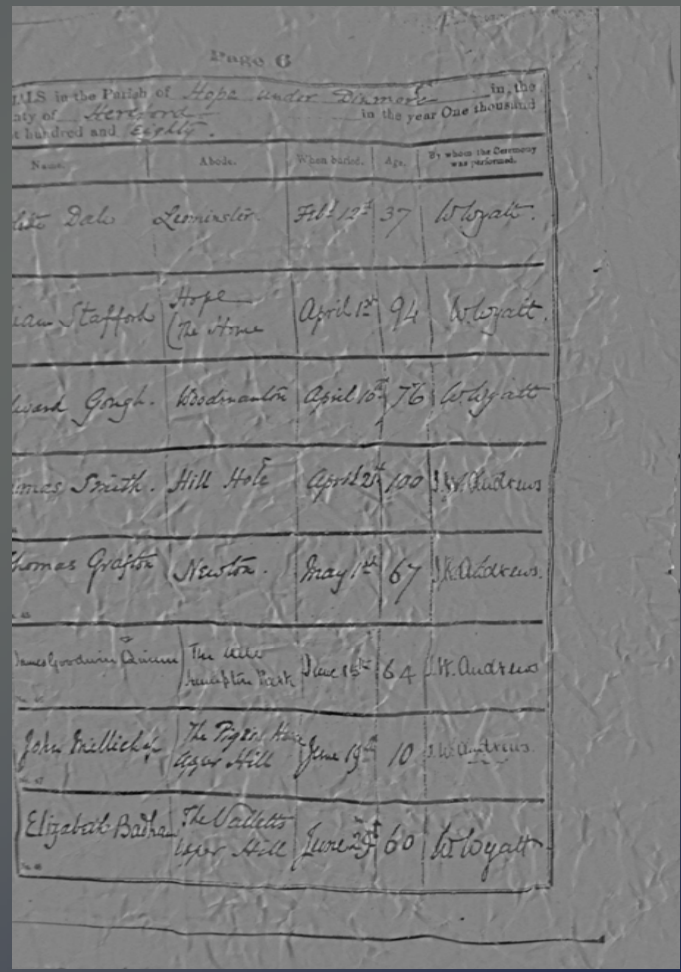




Page 6

...IS in the Parish of Hope under Dinmore in the ...ty of Hereford in the year One thousand ...t hundred and Eighty.

Name.	Abode.	When buried.	Age.	By whom the Ceremony was performed.
...ste Dale	Lemmister	Feb.y 12th	37	W. Loyatt.
...iam Stafford	Hope (the Home	April 12th	94	W. Loyatt.
...ward Gough.	Woodmanton	April 16th	76	W. Loyatt
...mas Smith.	Hill Hole	April 28th	100	J. W. Andrews
Thomas Grafton	Newton.	May 1st	67	J. W. Andrews.
...James Goodwin Quinn	The Alter ...mpton Park	June 15th	64	J. W. Andrews
John Mellichip	The Piper's Hole Ager Hill	June 19th	10	J. W. Andrews.
Elizabeth Bather	The Valletts Ager Hill	June 23d	60	W. Loyatt.



