Mining Historical Documents for Near-Duplicate Figures

Thanawin (Art) Rakthanmanon

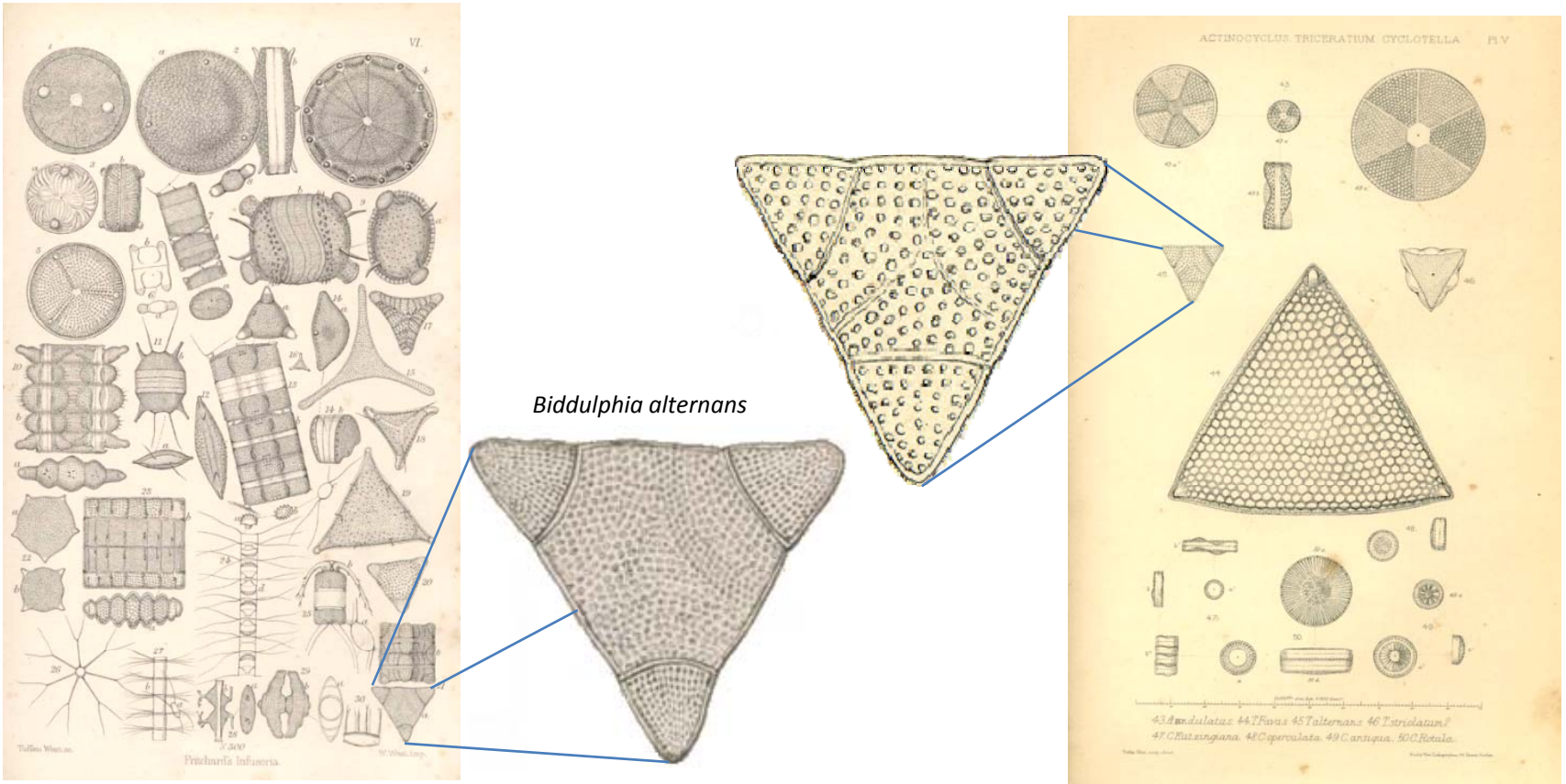
Qiang Zhu

Eamonn Keogh

UNIVERSITY OF CALIFORNIA
UCRIVERSIDE

What is a near-duplicate pattern?

Same Species of diatoms in different books



A History of Infusoria, including Desmidiaceae and Diatomaceae, 1861.

A Synopsis of the British Diatomaceae, 1853.

Motivation

- There are about 130 million books in the world (according to Google 2010).
- Many are now digitized.
- Finding repeated patterns can ..
  - allow us to trace the evolution of cultural ideas
  - allow us to discover plagiarism
  - allow us to combine information from two different sources

Problem Statement

- Given 2 books and user defined parameters (i.e. size of motifs), find similar pattern/figures between these books in reasonable amount of time.

What is a “reasonable amount of time”?

- It can take minutes to hours for scanning a book.
- We would like to be able to discover similar figures in minutes or tens of minutes.
- This could be done offline (a ‘screensaver’ could work on you personal library at night).

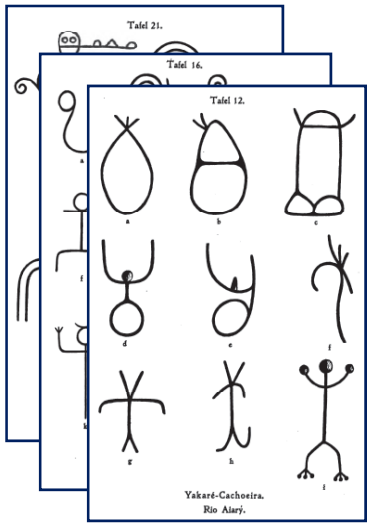
Objectives

- We propose an algorithm to discover similar patterns inside a manuscript or across 2 books.
- Our scalable method consider only *shape* so input documents can be b/w or color documents.
- Our method will return approximately repeated shape patterns in small amount of time.

