

# Towards Searchable Indexes for Handwritten Documents

Douglas J. Kennard  
and  
William A. Barrett

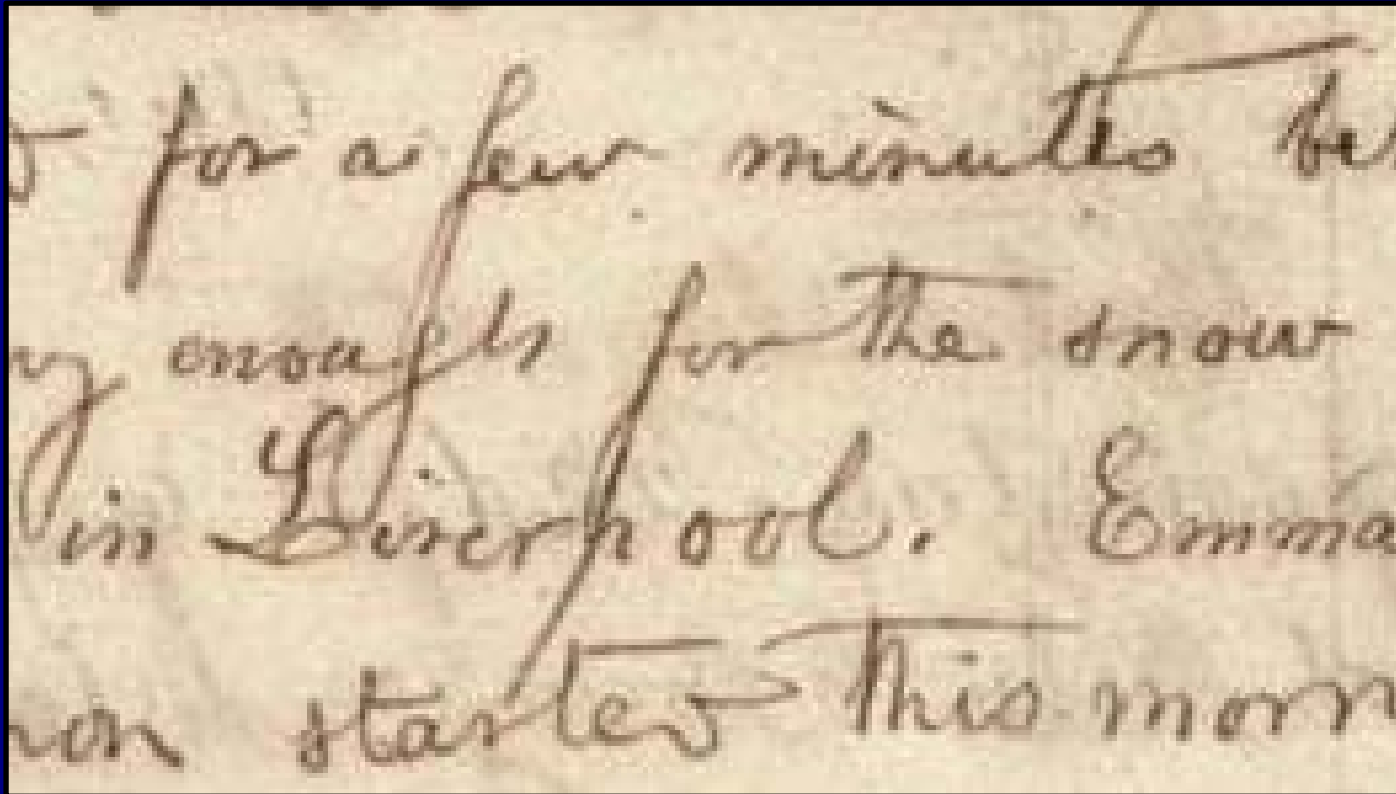
BYU Computer Science Department

Goal: Ability to “search” handwritten documents

Transcriptions are created manually:

- Time-consuming
- Costly

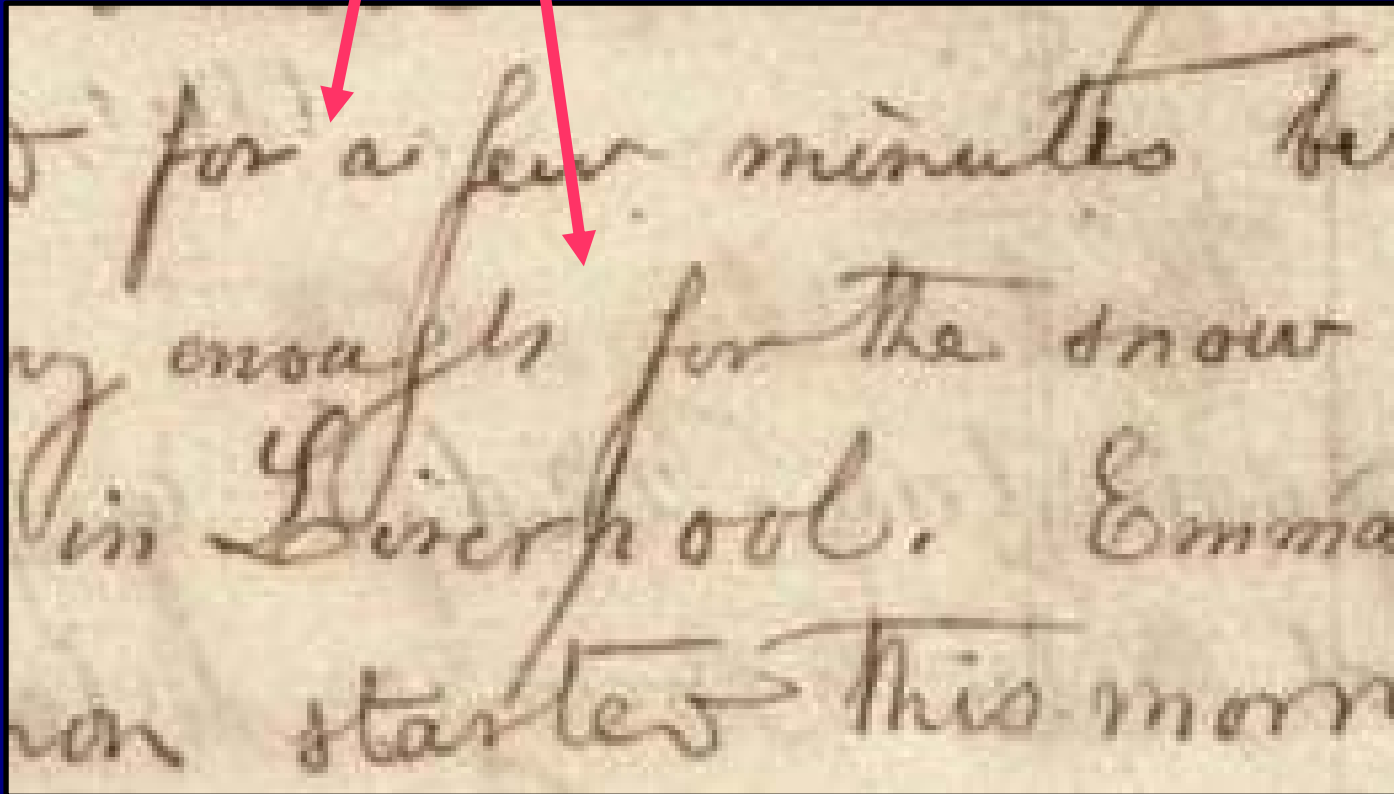
# Difficulties in Automatic Handwriting Recognition



“Trails of Hope: Overland Diaries and Letters, 1846-1869” (BYU Library online collection)

# Difficulties in Automatic Handwriting Recognition

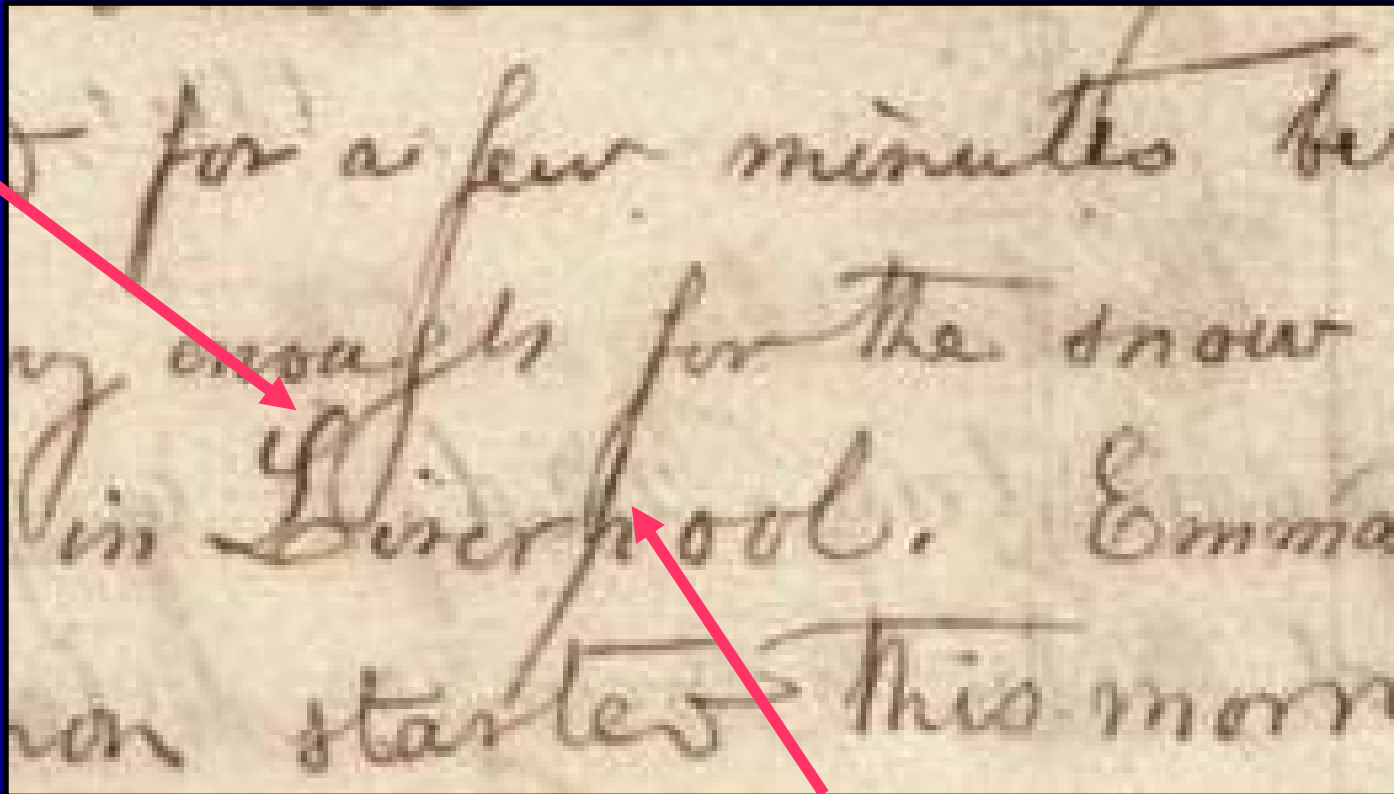
inconsistent spacing



“Trails of Hope: Overland Diaries and Letters, 1846-1869” (BYU Library online collection)