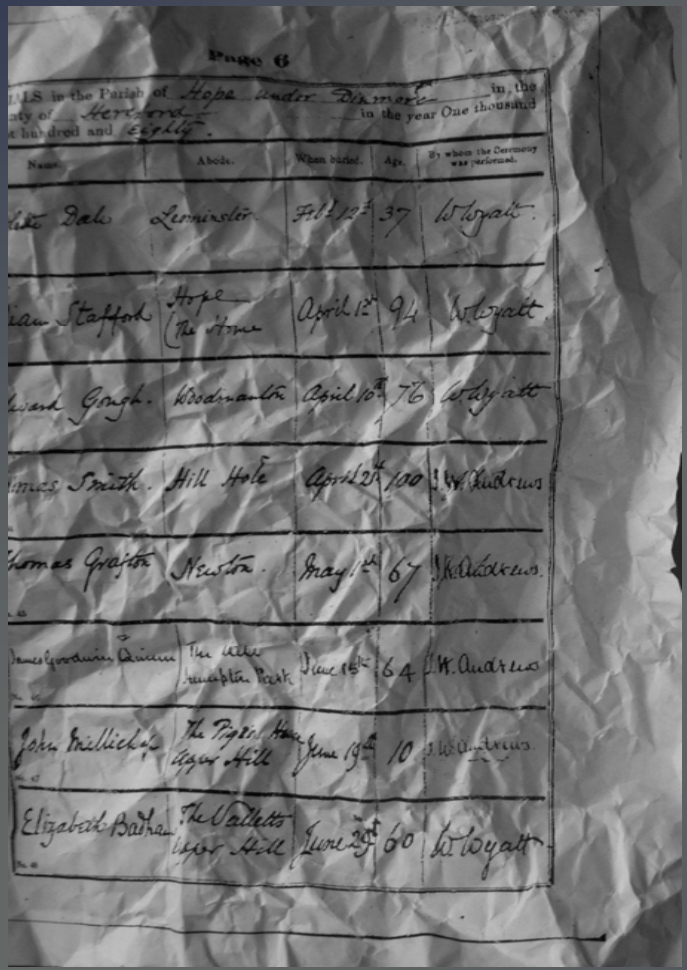
Page 6

...IS in the Parish of Hope under Dinmore in the ...ty of Hereford in the year One thousand ...t hundred and Eighty.

Name.	Abode.	When buried.	Age.	By whom the Ceremony was performed.
...ste Dale	Lemmister	Feb.y 12th	37	W. Loyatt.
...iam Stafford	Hope (the Home	April 12th	94	W. Loyatt.
...ward Gough.	Woodmanton	April 16th	76	W. Loyatt
...mas Smith.	Hill Hole	April 28th	100	J. W. Andrews
Thomas Grafton	Newton.	May 1st	67	J. W. Andrews.
...James Goodwin Quinn	The Alter ...mpton Park	June 15th	64	J. W. Andrews
John Mellichip	The Piper's Hole Ager Hill	June 19th	10	J. W. Andrews.
Elizabeth Bather	The Valletts Ager Hill	June 23d	60	W. Loyatt.

Original Image

Background Subtracted

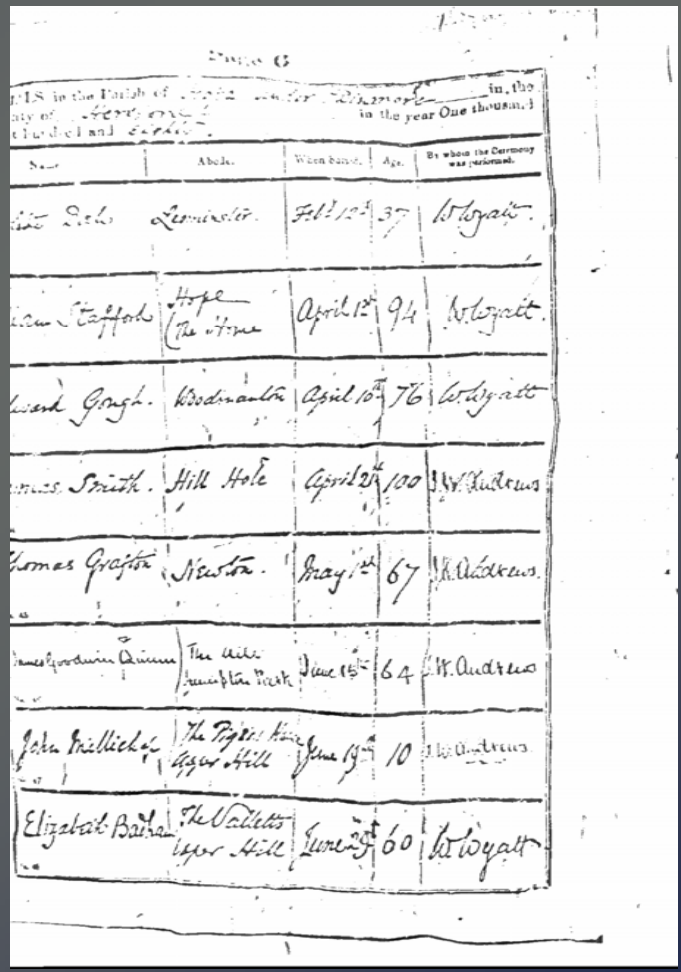


Page 6

...IS in the Parish of Hope under Dinmore in the ...ty of Hereford in the year One thousand ...t hundred and Eighty.