Example Results (1)

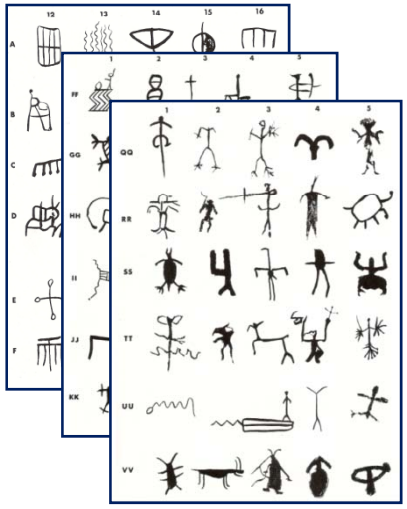
- Two Petroglyph Books

1st book



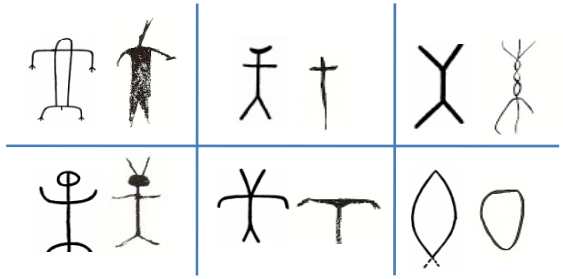
[1]

2nd book



[2]

Similar Figures

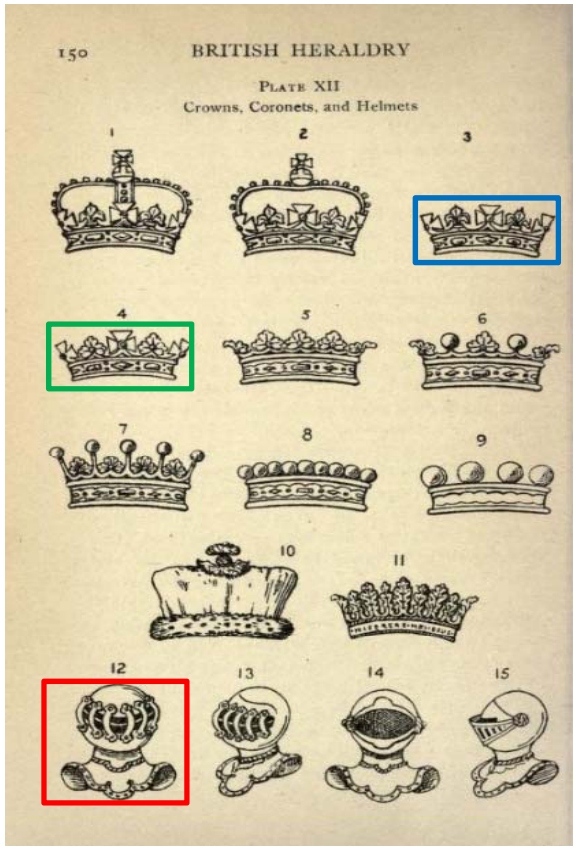


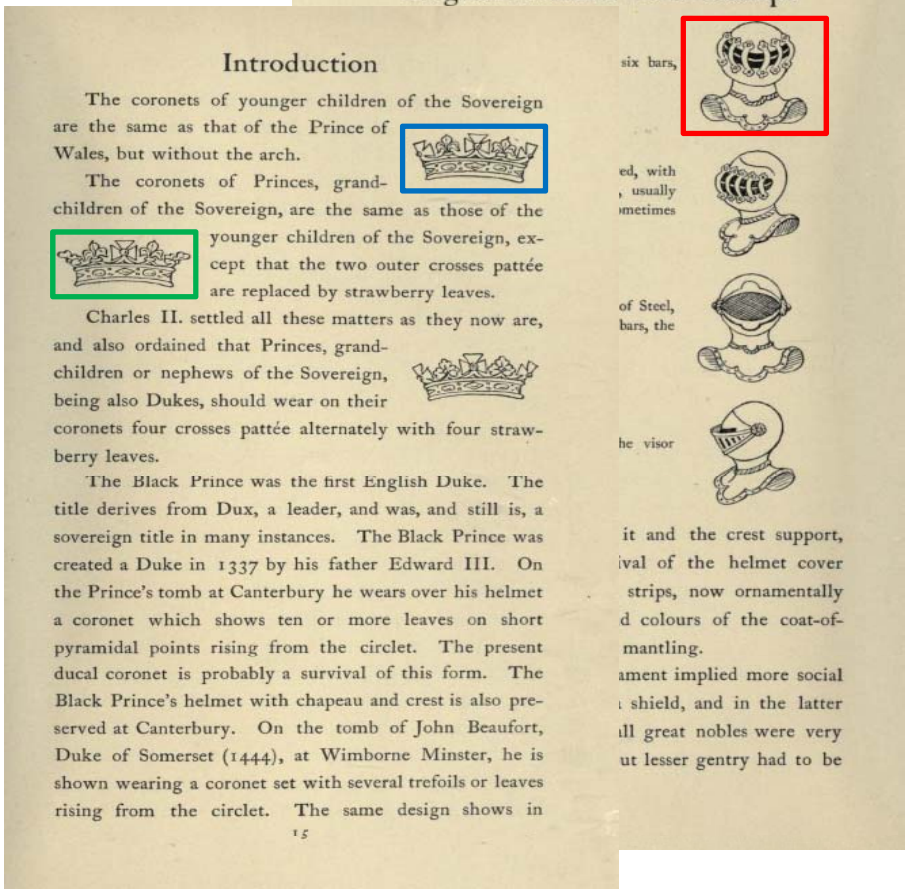
[1] Indian Rock Art of Southern California with Selected Petroglyph Catalog, 1975.

[2] Südamerikanische Felszeichnungen (South American Petroglyphs), Berlin, 1907.

Example Results (2)

- 25 seconds to find motifs across 2 books of size 478 pages and 252 pages.





Example Results (3)

- Similar figures from 4 different books are discovered.

Book1 [3]: Scottish Heraldry (243 pages)

Book2 [4]: Peeps at Heraldry (110 pages)

Book3 [5]: British Heraldry (252 pages)

