Image Mining of Historical Manuscripts to Establish Provenance

Bing Hu
Image Mining of Historical Manuscripts to Establish Provenance
Outline

• Introduction
• CK1 Distance Measure (Bilson and Dr. Keogh)
• Search Algorithm
• Experimental Evaluation
• Speed Up Techniques
• Conclusion and Future Work
Introduction

We want to know the provenance of historical manuscripts Why?

- Digitization of books
- Intact books and loose pages
- Media reported the story in 2005

Three Million Dollars in losses

A rare manuscript thief, Edward Forbes Smiley III
Introduction

We want to know the provenance of historical manuscripts Why?

• Cultural transmission

These groups who have spent lots of time and money care about cultural transmission

Several European Digital Libraries

<table>
<thead>
<tr>
<th>DLs</th>
<th>Books</th>
<th>Periods</th>
<th>Web links</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVH</td>
<td>14 000</td>
<td>16th</td>
<td><a href="http://www.bvh.univ-tours.fr">http://www.bvh.univ-tours.fr</a></td>
</tr>
<tr>
<td>Impact</td>
<td>2 200</td>
<td>16th-18th</td>
<td><a href="http://eclipsi.bib.ub.es/imp/impcat.htm">http://eclipsi.bib.ub.es/imp/impcat.htm</a></td>
</tr>
<tr>
<td>Mouriau</td>
<td>1 850</td>
<td>18th</td>
<td><a href="http://www.ornements-typo-mouriau.be/">http://www.ornements-typo-mouriau.be/</a></td>
</tr>
</tbody>
</table>
Introduction

The Hand-Press period starts around 1454

- Hand press
- A hand-press book
- Printing using woodblocks
- Wood curving
- Seal
Introduction

We propose to treat ornamental letters as “fingerprints”

*There is little known on how to compute invariants for indexing documents on this kind of features.

This task is not easy!

Introduction

Some difficulties…
Outline

• Introduction

• **CK1 Distance Measure** (Bilson and Dr. Keogh)

• Search Algorithm

• Experimental Evaluation

• Speed Up Techniques

• Conclusion and Future Work
**CK1 Distance Measure** [1]

**black box function**

\[
\text{CK1\_Dist} = \frac{\text{mpegSize}(x, y) + \text{mpegSize}(y, x)}{\text{mpegSize}(x, x) + \text{mpegSize}(y, y)} - 1;
\]

We use the MPEG-1 encoder to construct a video of two images

Why choose CK1 distance measure?

- Parameter free (can be easily used)
- Very accurate
- Robust to distortions

CK1 Distance Measure

Accurate: CK1 can handle very subtle differences

Why choose CK1 distance measure?

- Parameter free
- Very accurate
- Robust to distortions

Twelve ornamental initial letters, are clustered using CK1 with complete linkage hierarchical clustering
CK1 Distance Measure

An experiment to test the accuracy and robustness of CK1 distance measure

Experiment on benchmark dataset (created by experts, not us):

19 classes, 571 images

<table>
<thead>
<tr>
<th>Method</th>
<th>CK1 Distance</th>
<th>F-RLE [1]</th>
<th>SIFT [2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error Rate</td>
<td>0%</td>
<td>39.42%</td>
<td>10.21%</td>
</tr>
</tbody>
</table>

Resizing from 90% to 110%

Error Rate 3.81%

Rotation from -10 degree to +10 degree

Error Rate 12.21%

CK1 Distance Measure

Robust: It can handle rotated, resized, distorted images

Why choose CK1 distance measure?

- Parameter free
- Very accurate
- Robust to distortions

Invariance to these difficulties
Informal Problem Statement

Annotated initial letter dataset pool with known provenance

(i.e. Reference data)
Brute Force Search

Assume that there is a threshold, such that any distance that is smaller than this threshold has a very high possibility to be a true positive.

\[
\text{if } \text{dist} < \text{threshold} \\
\text{`found true positive`} \\
\text{end}
\]

If there is a true positive on this page, our algorithm will report this true positive.
Brute Force Search

More than 6,000 annotated initial letters

For a single page (more than 100,000 windows) and a dataset pool with 6,000 annotated initial letters, this search will take about **32** years to finish.