# Difficulties in Automatic Handwriting Recognition

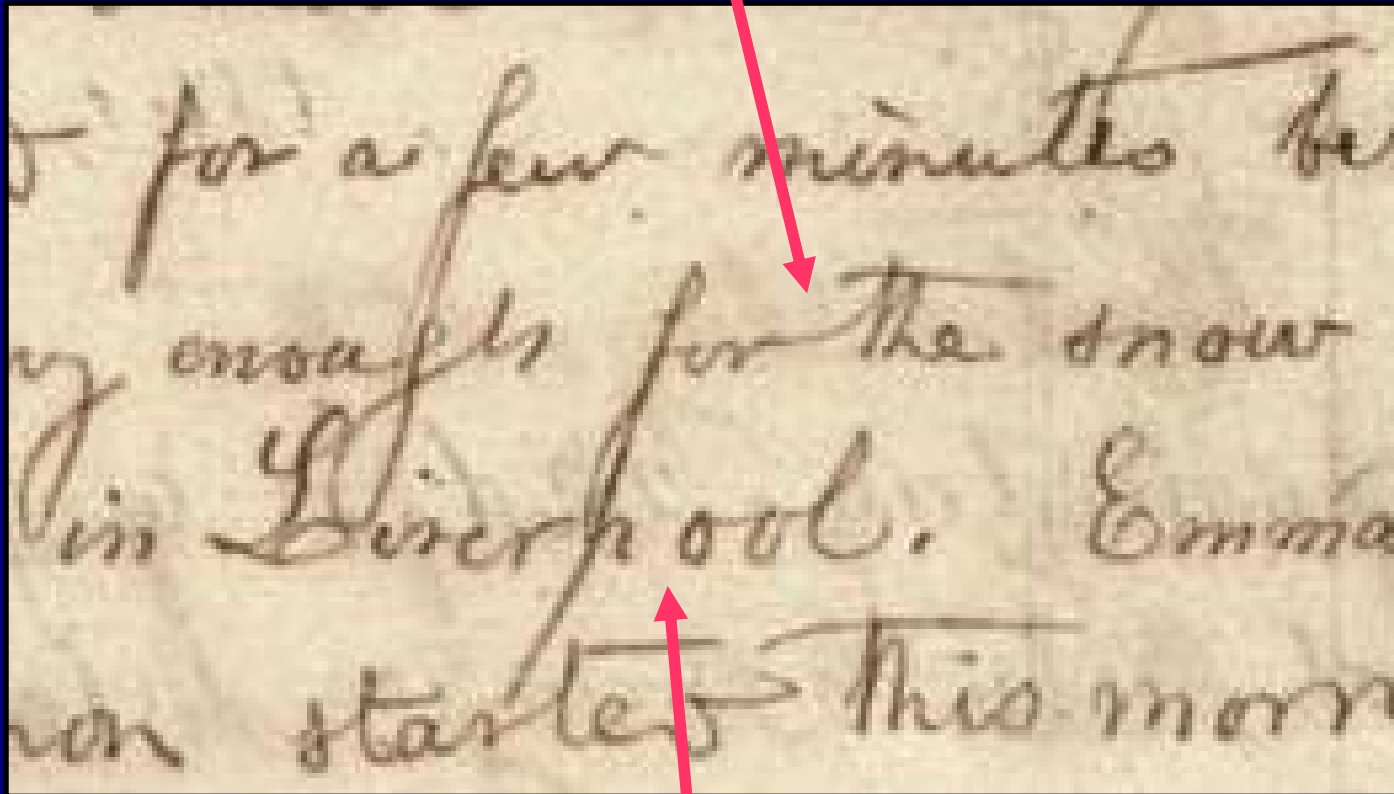
Ascenders/Descenders  
touching other lines of text



“Trails of Hope: Overland Diaries and Letters, 1846-1869” (BYU Library online collection)

# Difficulties in Automatic Handwriting Recognition

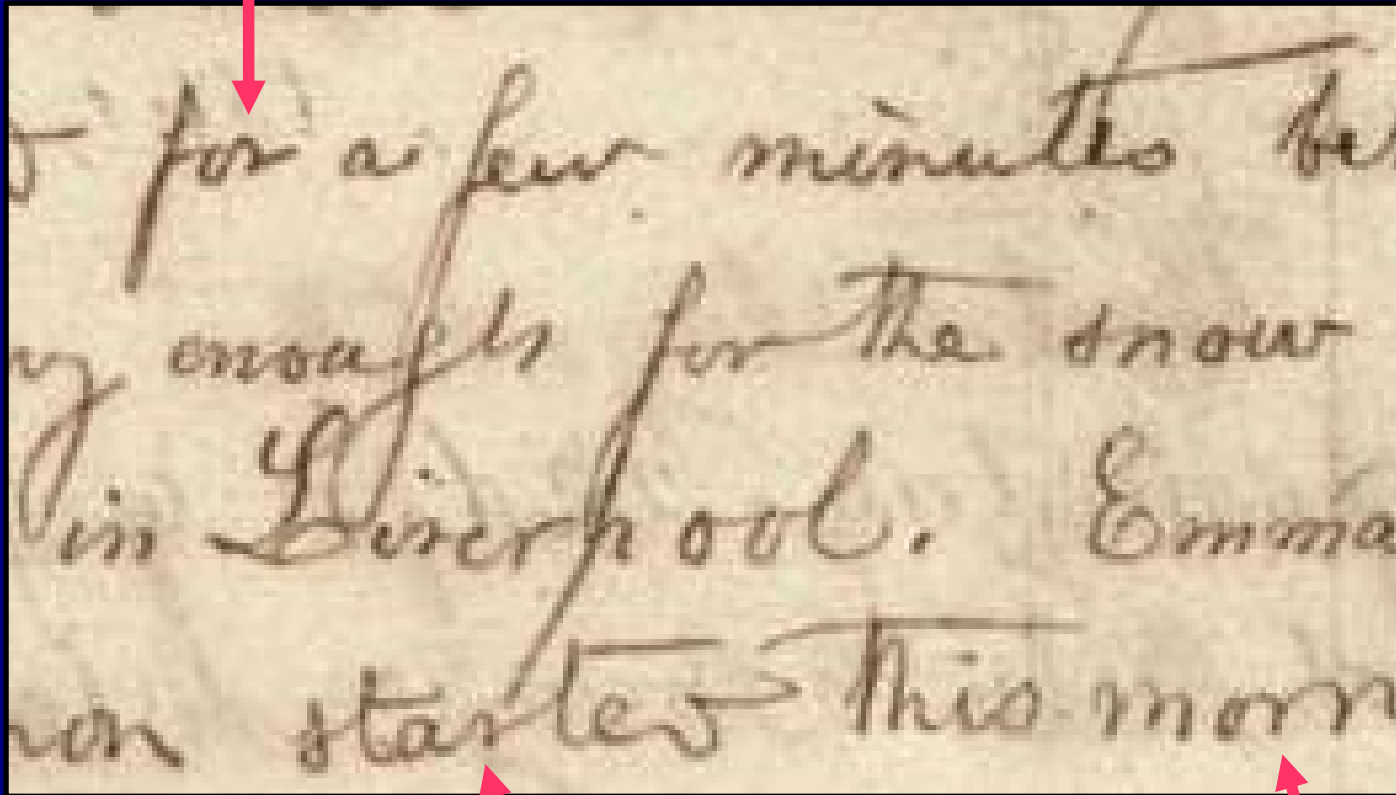
No space between words,  
space within a single word



“Trails of Hope: Overland Diaries and Letters, 1846-1869” (BYU Library online collection)

# Difficulties in Automatic Handwriting Recognition

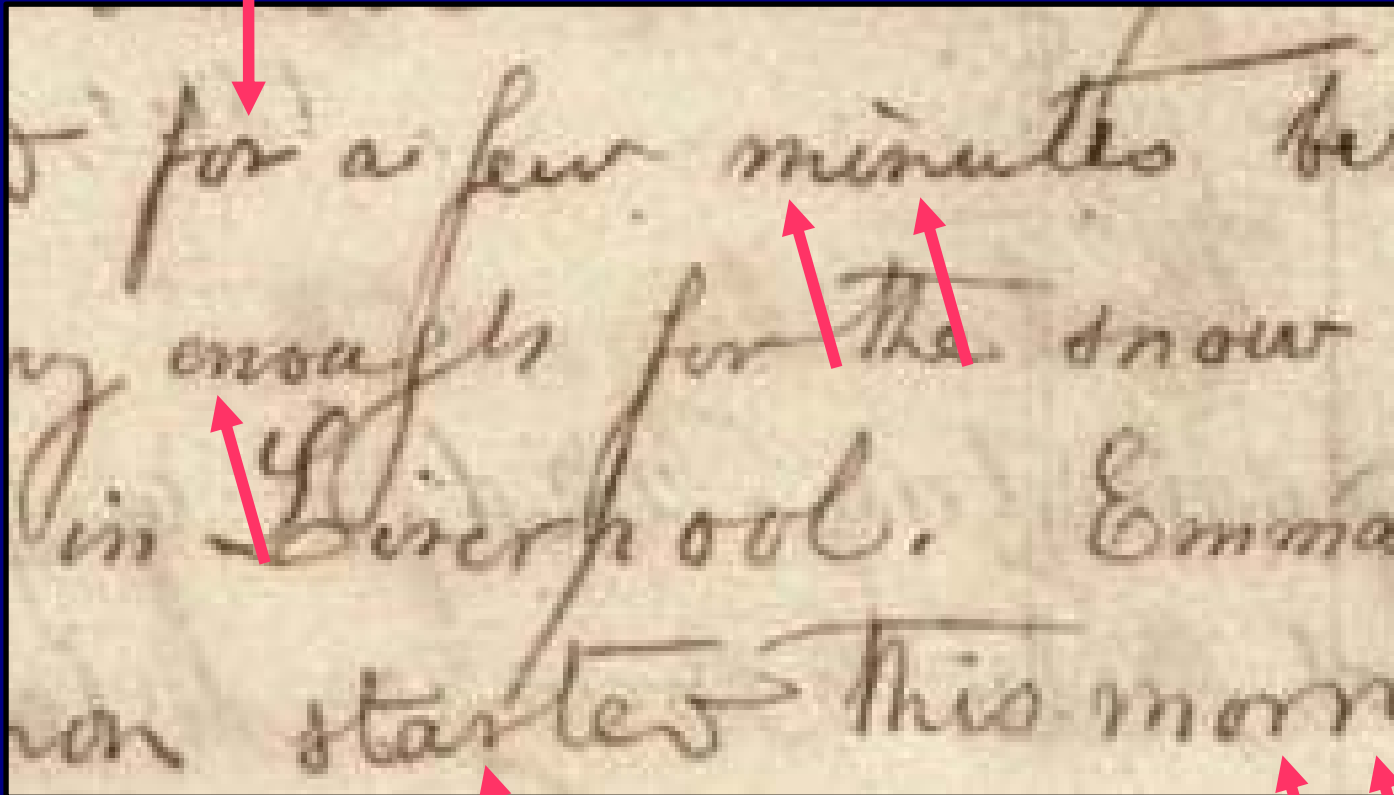
Same letter shaped differently



“Trails of Hope: Overland Diaries and Letters, 1846-1869” (BYU Library online collection)

# Difficulties in Automatic Handwriting Recognition

Different letters shaped  
similarly (n, m, r, ...)