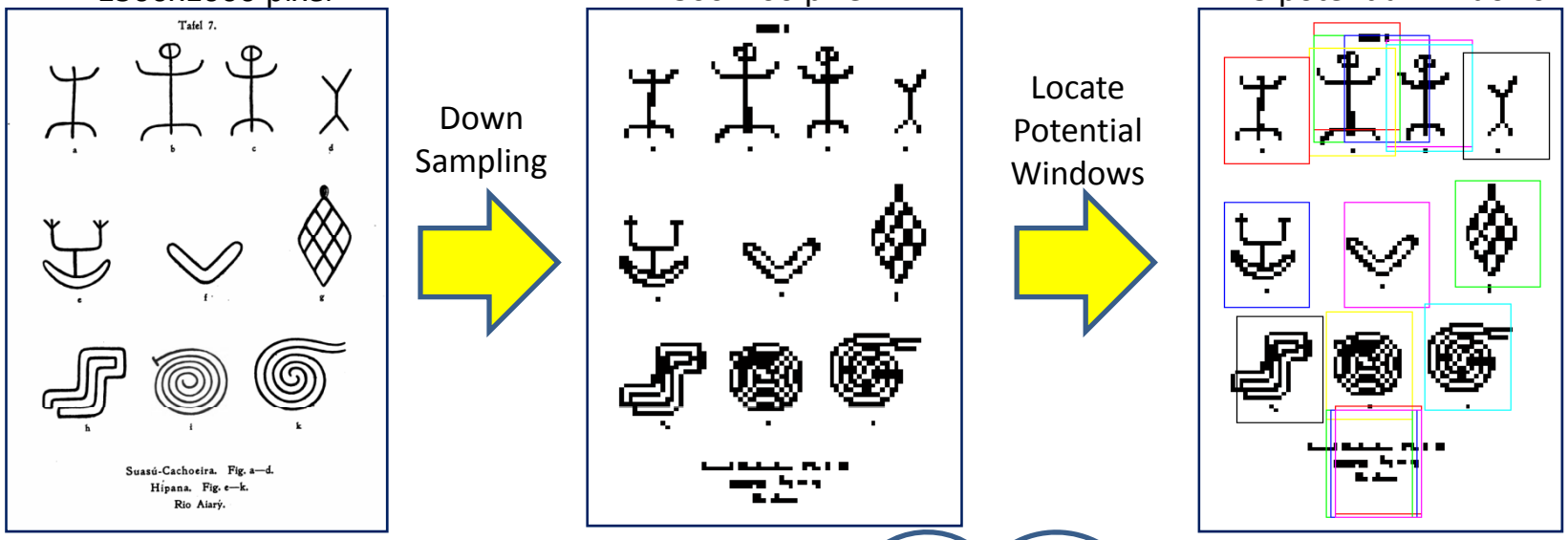
Book4 [6]: English Heraldry (487page)

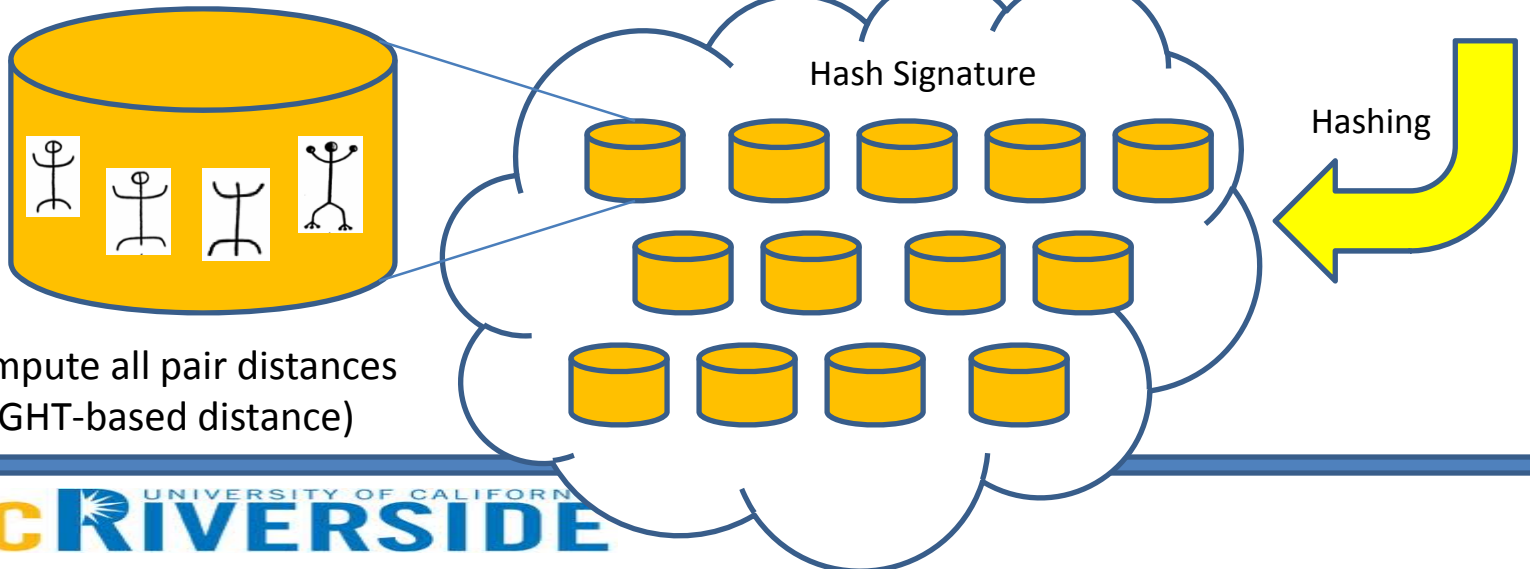
Example Results (4)

- Also works well for handwritten documents.

IAM dataset from Research Group on Computer Vision and Artificial Intelligence, University of Bern