We know the provenance of those annotated initial letters, such as the date of publication, the print shop, the printer, etc.

The reference dataset pool keeps growing......
Can we do this faster?

Choices

We could:

- **Use better hardware** (parallel processing, cloud etc.)
  - But expensive
- **Index the data**
  - CK1 is non-metric, thus hard to index
- **Do lower bounding search?**
  - Used in many data mining problems…
Can we do this faster?

Can we lower bound of CK1?

CK1 is a \textit{black box} function
No explicit feature extraction, making lower bounding difficult

Trivial to get a \textit{fast} lower bound: \( \text{LB}_{\text{CK1}}(x,y) = \text{zero} \)
Trivial to get a \textit{tight} lower bound: \( \text{LB}_{\text{CK1}}(x,y) = \text{CK1}(x,y) \)

However, we are unable to come up with a \textit{fast} and \textit{tight} lower bound for CK1

*\textbf{We strongly suspect that no one can achieve the lower bound of CK1}*

MPEG-1 encoders are among the most tightly optimized compression methods available

*S. Pigeon, S. Coulombe: \textit{Very Low Cost Algorithms for Predicting the File Size of JPEG Images}. DCC, pp.538, 2008.*
An Approximate Lower Bound

Intuition

black pixel density: "dense" 70.9%  "sparse" 38.15% "text" 18.64%

Goal: To reduce the number of CK1 calculations, since the algorithm is dominated by CK1.
An Approximate Lower Bound

Intuition

Goal: To reduce the number of CK1 calculations, since the algorithm is dominated by CK1.

Black pixel density of P is 44.82%

A page from “Comediae” (1552)
An Approximate Lower Bound

How to learn the lower bound curve for every annotated initial letter

(a)              (b)            (c)                      (d)                    (e)

Goal: To reduce the number of CK1 calculations, since the algorithm is dominated by CK1.
An Approximate Lower Bound

How to use threshold distance for pruning

Four candidate windows

We learn the lower bound curve of the annotated initial letter ahead of time.

Goal: To reduce the number of CK1 calculations, since the algorithm is dominated by CK1.
Algorithm: Lower Bound Algorithm

Input:  
  - \( P \), A page
  - \( IL \), Target annotated initial letter
  - \( t \), A threshold

Output:   exists-flag, \( W_i \)  // Where \( W_i \) encodes location

1. \( W \) = set of all possible windows in \( P \)
2.   for \( i = 1 : |W| \)
3.     if \( \text{LB}(W_i, IL_{\text{lower_bound_info}}) < t \)
4.        current\_dist = \text{CK1\_dist}(W_i, IL); 
5.     endif
6.     endif
7.   endfor
8.   disp('target initial letter not found');
An Approximate Lower Bound

How to use threshold distance for pruning

Goal: reduce the number of CK1 calculations

Use four candidate windows to illustrate how to use threshold for pruning

If \textit{threshold} is too far away, there will be no pruning.
If \textit{threshold} is too close, there will be many false positives.
An Approximate Lower Bound

How to learn the threshold

Requirement for a threshold distance:

If \textit{threshold} is too far away, there will be no pruning.
If \textit{threshold} is too close, there will be many false positives.

\[ t = \text{mean (PairwiseDist)} + 3 \times \text{std} \]

Ten examples of ‘\( \mathcal{L} \)’ from one book published in 1626
This ‘\( \mathcal{L} \)’ appears 109 times in this book.
An Approximate Lower Bound

What if there is only one letter?

Two choices:
1. Take the thresholds learned for common letters
2. Produce slightly distorted versions of them
Outline

• Introduction
• CK1 Distance Measure (Bilson and Dr. Keogh)
• Search Algorithm
• Experimental Evaluation
• Speed Up Techniques
• Conclusion and Future Work
Experimental Evaluation

Empirical evaluation

We learned the lower bound curve for each letter; then calculate the *true* CK1 distance of each letter with 10 million windows.