“Trails of Hope: Overland Diaries and Letters, 1846-1869” (BYU Library online collection)



# Difficulties in Automatic Handwriting Recognition

## Other Problems:

Undulating / curved lines

Poor penmanship

Digitization artifacts / lens distortion

Faded ink

Smears, blobs, uneven background

Deteriorated pages

Bleed-through / shine-through

## Conclusion: Handwriting Recognition is Hard!

# A Small Sampling of HR Approaches:

## Dynamic Programming

- Split words into segments
- Use DP to match letters to the segments

## Hidden Markov Models

- Hidden states representing “letters of a possible interpretation”
- Probability of state transitions producing the observed features

## Human Reading Models

- Top-down and Bottom-up combined
- We can't fully segment without some recognition, can't fully recognize without segmentation.

## Holistic (word-level) Features

- Avoid segmenting words

(See references in syllabus)

# Perfect Transcriptions Aren't Necessary

Work done by researchers in France:

- Automatic “annotation”
- Made Available Online
- Users correct errors as they find them

# Handwriting Recognition is Still Hard!

\_i\_e

five

live

time

dime

jive

hive

⋮

\_on\_

bone

gone

pony

⋮

What are these words?  
(recognition / transcription)

# Handwriting Recognition is Still Hard!

\_i\_e      \_on\_

Find the word “lime”

(We don't need a transcription,  
just a “search” for probable matches.)

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop

[http://ciir.cs.umass.edu/cgi-bin/irdemo/handwriting-demo/retrieve\\_1word.pl](http://ciir.cs.umass.edu/cgi-bin/irdemo/handwriting-demo/retrieve_1word.pl)

Search

Print

Home Bookmarks Members WebMail Connections Biz Journal SmartUpdate Mktplace

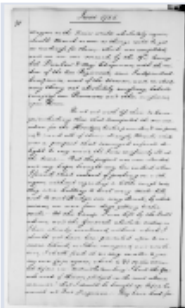
Query (one word only): provision

Search

Query confidence scores ([what's this?](#)):

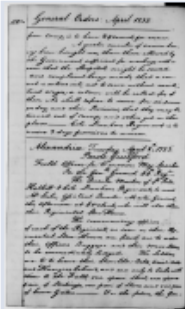
provision

Ranked result list: on the left is the matching page (click to enlarge) and on the right you can see the matching document portion with the query term centered in the middle (click to display the page with the term highlighted)

Result  
1

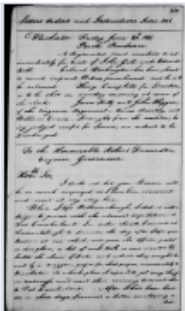
Snippets:

carrying our Provisions and other

Result  
2

Snippets:

3 days provisions to me

Result  
3

Snippets:

three days, I receive a Letter enclosing a list

Handwritten Text Retrieval Demo (c) CIIR 2003 - Mozilla

File Edit View Go Bookmarks Tools Window Help

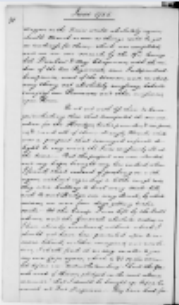
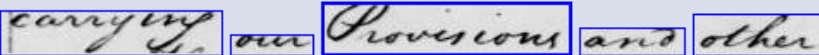
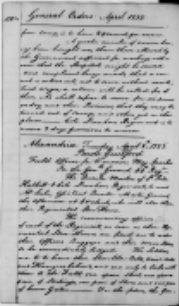
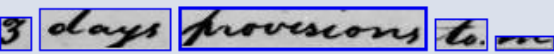
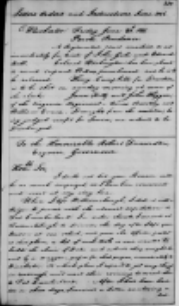

Back Forward Reload Stop  Search Print

Home Bookmarks Members WebMail Connections Biz Journal SmartUpdate Mktplace

Query (one word only):  Search

Query confidence scores ([what's this?](#)):

Ranked result list: on the left is the matching page (click to enlarge) and on the right you can see the matching document portion with the query term centered in the middle (click to display the page with the term highlighted)

Result 1		Snippets: 
Result 2		Snippets: 
Result 3		Snippets: 

Excellent Penmanship  
Relatively "Clean" Images  
100 Pages of Training

# Our Recent Work

Improve Input to HR or Search Systems:

- Improve Text Line Segmentation
- Mark Ambiguities



# Line Segmentation – Simple Profile Method

and all hearts merry, but none more glad than ours.

Next morning 3 of us went back and got each about 40 lbs. of the best meat. The wolves had just taken their first choice of all game; the insides.

The day we reached the upper crossing of the Platt, Aug. 4<sup>th</sup> a high mountain lay in faint view, through the smoke which bedecked the atmosphere, which seemed to be one, some two, some three miles distant, and so on. Next morning in company with my previous hunting companions and one horse I stepped the distance which I found to be about 8 miles as near as we could judge.

It was up hill 6 or 7 miles, where we parted and Gibson and I went together and to the horse with us about 1/4 a mile up the mountain side, as steep as we could climb with the horse and there left him being too steep for him to go farther.

We climbed about one and a half mile farther and were at the summit or near it and thought we were nearer heaven than ever before.



and all hearts merry, but none more glad than ours.

Next morning 3 of us went back and got each about 40 lbs. of the best meat. The wolves had just taken their first choice of all game; the insides.

The day we reached the upper crossing of the Platt, Aug. 4<sup>th</sup> a high mountain lay in faint view, through the smoke which bedecked the atmosphere, which seemed to be one, some two, some three miles distant, and so on. Next morning in company with my previous hunting companions and one horse I stepped the distance which I found to be about 8 miles as near as we could judge.

It was up hill 6 or 7 miles, where we parted and Gibson and I went together and to the horse with us about 1/4 a mile up the mountain side, as steep as we could climb with the horse and there left him being too steep for him to go farther.

We climbed about one and a half mile farther and were at the summit or near it and thought we were nearer heaven than ever before.

# Line Segmentation – Simple Profile Method

view, through the atmosphere, which  
~~seemed to be one, some three~~  
~~some guessed to be one, some three~~  
some three miles distant, and so  
on. Next morning in company  
with my previous hunting companions  
and one horse I stepped the distance  
which I found to be about 8 miles  
as near as we could judge.

# Our Text Line Separation Method

- Preprocess
- Find Locations of Text Lines
- Split / Merge Text Lines
- Output Text Line Images

# Preprocessing: Background Removal

10<sup>14</sup>

January 16<sup>th</sup>, 1862.

Sunday. Commenced putting my scrap-book this evening.

Thursday, 16<sup>th</sup>. Received a letter from John Parry,  
Birmingham, who had written Mr. Ayler. Engaged in reading  
proofs of the Journal with bro. Whittall, and also in  
translating the minutes of the Birmingham Council.  
A very neat little work. It was sent by the Valley Press from Newcastle, by which I found it  
cost 200 pages - less than usual.

Friday, 17<sup>th</sup>. Wrote a letter to bro. J. Griffiths,  
Newcastle, respecting some amount of his labors. Bro. Davies, plasterer,  
Swansea, ~~wrote~~ said he had paid me on the 24<sup>th</sup>  
of March, 1861, being a part of the amount due me by  
the book agency of Swansea Branch. Of this I have  
not the slightest recollection, and I think that bro.  
Davies is mistaken. Emma and I attended a  
party this evening at bro. Whittall's, where we enjoyed  
ourselves well. Each had to pay for his provisions. Among  
the numbers present were bro. S. C. Graham, The Talbot &  
Ross, J. Priestley, J. Bulbridge, Joseph Williams,  
a wife & son. Also, his bro. John, Sisters Corington,  
Sophia Ross, Morris, and Personage were also in the  
company. We did not return home until about ½ past  
3 in the morning.

Saturday, 15th. A  
one. The weather hitherto has been  
of the year. We have had only  
three, to-day, this winter. At  
2, 3, p. m., to-day, but not  
continued in the ground, at least  
equally. Cross Lyman and  
Norwich.

Brigham Young University Harold B. Lee

think that bro  
I attended  
where we enjoy  
provisions. A  
Mr. Faller, &  
Miss M. W.  
Sisters Covington