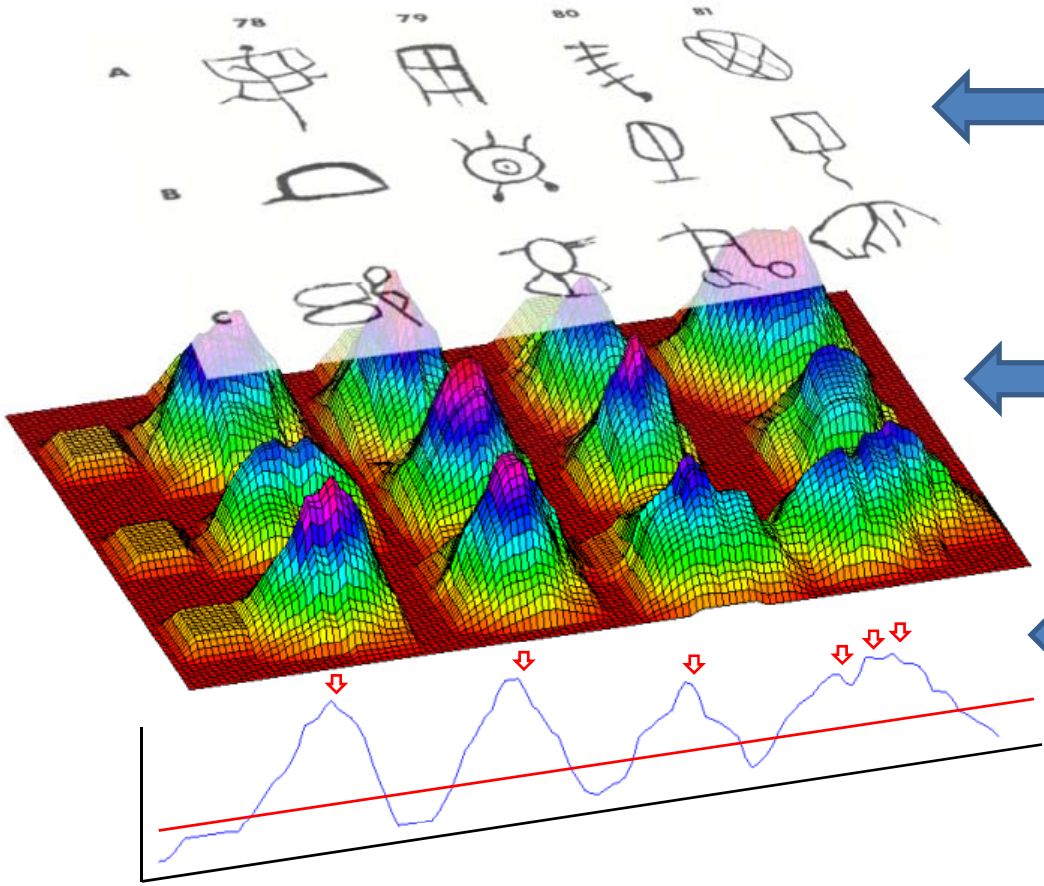
Overview of Our Algorithm





Locating Potential Windows

- Humans easily locate the figures. How?



Our observations ...

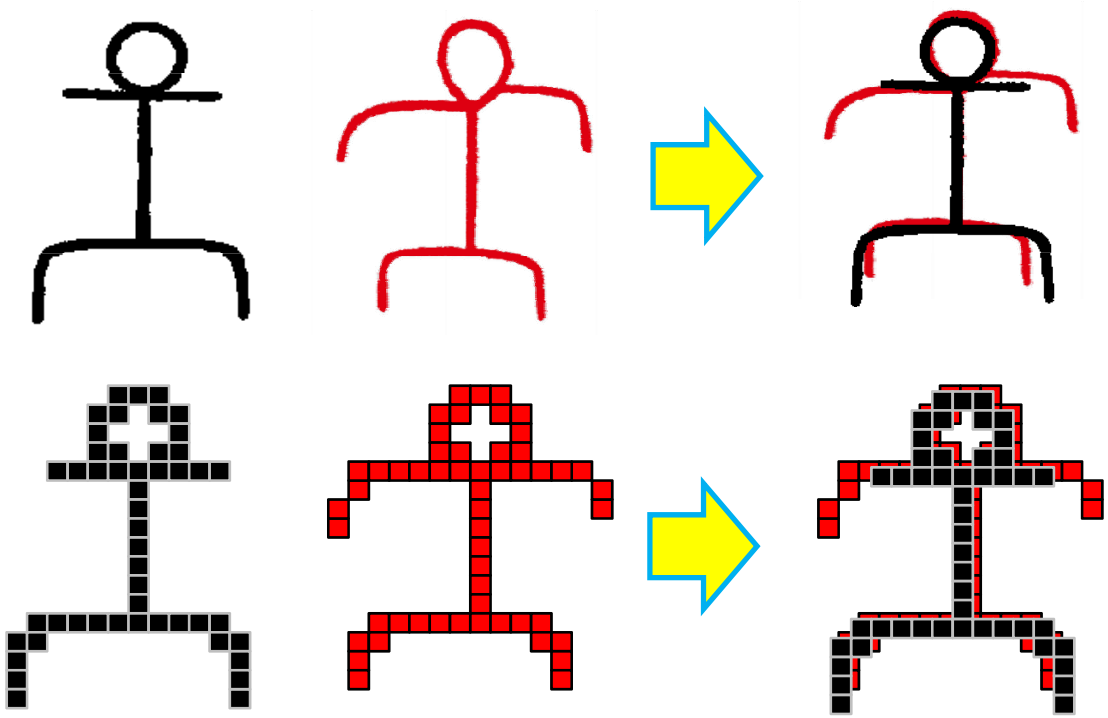
Document contain black and white pixels.

Count the number of black pixels around a fix pixel.

Locate figures by the peaks. Remove noise by threshold.

50,000 windows reduce to 20 potential windows.