Name.	Abode.	When buried.	Age.	By whom the Ceremony was performed.
...ste Dale	Leominster	Feb.y 12th	37	W. Wyatt.
...iam Stafford	Hope (The Home	April 12th	94	W. Wyatt.
...ward Gough.	Woodmanton	April 16th	76	W. Wyatt
...mes Smith.	Hill Hole	April 28th	100	W. Andrews
...homas Grafton	Newton.	May 1st	67	W. Andrews.
...James Goodwin Quinn	The Alley Hampton Park	June 15th	64	W. Andrews
John Mullich...n	The Tyger's Head Upper Hill	June 19th	10	W. Andrews.
Elizabeth Bathan	The Valletts Upper Hill	June 23d	60	W. Wyatt.

Original Image

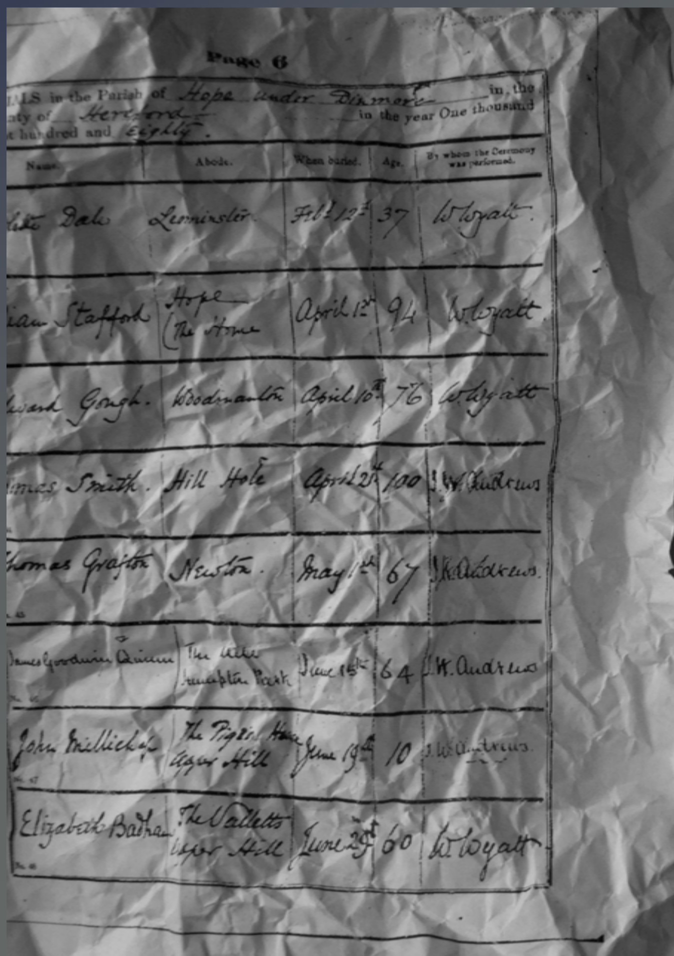


Page 6

...IS in the Parish of ... in the ...ty of ... in the year One thousand ...t hundred and Eighty.

Name.	Abode.	When buried.	Age.	By whom the Ceremony was performed.
...ste Dale	Leominster	Feb.y 12th	37	W. Wyatt.
...iam Stafford	Hope (The Home	April 12th	94	W. Wyatt.
...ward Gough.	Woodmanton	April 16th	76	W. Wyatt
...mes Smith.	Hill Hole	April 28th	100	W. Andrews
...homas Grafton	Newton.	May 1st	67	W. Andrews.
...James Goodwin Quinn	The Alley Hampton Park	June 15th	64	W. Andrews
John Mullich...n	The Tyger's Head Upper Hill	June 19th	10	W. Andrews.
Elizabeth Bathan	The Valletts Upper Hill	June 23d	60	W. Wyatt.