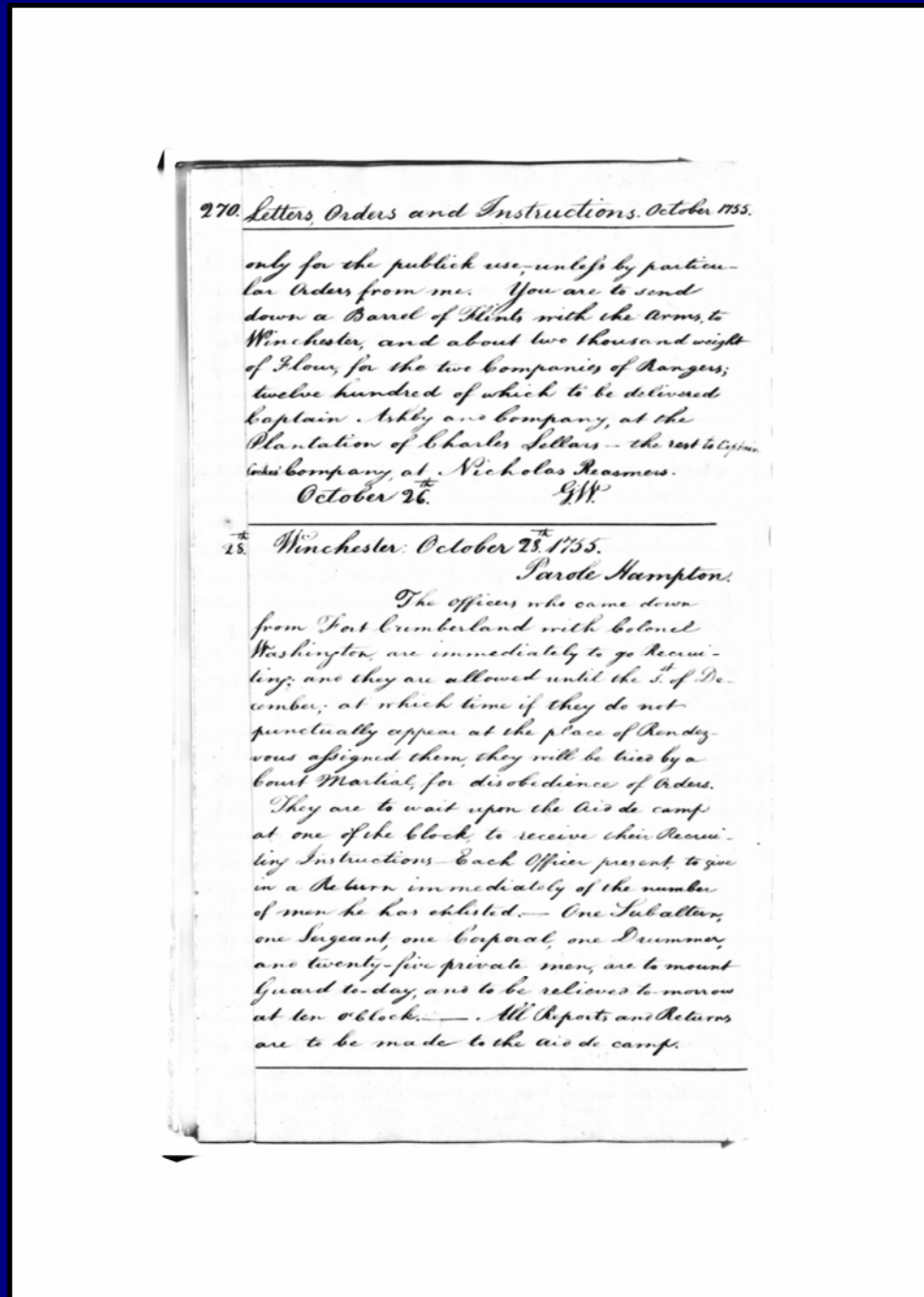
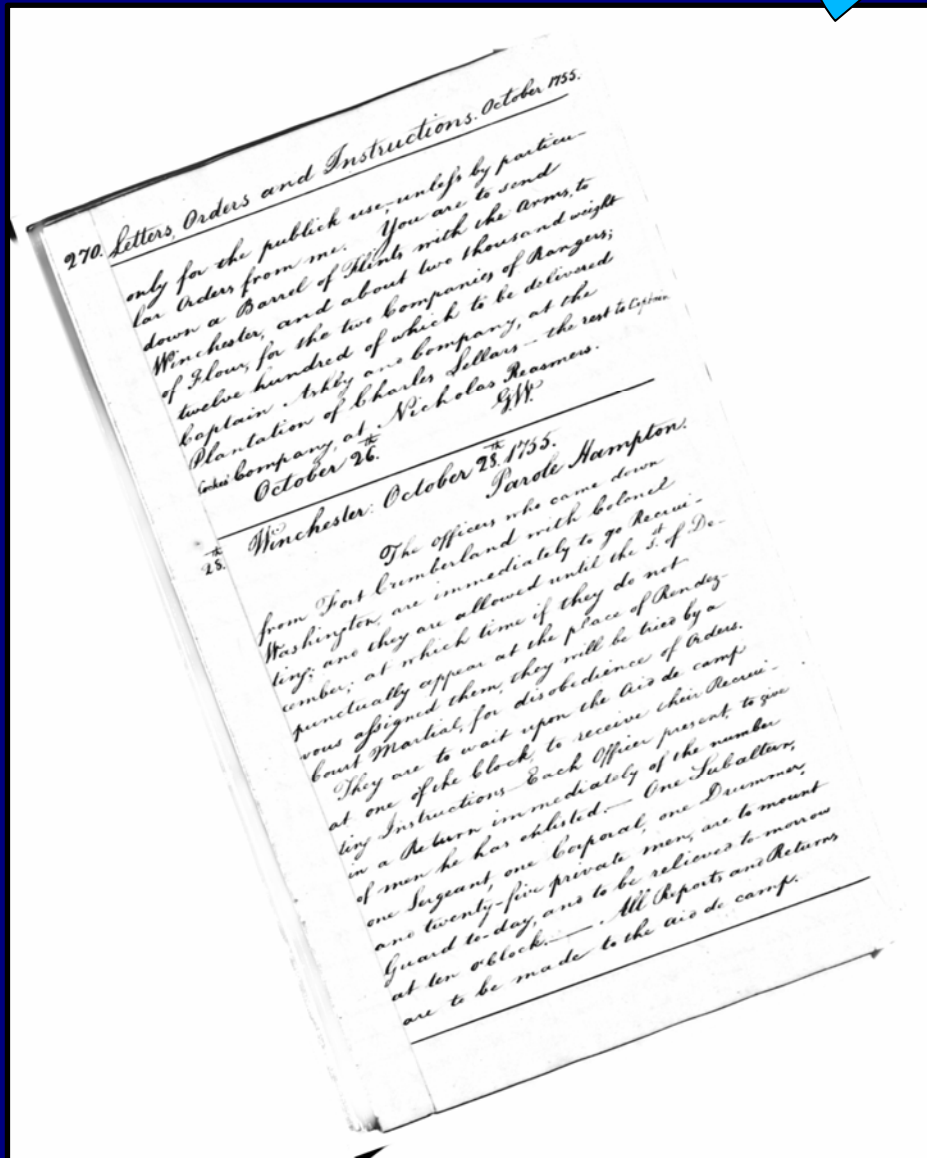
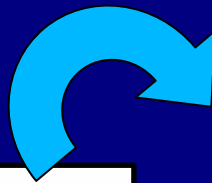
101A  
January 16th, 1862.  
Evening. Commenced putting my scrap-book this evening.  
Tuesday, 16th. Received a letter from John Parry,  
Brynmaer, Merioneth, via Rhyl. Engaged in reading  
proofs of the Journal: wrote bro. Whittall, and also in  
translating the minutes of the Birmingham Council.  
I was most disagreeably surprised to find that the meeting held on the 14th, by which I was  
of 1000 persons, was the least.  
Friday, 17th. Wrote a letter to bro. J. Griffiths,  
Morristown, respecting the same, which bro. Wm. Davies, plasterer,  
Swansea, ~~sent~~ said he had paid me on the 24th  
of March, 1861, being a part of the amount due me by  
the book-agency of Swansea Branch. Of this I have  
not the slightest recollection, and I think that bro.  
Davies is mistaken. Emma and I attended a  
party this evening at bro. Whittall's, where we enjoyed  
ourselves well. Each had to pay for his provisions. Among  
the numbers present were bro. J. Colclough, the Talbot  
Apostle, J. Bradley, J. Tullidge, <sup>Sister</sup> M. Williams,  
his wife & sister. Also, his bro. John, Sister Gorington,  
Sophia Ross, Morris, and Personage were also in the  
throng. We did not return home until about 1/2 past  
3 in the morning.

Saturday, 18th. A  
one. The mother & father had  
of the year. We have had only  
before. In, & this winter. It  
2 44 3, p. m., today, but  
continued on the ground, at le  
equally. Mrs. Lyman and  
or Norwich.

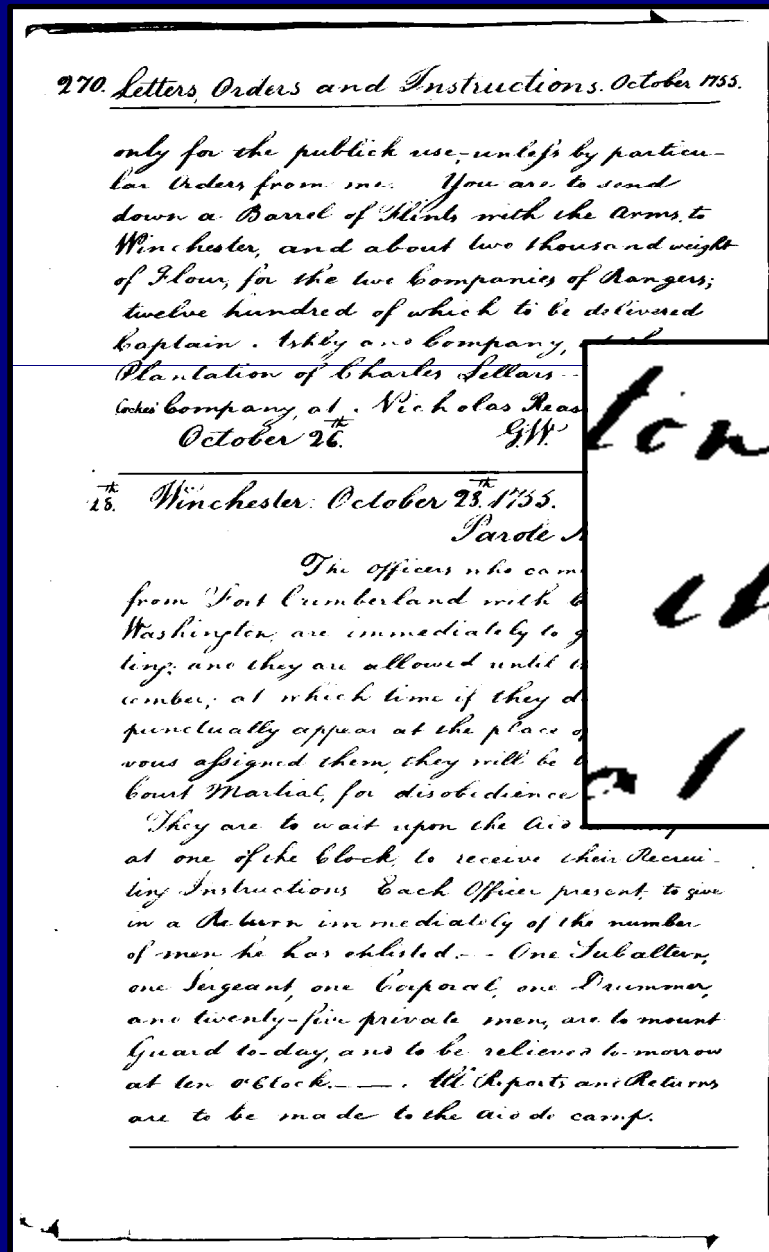
think that bro  
I attended  
where we enjoyed  
provisions. A  
Mr. Faller, the  
share of Miss  
Sisters Covington