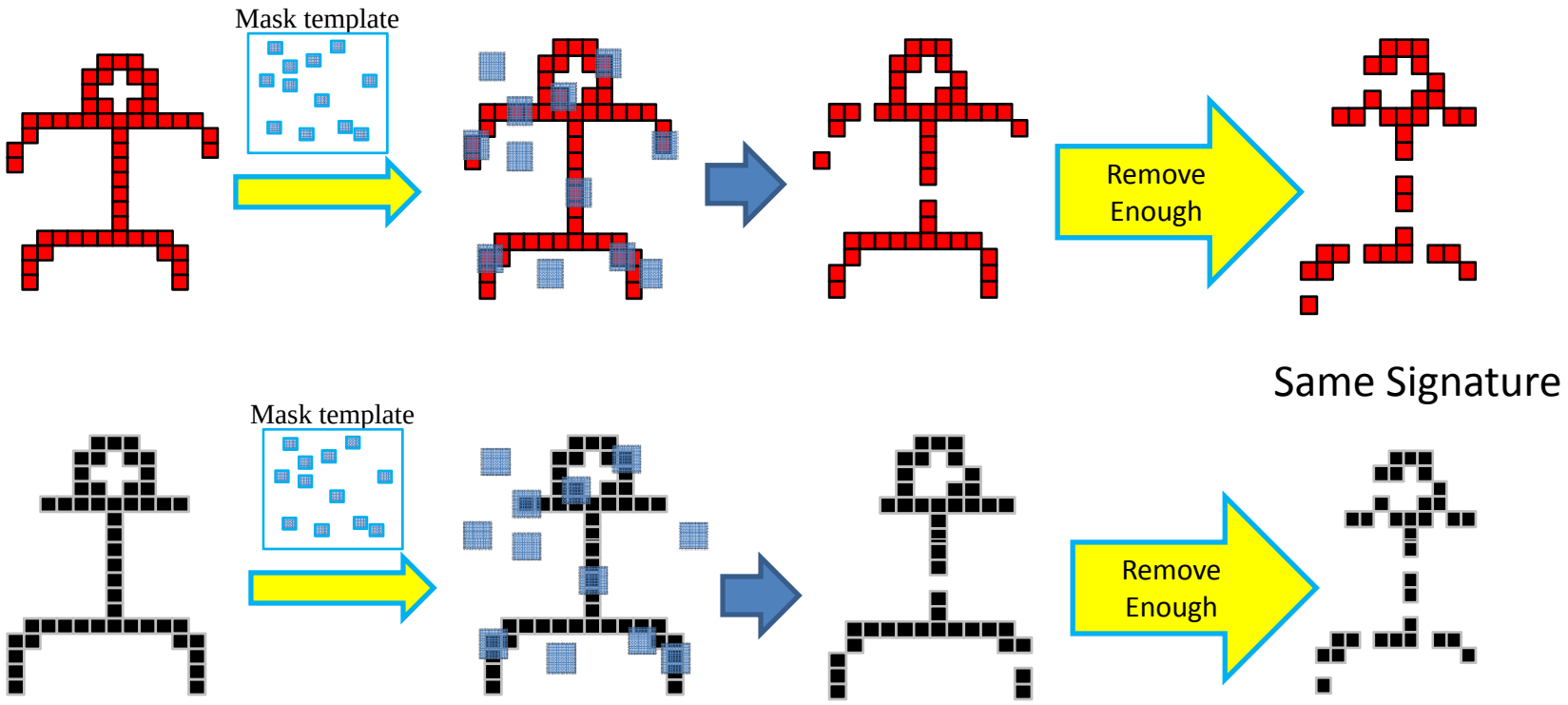
Down Sampling



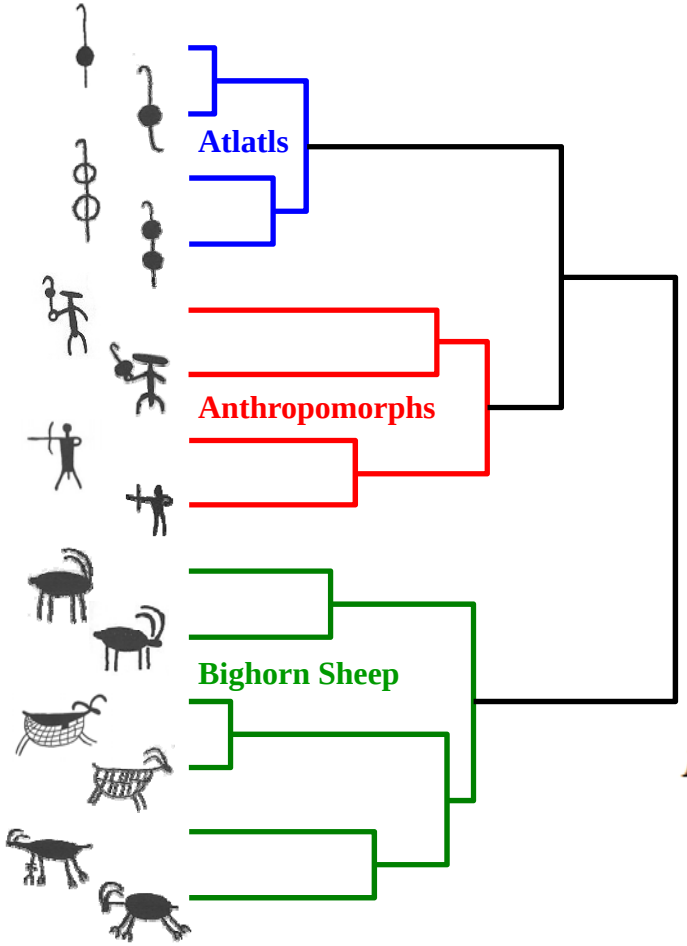
- Reduce search space.
- Increase the quality of matching.

Random Projection

- Hashing is an efficient way to reduce the number of expensive real distance calculations.



GHT-based Distance Calculation



GHT = Generalized Hough Transform

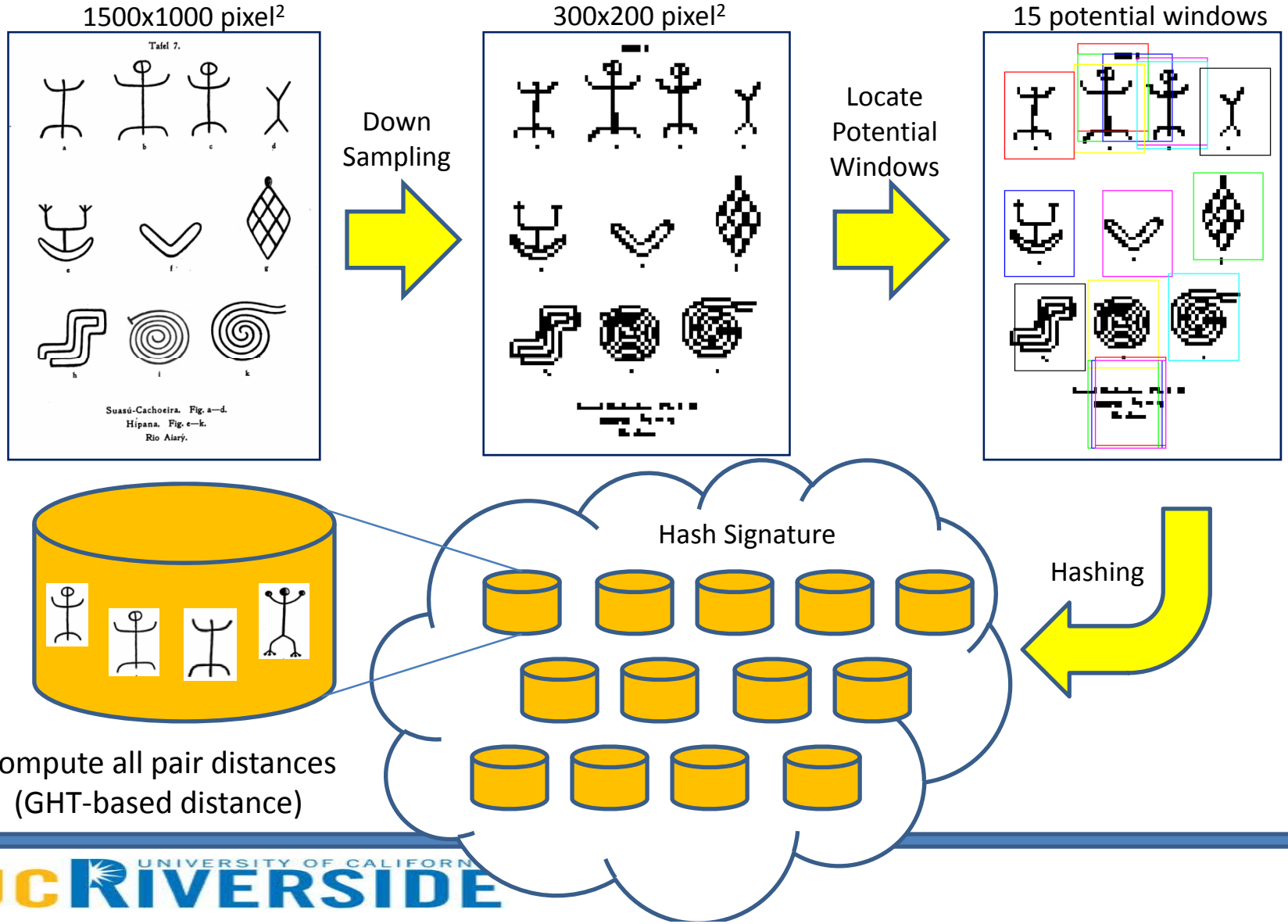
(1) GHT-based distance measure correctly groups all seven pairs.