S1

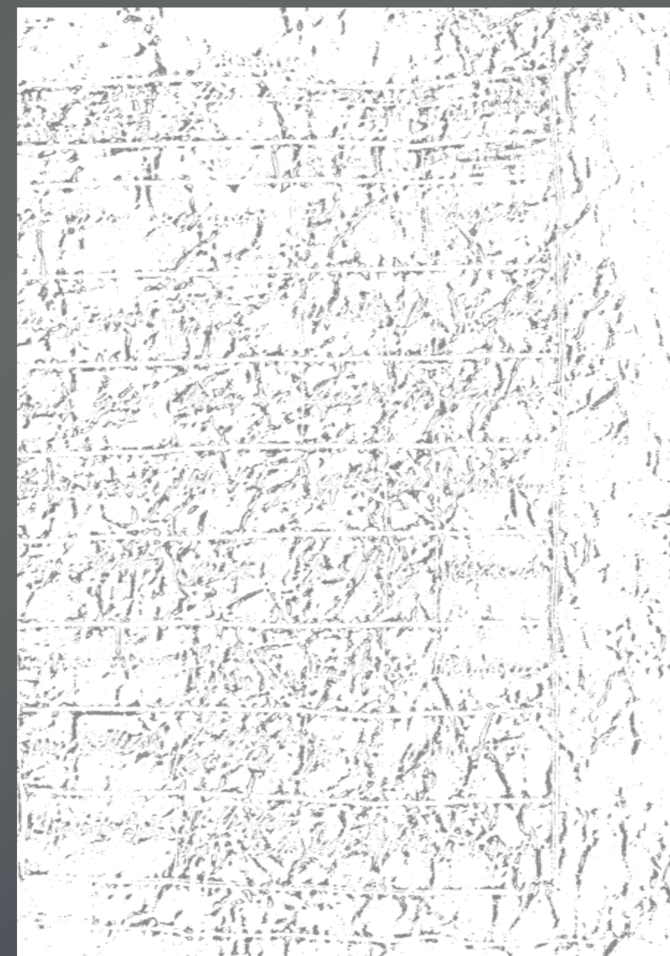


Page 6

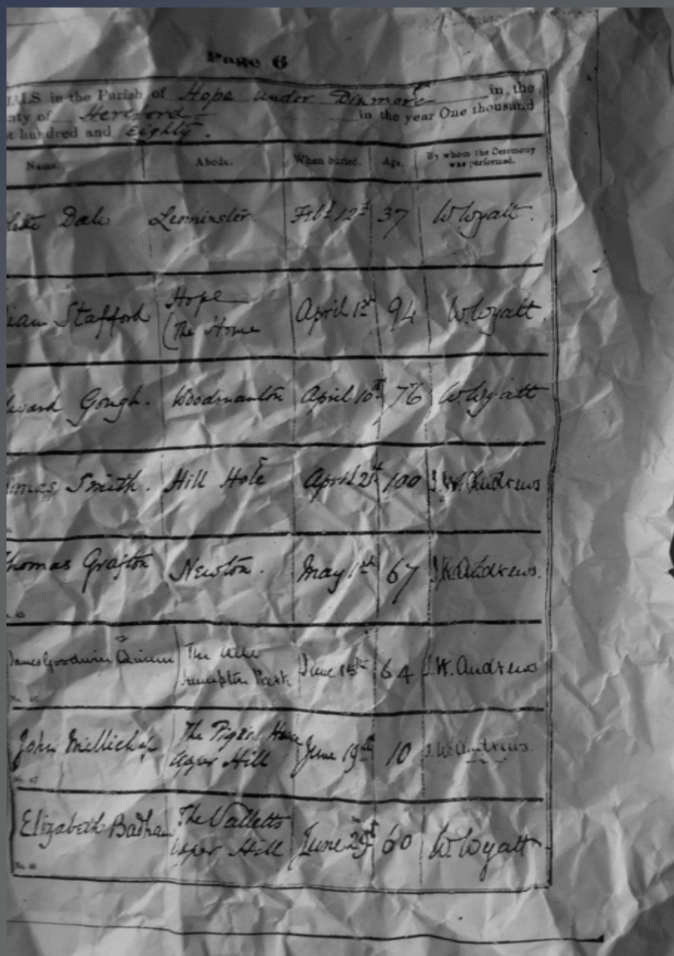
…ALS in the Parish of Hope under Dinmore in the …nty of Hereford in the year One thousand …t hundred and Eighty.