# Preprocessing: Deskew Page



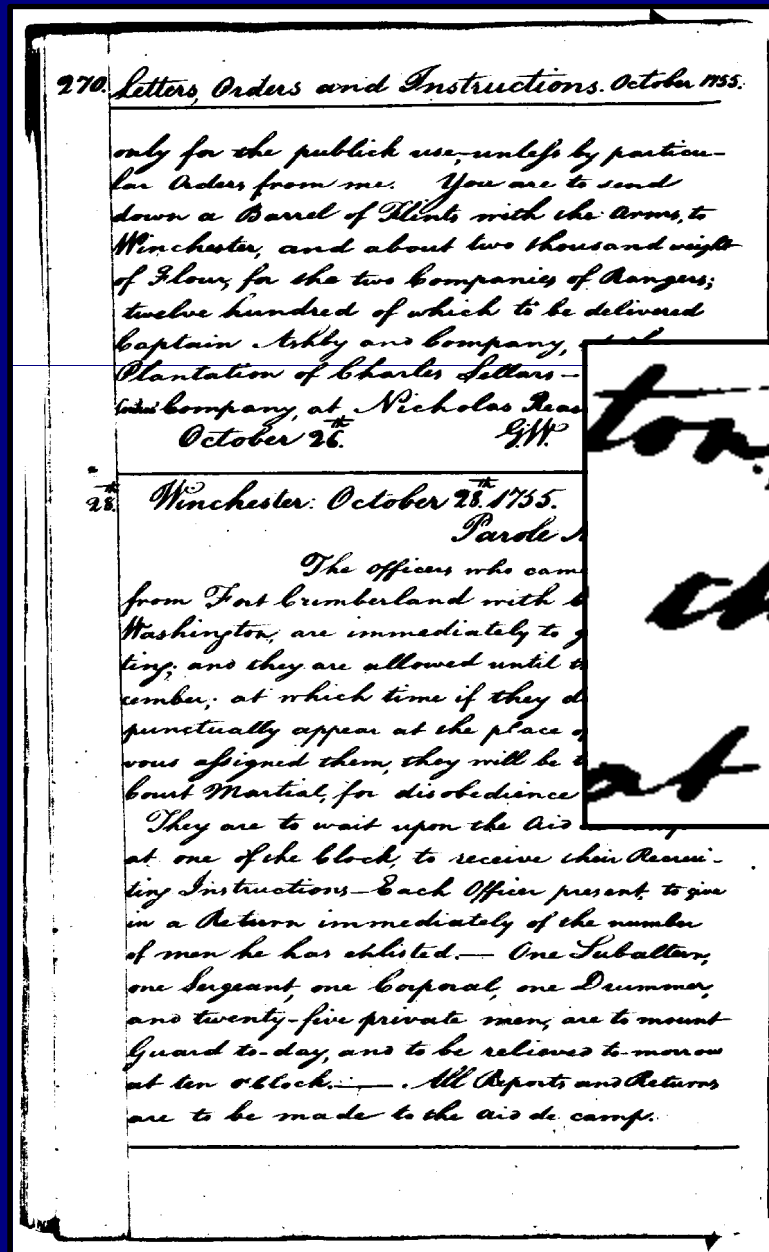
# Preprocessing: Choose Threshold



ton, are immediate  
they are allowed  
at which time

Otsu's Method: Threshold too low

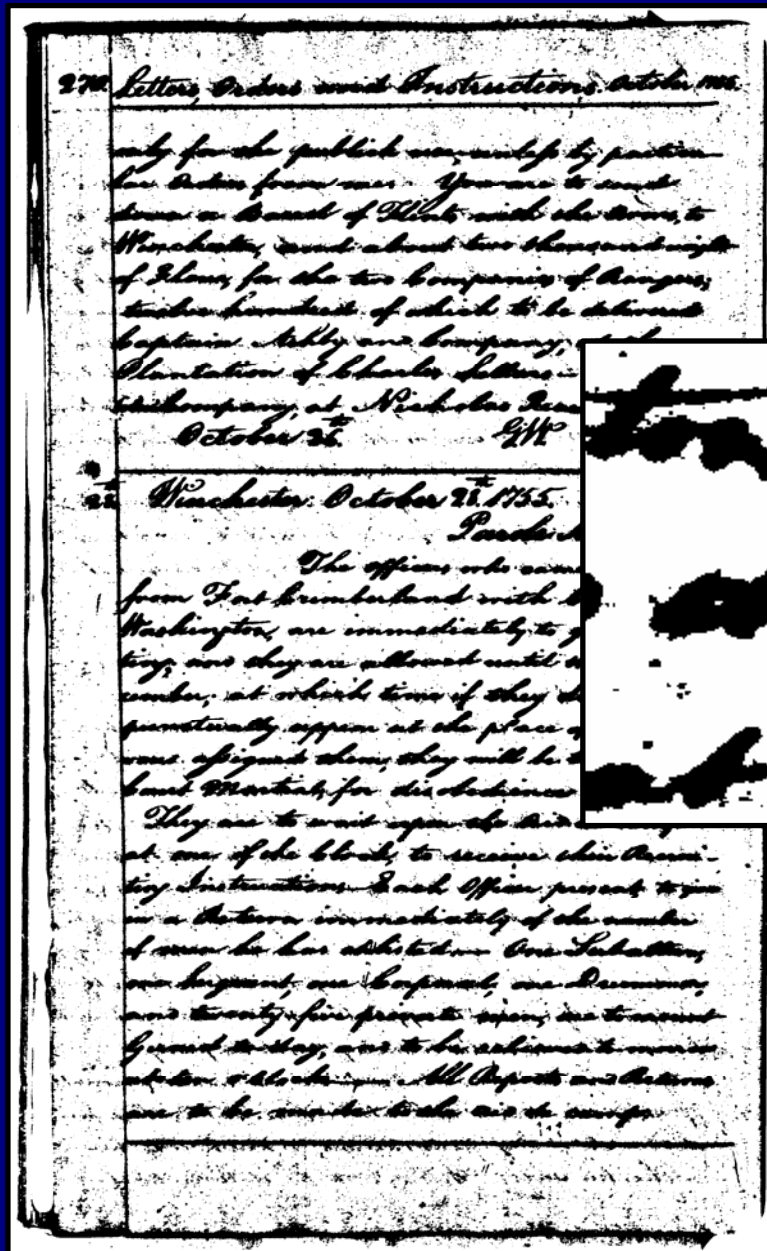
# Preprocessing: Choose Threshold



ton, are imme  
they are allow  
at which time

Good Threshold

# Preprocessing: Choose Threshold

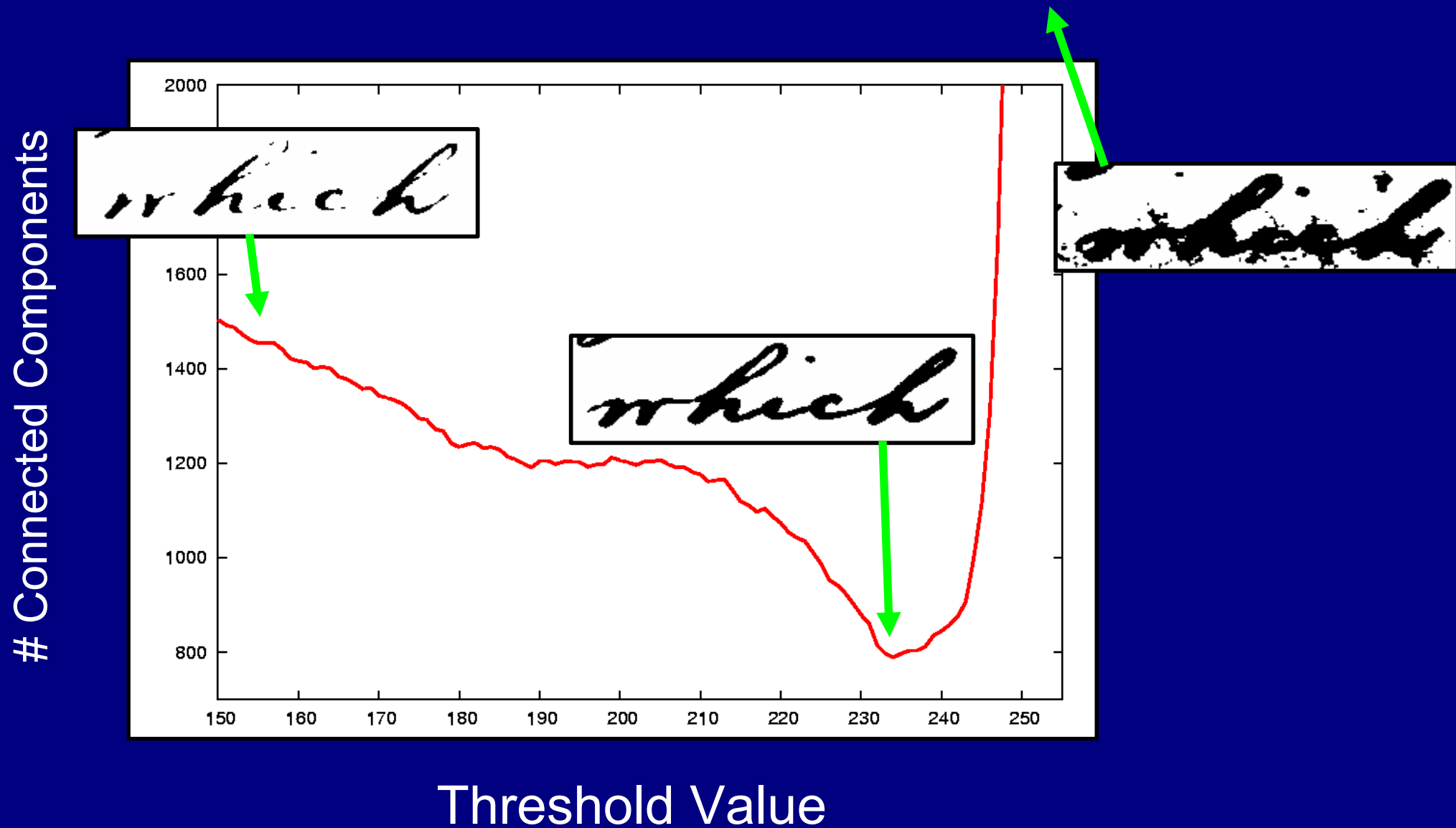


ton, are in  
they are  
at which time

Threshold too high



# Preprocessing: Choose Threshold



# Preprocessing: Remove Rule Lines

270. Letters, Orders and Instructions. October 1755.

only for the publick use, unless by particular Orders from me. You are to send down a Barrel of Flint with the Arms, to Winchester, and about two thousand weight of Flour, for the two Companies of Rangers; twelve hundred of which to be delivered Captain Arkley and Company, at the Plantation of Charles Sellers — the rest to Captain Cooks Company, at Nicholas Neasmons.

October 26. G.H.