(2) The higher level structure of the dendrogram also correctly groups similar petroglyphs.

$$Dnn(Q,C) = \begin{cases} \dfrac{1}{N_Q - MUE(Q,C)} \sqrt{N_C / N_Q} & \text{if } N_C > N_Q \\ \dfrac{1}{N_Q - MUE(Q,C)} & \text{otherwise} \end{cases}$$

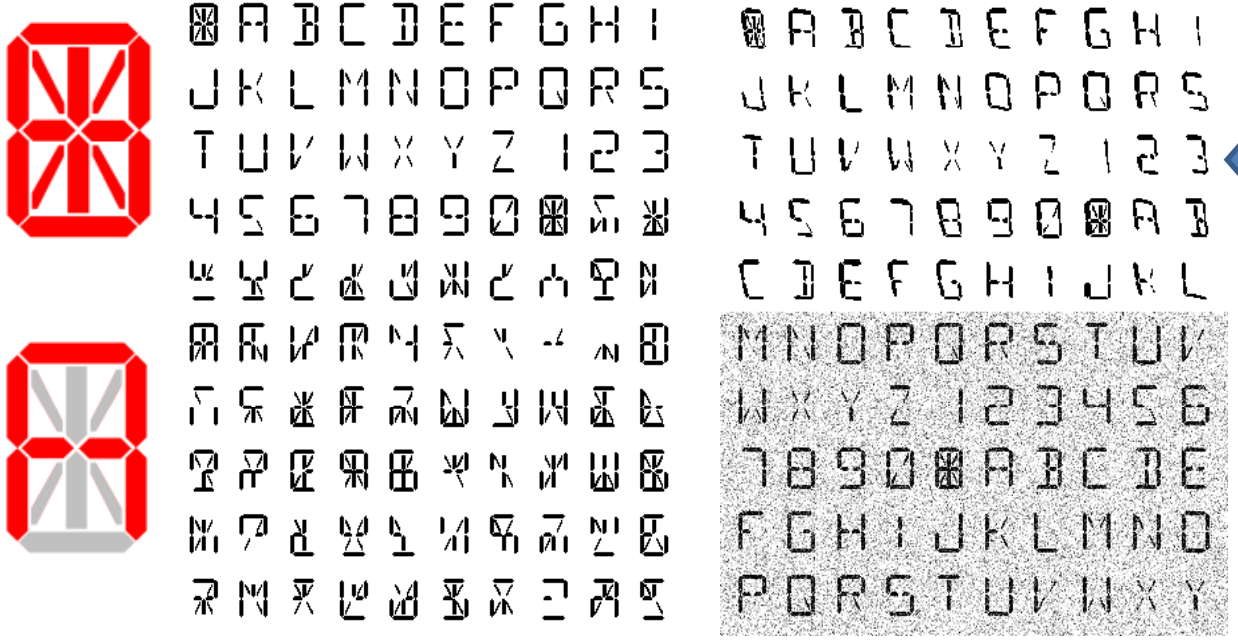
Figure and Equation from [14] Q. Zhu, X. Wang, E. Keogh and S.H. Lee, "Augmenting the Generalized Hough Transform to Enable the Mining of Petroglyphs," SIGKDD, 2009

Overview of Our Algorithm



Experimental Results

- The performance of our algorithm depends on dataset.
- We created artificial “books” to test on.
- Each page of book contains 100 random characters.
- Each characters contains 14 segments.