Name.	Abode.	When buried.	Age.	By whom the Ceremony was performed.
…Dale	Leominster	Feb 12th	37	W. Wyatt
…Stafford	Hope (The Home	April 12th	94	W. Wyatt
…ward Gough.	Woodmanton	April 16th	76	W. Wyatt
…mes Smith.	Hill Hole	April 21st	100	J. W. Andrews
Thomas Grafton	Newton.	May 1st	67	J. W. Andrews
James Goodwin [illegible]	[illegible]	June 15th	64	J. W. Andrews
John Millichap	The Pigeon House Upper Hill	June 19th	10	J. W. Andrews
Elizabeth Badham	The Valletts Upper Hill	June 29th	60	W. Wyatt

Original Image



S3



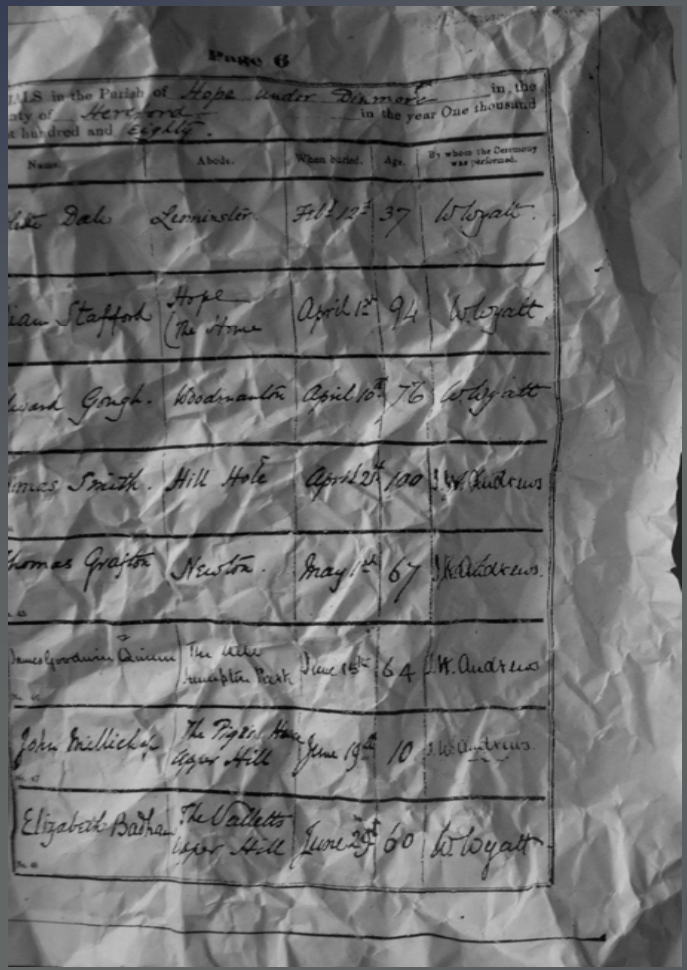
Page 6

...ALS in the Parish of Hope under Dinmore in the ...nty of Hereford in the year One thousand ...t hundred and Eighty.