25. Winchester. October 23<sup>rd</sup> 1755.

Parole Hampton.

The officers who came down from Fort Cumberland with Colonel Washington, are immediately to go Recruiting; and they are allowed until the 1<sup>st</sup> of December; at which time if they do not punctually appear at the place of Rendezvous assigned them, they will be tried by a Court Martial, for disobedience of Orders.

They are to wait upon the *vis de camp* at one of the block to receive their Recruiting Instructions. Each Officer present, to give in a Return immediately of the number of men he has enlisted. — One Subaltern, one Sergeant, one Corporal, one Drummer, and twenty-five private men, are to mount Guard to-day, and to be relieved to-morrow at ten o'clock. — All Reports and Returns are to be made to the *vis de camp*.

270. Letters, Orders and Instructions. October 1755.

only for the publick use, unless by particular Orders from me. You are to send down a Barrel of Flint with the Arms, to Winchester, and about two thousand weight of Flour, for the two Companies of Rangers; twelve hundred of which to be delivered Captain Arkley and Company, at the Plantation of Charles Sellers — the rest to Captain Cooks Company, at Nicholas Neasmons.

October 26. G.H.

25. Winchester. October 23<sup>rd</sup> 1755.

Parole Hampton.

The officers who came down from Fort Cumberland with Colonel Washington, are immediately to go Recruiting; and they are allowed until the 1<sup>st</sup> of December; at which time if they do not punctually appear at the place of Rendezvous assigned them, they will be tried by a Court Martial, for disobedience of Orders.

They are to wait upon the *vis de camp* at one of the block to receive their Recruiting Instructions. Each Officer present, to give in a Return immediately of the number of men he has enlisted. — One Subaltern, one Sergeant, one Corporal, one Drummer, and twenty-five private men, are to mount Guard to-day, and to be relieved to-morrow at ten o'clock. — All Reports and Returns are to be made to the *vis de camp*.

# Find Lines of Text

and all hearts merry, but none  
more glad than ours.  
Next morning 3 of us went  
back and got each about 90 lbs.  
of the best meat. The workers had  
just taken their first choice of all  
game; the inside.  
The day we reached the upper  
crossing of the Platt, Aug. 4<sup>th</sup>  
a high mountain lay in faint  
view, through the smoke which  
beckoned the atmosphere, which  
some guessed to be one some two,  
some three miles distant. and so

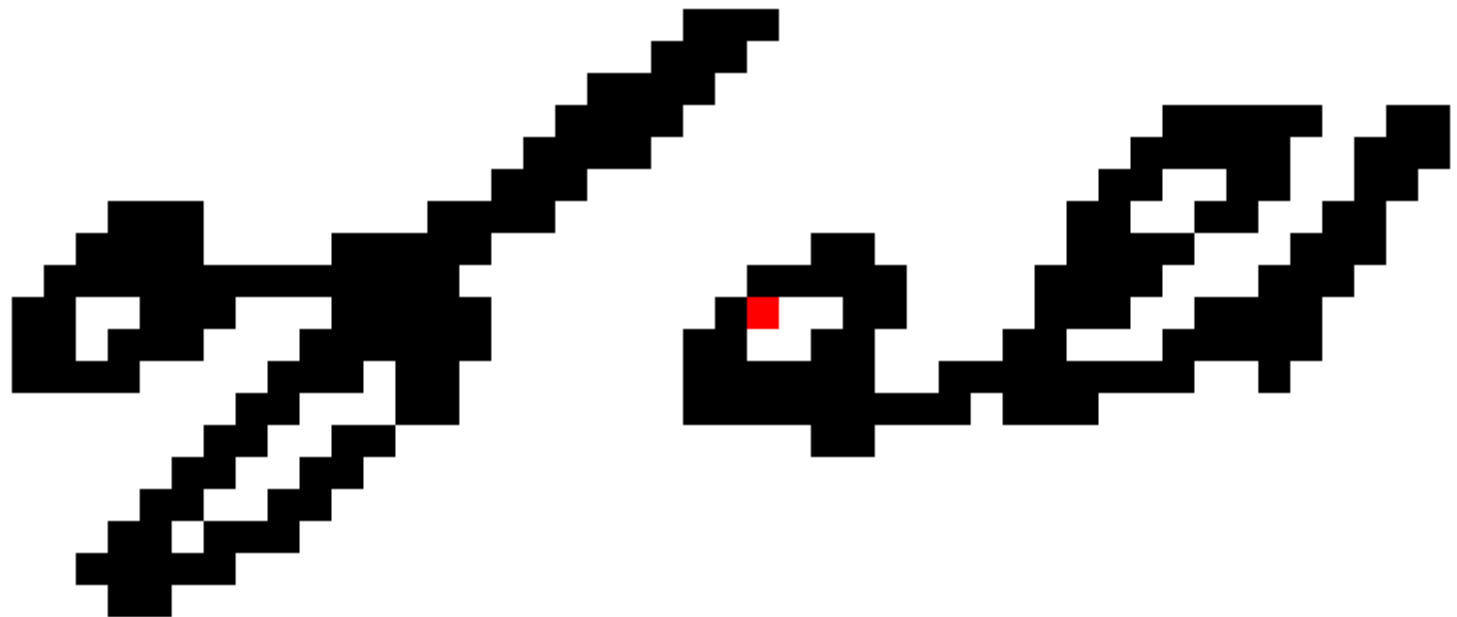


Bitonal (Black / White)

Transition Count Map

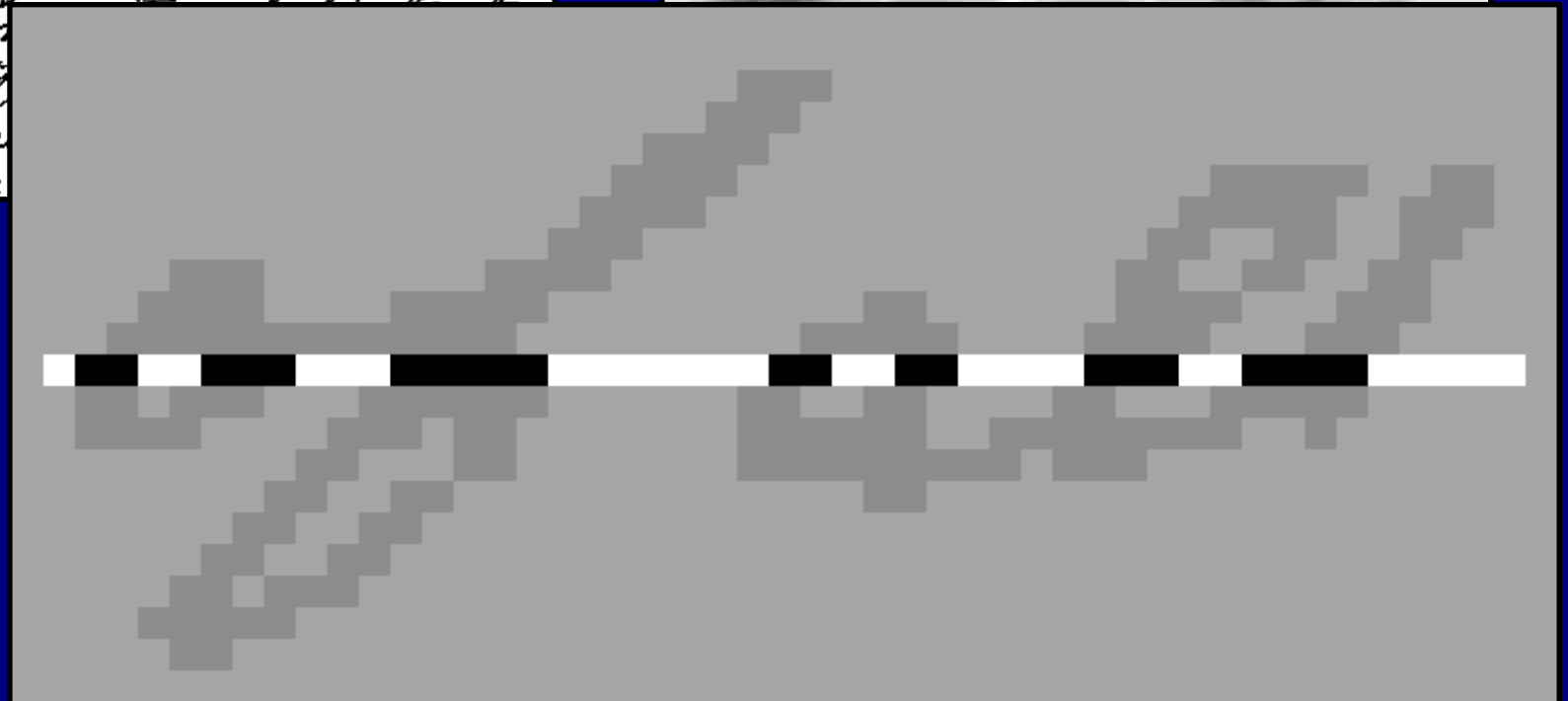
# Find Lines of Text

and all hearts merry, but none  
more glad than ours.  
Next morning 3 of us went  
back and got each about 90 lbs.  
of the best meat. The workers had  
just taken their first choice of all  
game; the inside.  
The day we reached the upper  
crossing of the Platt, Aug. 4<sup>th</sup>  
a high  
view, through  
hillsides  
some grass  
some trees