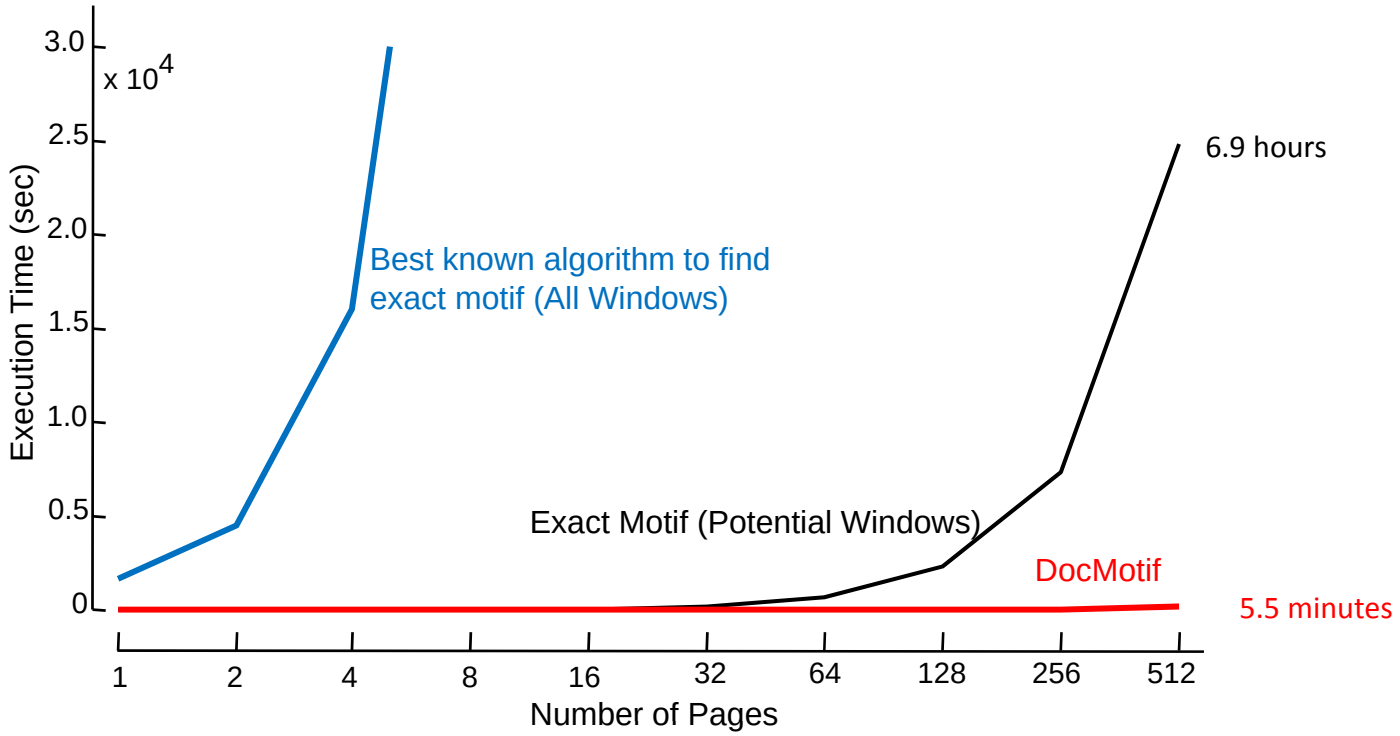


Polynomial Distortion

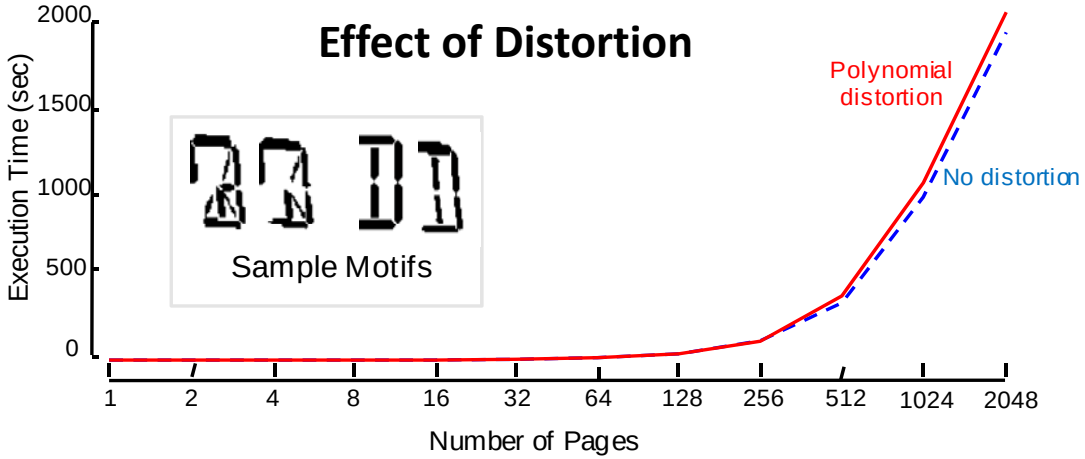
Gaussian Noise

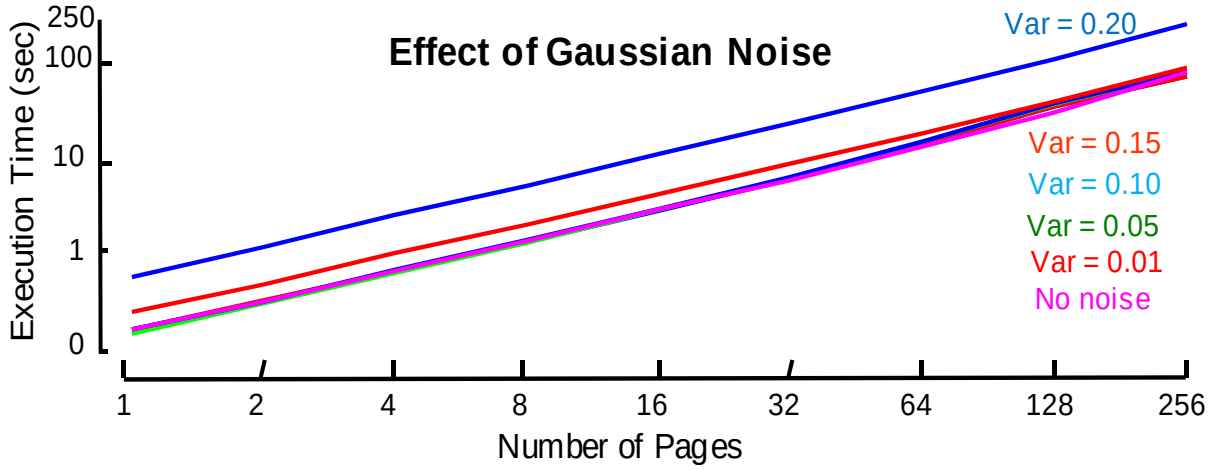
Experimental Results

- Our algorithm can find similar figures (motifs) from 100-page book in less than a minute.