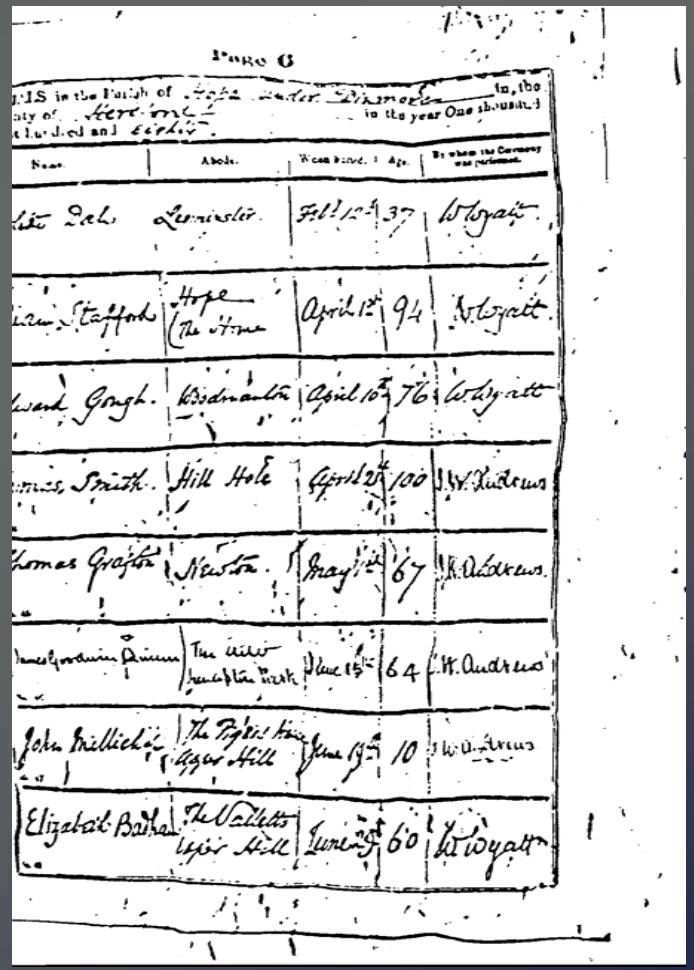
Name.	Abode.	When buried.	Age.	By whom the Ceremony was performed.
...[illegible] Dale	Leominster	Feb 12th	37	W Wyatt
...iam Stafford	Hope (The Home	April 12th	94	W Wyatt
...ward Gough	Woodmanton	April 16th	76	W Wyatt
...mes Smith	Hill Hole	April 21st	100	J.W. Andrews
Thomas Grafton	Newton	May 1st	67	J.W. Andrews
James Goodwin [illegible]	The [illegible]	June 15th	64	J.W. Andrews
John Millichap	The Pigeon House Upper Hill	June 19th	10	J.W. Andrews
Elizabeth Batham	The Vallets Upper Hill	June 29th	60	W Wyatt