In these **100 million** CK1 calculations, not a single violation of the lower bound was observed.
Experimental Evaluation

Experiments on historical manuscripts

- Published in 1559
- Published in 1576
- Printed in the same printing house in Paris

98.04% of the candidate windows were pruned by the lower bound
Experimental Evaluation

Experiments on Historical Manuscripts

Published in 1559

Test another book that is similar in age, style, language and use of initial letters

Published in 1572

But No False Positives!
Published in 1533, by Heinrich Petri

The three books were printed by the same person.

in total 6,956 pages
pruning ratio: 99.38%
160 times faster than the brute force search

twenty books from the 15th and 16th century
Experimental Evaluation

A test for robustness

Crumpled ball
Is my algorithm **fast** enough?

• The digitization process itself is not **fast**.
  – *For rare historical manuscripts, it takes perhaps hours to scan one book.*

• *Almost*, and we can make it faster…
  – Downsampling (for unlabeled pages)
  – Indexing the reference set (for unlabeled initial letters)

Speed Up Techniques

**Downsampling**

Here we are ten times faster, with only a slight increase in the error rate.

CK1 Distance Measure

- Experiment on one dataset: 19 classes, 371 images

<table>
<thead>
<tr>
<th>Method</th>
<th>CK1 Distance</th>
<th>CSR</th>
<th>SRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error Rate</td>
<td>0%</td>
<td>39.42%</td>
<td>10.21%</td>
</tr>
</tbody>
</table>

NEED SOME FIGURES HERE!
Speed Up Techniques

Can we make it *faster*? Since there will be

- more and more books online ....
- increasing size of the annotated initial letter dataset pool....

### Annotated initial letter dataset pool

<table>
<thead>
<tr>
<th>DLs</th>
<th>Books</th>
<th>Periods</th>
<th>Web links</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVII</td>
<td>14 000</td>
<td>16th</td>
<td><a href="http://www.bvh.univ-tours.fr">http://www.bvh.univ-tours.fr</a></td>
</tr>
<tr>
<td>Mouriau</td>
<td>1 850</td>
<td>18th</td>
<td><a href="http://www.ornements-typo-mouriau.be/">http://www.ornements-typo-mouriau.be/</a></td>
</tr>
</tbody>
</table>
**Speed Up Techniques**

**Intuition**

Suppose we have an unlabeled initial letter...

*lower bound search of the reference collection*

---

![Graph showing black pixel density distribution](image)

The black pixel density distribution for 6,395 annotated initial letters collection
Speed Up Techniques

Intuition

We can produce *generic versions* of these curves for any image that has $K$ percent black pixel density.

![Graph showing black pixel density vs CK1 distance for LB_Red, LB_Blue, and LB_Green](image)

*LB_Red* and *LB_Blue* and taking their *minimum* values at each location in the X-axis.
The search order is by the **absolute value of the differences** of the query black pixel density and the reference object’s black pixel density.

*top*) When the `best_so_far` is initialized to 1.2, nothing can be pruned.

*middle*) The first item examined reduces the `best_so_far` to 0.82, **pruning 28% of the data**, shown in light gray.

*bottom*) After examining a handful more candidates, a match is found with a distance of 0.37 allowing us to **prune 81% of the data**.
The example in previous page is perhaps the worst case for pruning. This is because that “V” shape is close to the mean of the distribution. Because both the sparse and dense queries are centered far from mean of the distribution, the pruning is much more effective for these queries, assuming the best match is about the same distance away.
Conclusion

• **Introduction**
  – To establish the provenance of historical manuscripts by searching for occurrences of initial letters from a referenced library

• **CK1 Distance Measure**
  – Parameter free, accurate, robust

• **An Approximate Lower Bound**

• **Experimental Evaluation**

• **Speed Up Techniques**
  – Downsampling (for unlabeled pages)
  – Indexing the reference set (for unlabeled initial letters)
Future Work

• It is possible to group together similar elements
  – Currently, each initial letter is searched independently

We can group ‘O’ and ‘Q’, ‘L’ and ‘I’, then conduct a single search for the group representative.