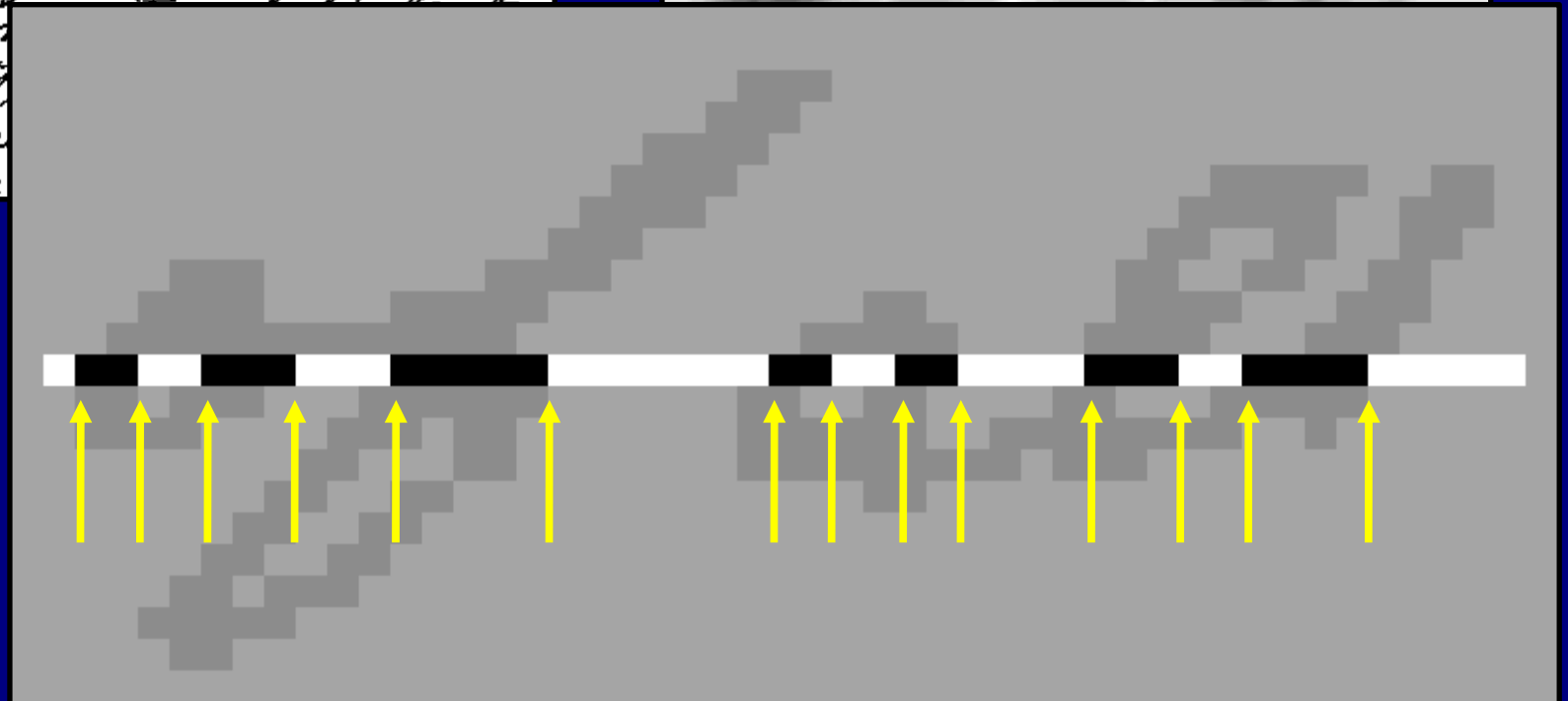
# Find Lines of Text

and all hearts merry, but none  
more glad than ours.  
Next morning 3 of us went  
back and got each about 90 lbs.  
of the best meat. The workers had  
just taken their first choice of all  
game; the inside.  
The day we reached the upper  
crossing of the Platt, Aug. 4<sup>th</sup>  
a high  
view, through  
hundreds  
some grass  
some trees



# Find Lines of Text

and all hearts merry, but none  
more glad than ours.  
Next morning 3 of us went  
back and got each about 90 lbs.  
of the best meat. The workers had  
just taken their first choice of all  
game; the inside.  
The day we reached the upper  
crossing of the Platt, Aug. 4<sup>th</sup>  
a high  
view, through  
hundreds  
some grass  
some trees



# Find Lines of Text

and all hearts merry, but none  
more glad than ours.  
Next morning 3 of us went  
back and got each about 90 lbs.  
of the best meat. The workers had  
just taken their first choice of all  
game; the inside.  
The day we reached the upper  
crossing of the Platt, Aug. 4<sup>th</sup>  
a high mountain lay in faint  
view, through the smoke which  
beckoned the atmosphere, which  
some guessed to be one some two,  
some three miles distant. and so

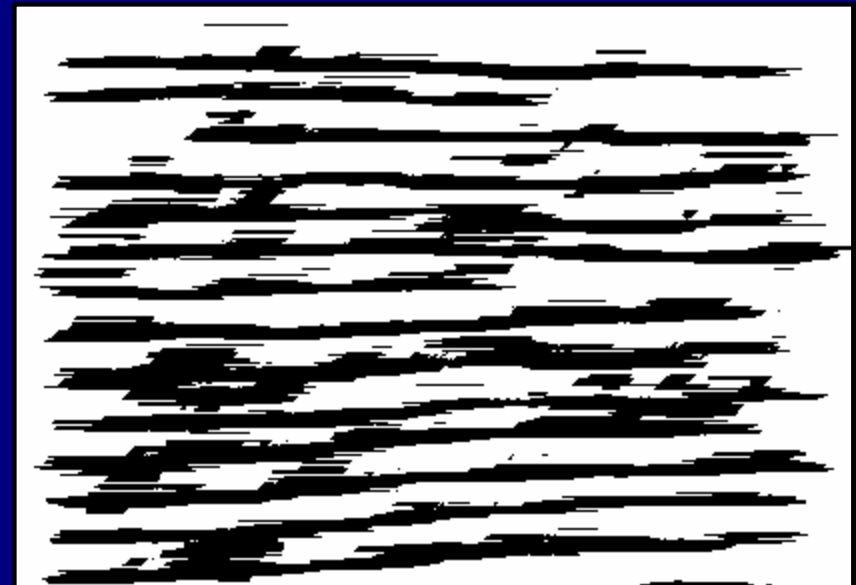


Bitonal (Black / White)

Transition Count Map

# Find Lines of Text

and all hearts merry, but none  
more glad than ours.  
Next morning 3 of us went  
back and got each about 90 lbs.  
of the best meat. The workers had  
just taken their first choice of all  
game; the inside.  
The day we reached the upper  
crossing of the Platt, Aug. 4<sup>th</sup>  
a high mountain lay in faint  
view, through the smoke which  
beckoned the atmosphere, which  
some guessed to be one some two,  
some three miles distant. and so



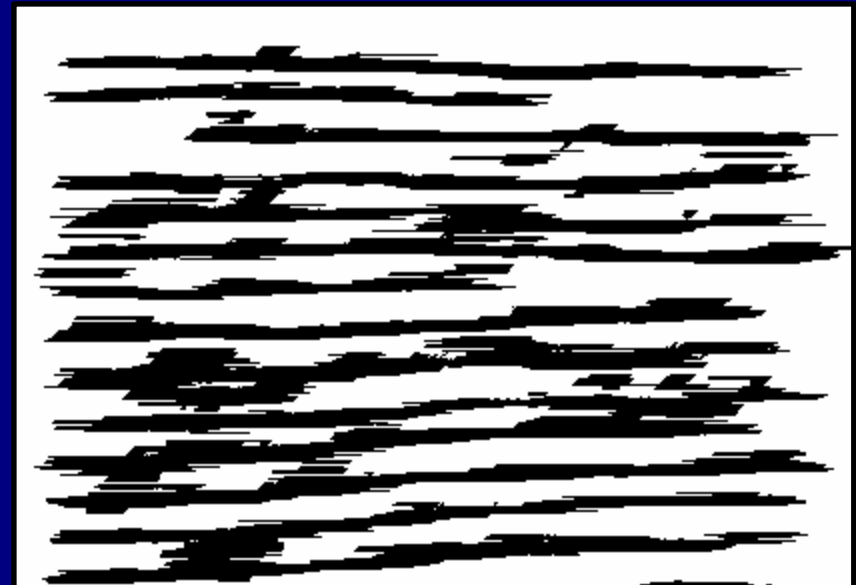
Bitonal (Black / White)

Thresholded  
Transition Count Map



# Find Lines of Text

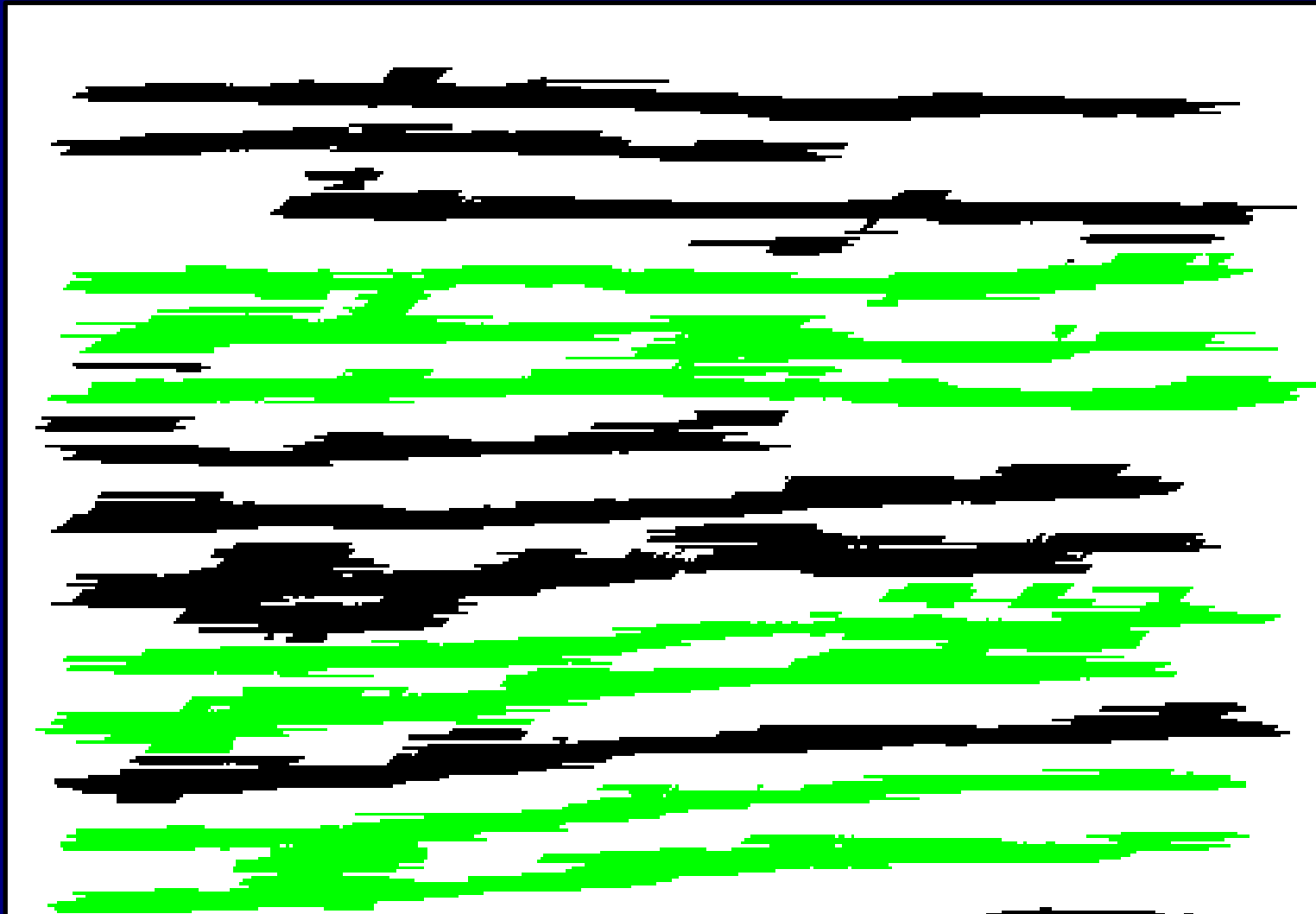
and all hearts merry, but none  
more glad than ours.  
Next morning 3 of us went  
back and got each about 90 lbs.  
of the best meat. The workers had  
just taken their first choice of all  
game; the inside.  
The day we reached the upper  
crossing of the Platt, Aug. 4<sup>th</sup>  
a high mountain lay in faint  
view, through the smoke which  
beckoned the atmosphere, which  
some guessed to be one some two,  
some three miles distant. and so