Original Image

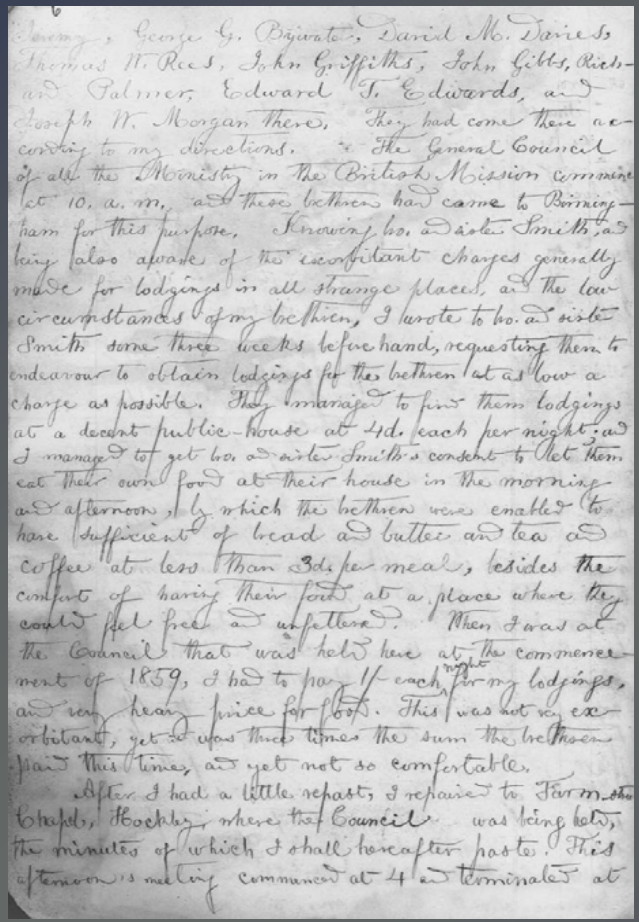
S3



Original Image



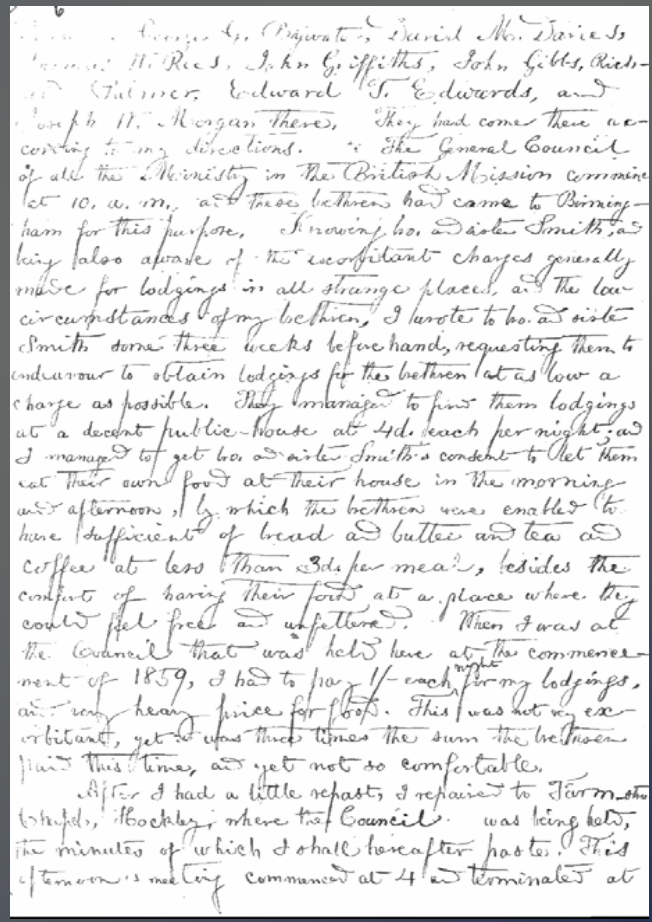
S1 + S2 + S3



George G. Bywater, David M. Davies, Thomas W. Rees, John Griffiths, John Gibbs, Richard Palmer, Edward T. Edwards, and Joseph W. Morgan there. They had come there according to my directions. The General Council of all the Ministry in the British Mission commenced at 10. a. m., as those brethren had came to Birmingham for this purpose. Knowing bro. and sister Smith, and being also aware of the exorbitant charges generally made for lodgings in all strange places, and the low circumstances of my brethren, I wrote to bro. and sister Smith some three weeks before hand, requesting them to endeavour to obtain lodgings for the brethren at as low a charge as possible. They managed to find them lodgings at a decent public-house at 4d. each per night; and I managed to get bro. and sister Smith's consent to let them eat their own food at their house in the morning and afternoon; by which the brethren were enabled to have sufficient of bread and butter and tea and coffee at less than 3ds. per meal, besides the comfort of having their food at a place where they could feel free and unfettered. When I was at the Council that was held here at the commencement of 1859, I had to pay 1/- each night for my lodgings, and very hearty price for food. This was not very exorbitant, yet it was three times the sum the brethren paid this time, and yet not so comfortable.

After I had a little repast, I repaired to Farm-stead, Hockley, where the Council was being held, the minutes of which I shall hereafter paste. This afternoon's meeting commenced at 4 and terminated at

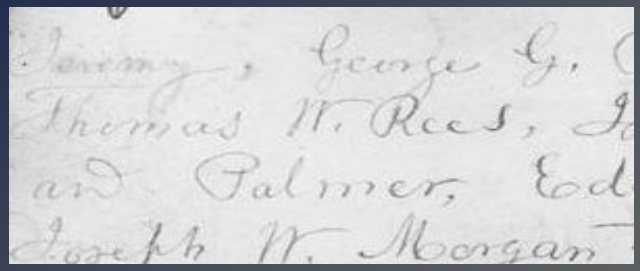
Original Image

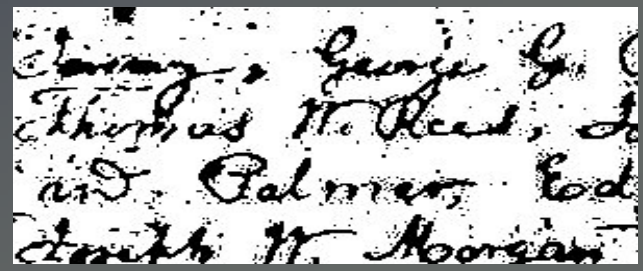


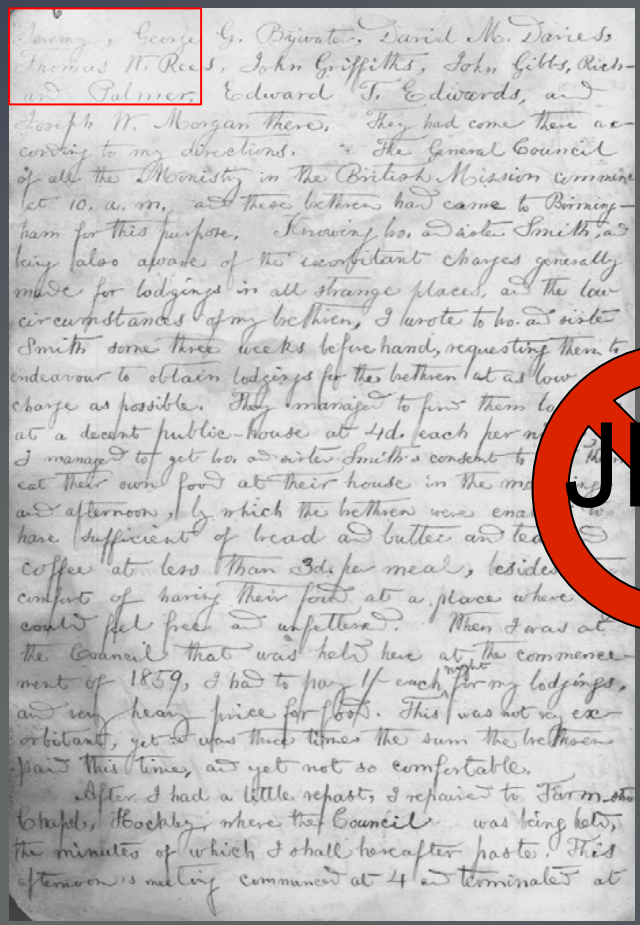
George G. Bywater, David M. Davies, Thomas W. Rees, John Griffiths, John Gibbs, Richard Palmer, Edward T. Edwards, and Joseph W. Morgan there. They had come there according to my directions. The General Council of all the Ministry in the British Mission commenced at 10. a. m., as those brethren had came to Birmingham for this purpose. Knowing bro. and sister Smith, and being also aware of the exorbitant charges generally made for lodgings in all strange places, and the low circumstances of my brethren, I wrote to bro. and sister Smith some three weeks before hand, requesting them to endeavour to obtain lodgings for the brethren at as low a charge as possible. They managed to find them lodgings at a decent public-house at 4d. each per night; and I managed to get bro. and sister Smith's consent to let them eat their own food at their house in the morning and afternoon; by which the brethren were enabled to have sufficient of bread and butter and tea and coffee at less than 3ds. per meal, besides the comfort of having their food at a place where they could feel free and unfettered. When I was at the Council that was held here at the commencement of 1859, I had to pay 1/- each night for my lodgings, and very hearty price for food. This was not very exorbitant, yet it was three times the sum the brethren paid this time, and yet not so comfortable.

After I had a little repast, I repaired to Farm-stead, Hockley, where the Council was being held, the minutes of which I shall hereafter paste. This afternoon's meeting commenced at 4 and terminated at

S1







JPG

Original Image

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?