Bitonal (Black / White)

“Cleaned-Up”  
Transition Count Map  
(small components removed)

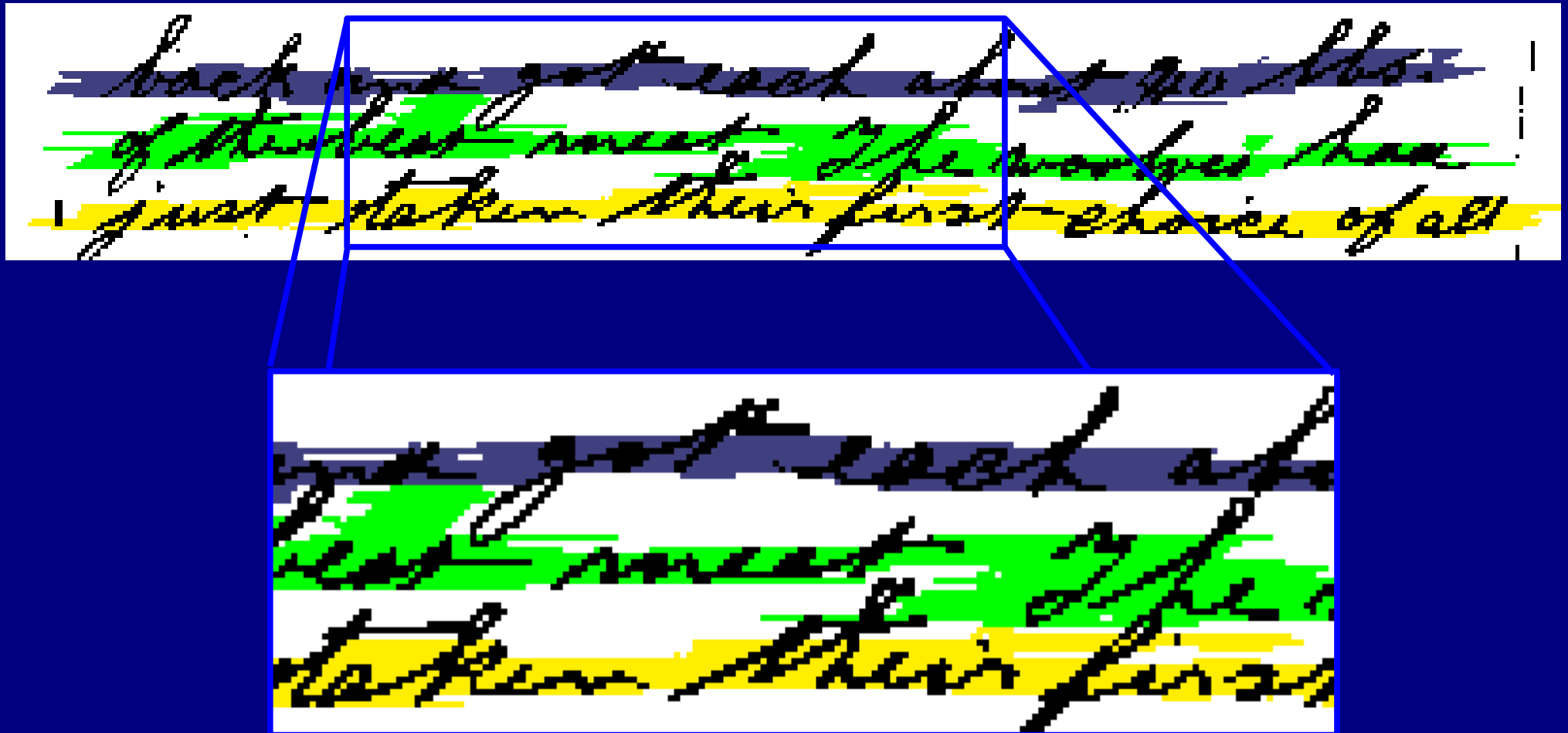
# Split Lines of Text



# Split Lines of Text

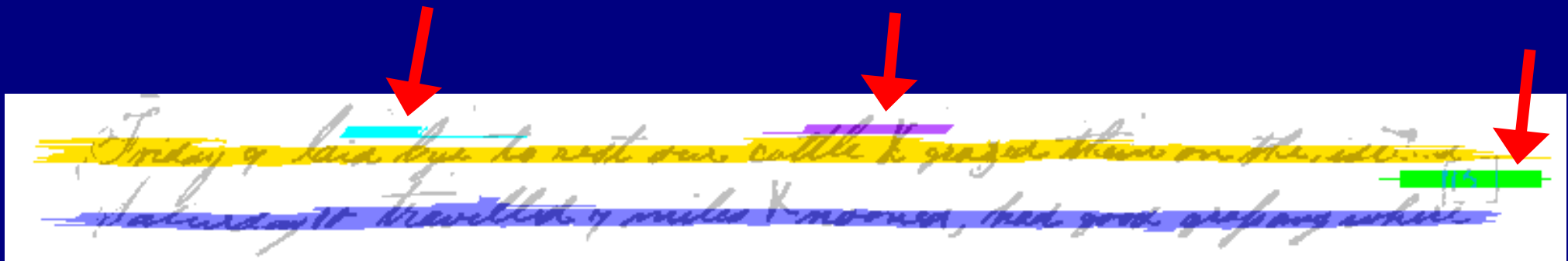


# Split Lines of Text



“Min-Cut / Max-Flow” Graph Cut used iteratively to split lines

# Merge Spurious Lines of Text



Today I laid by to rest our cattle & grazed them on the hill  
Saturday I travelled 7 miles & nooned, had good grazing where

The image shows two lines of handwritten text. The first line is highlighted in yellow and contains the text 'Today I laid by to rest our cattle & grazed them on the hill'. The second line is highlighted in blue and contains the text 'Saturday I travelled 7 miles & nooned, had good grazing where'. Three red arrows point to specific areas: the first arrow points to the word 'laid' in the first line, the second arrow points to the word 'grazed' in the first line, and the third arrow points to the end of the first line, specifically to a small green box containing the number '115'.



Today I laid by to rest our cattle & grazed them on the hill  
Saturday I travelled 7 miles & nooned, had good grazing where

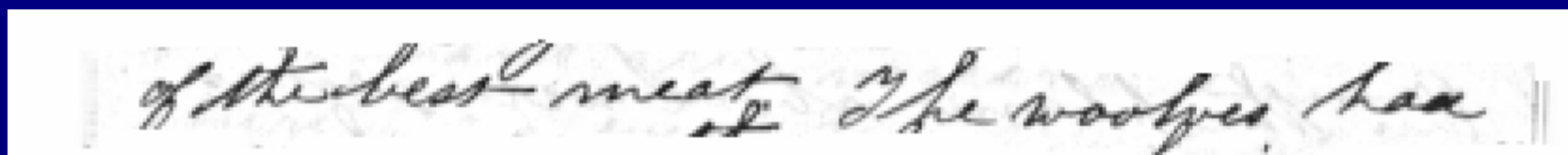
The image shows the same two lines of handwritten text as above. The first line is highlighted in yellow and contains the text 'Today I laid by to rest our cattle & grazed them on the hill'. The second line is highlighted in blue and contains the text 'Saturday I travelled 7 miles & nooned, had good grazing where'. A small green box containing the number '115' is visible at the end of the first line.

# Output Line Images



- Expand component region
- Ignore outside of expanded region
- Anything touching another line component considered ambiguous (within angle constraint)

# Output Line Images

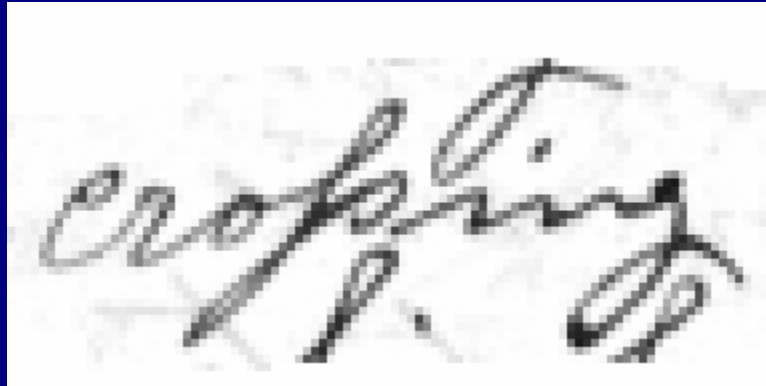


Grayscale Output Image

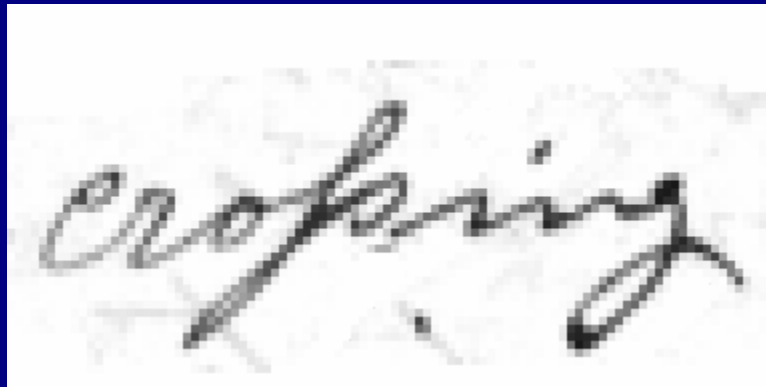


Output Mask Image

# Motivation for Ambiguous component information



?



crossing



# Planned Future Work

Reduce amount of manual training:

- Train interactively instead of transcribing  
(many words get used over and over)

# Planned Future Work

Reduce amount of manual training:

- Train interactively instead of transcribing  
(many words get used over and over)

Example: (from 36 pages of an Overland Trails diary)

“and” = 311 times

“the” = 286 times

6,212 words total

860 distinct words

86% of the total words are redundant!

# Planned Future Work

## Reduce amount of manual training:

- Train interactively instead of transcribing (many words get used over and over)
- Sub-word matching (letters and combinations of letters)
- Existing methods for generating artificial training data

# Conclusions

Current Technology permits searching handwritten documents (at least for good quality, large collections)

Won't work perfectly.

Still very useful— much better than nothing at all!

Current and future work will reduce amount of training needed, and improve accuracy by providing better input to the systems.

# Questions