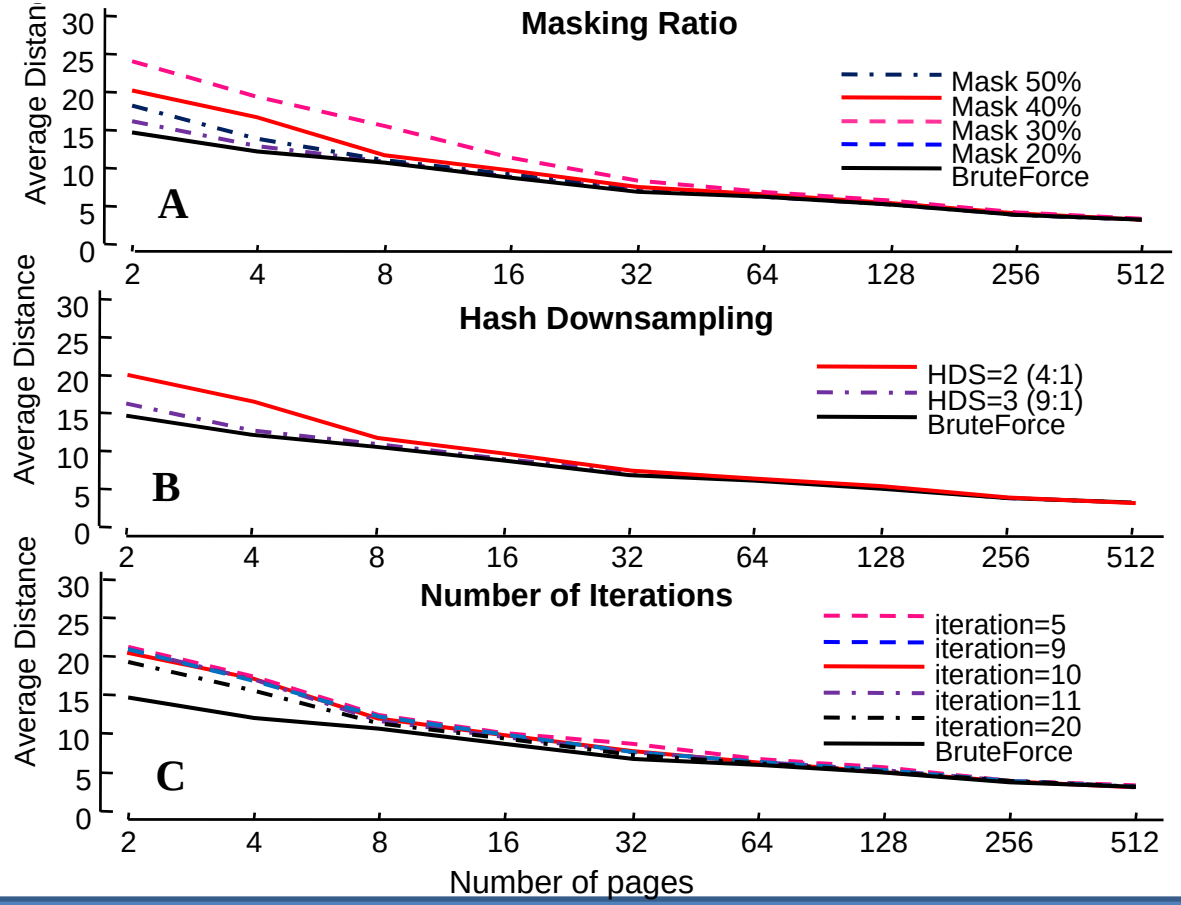
Scalability



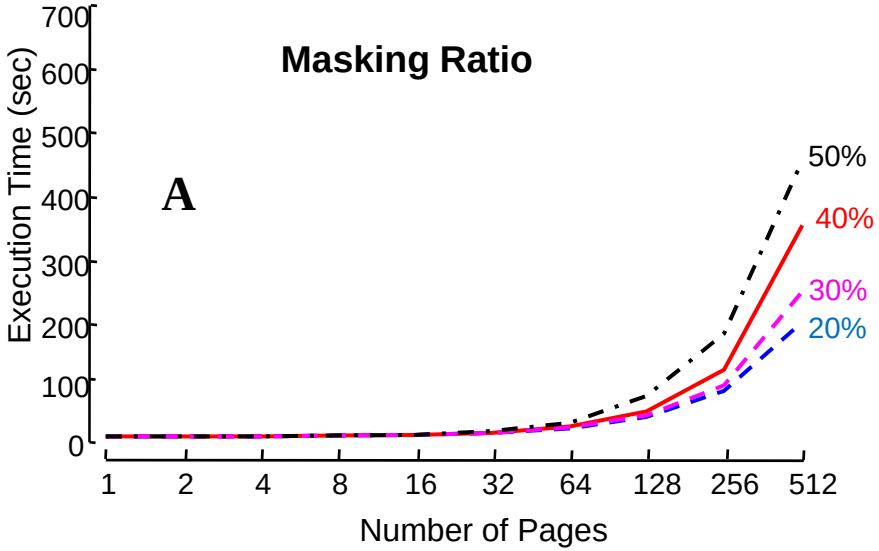


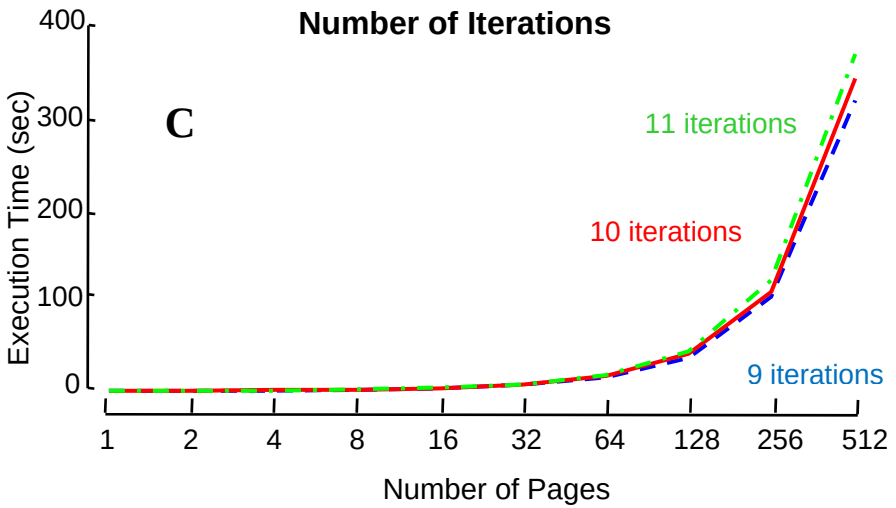
How Good of the Results?

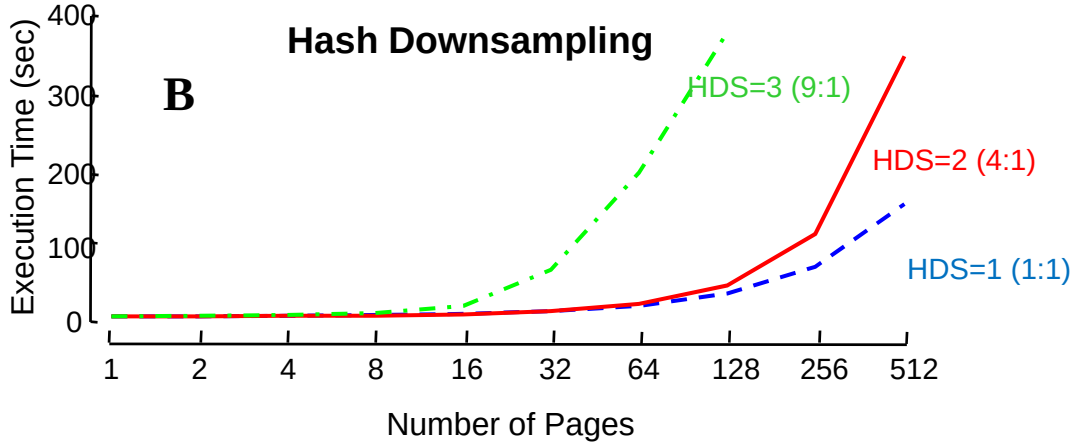
- The average distances from top 20 motifs are not much different among different parameter choices.



Parameter Effects







Conclusion

- An algorithm to find similar figures across two manuscripts.
  - Approximation algorithm
  - Work pretty well on both figures and text
  - Practical: very fast and very similar
- Key Ideas
  - Locating potential windows
  - Down Sampling
  - Random Projection
  - GHT-based Distance
- Drawbacks
  - Not support rotation invariance
  - Many parameters but not much sensitive

References

1. G. Ramponi, F. Stanco, W. D. Russo, S. Pelusi, and P. Mauro, “Digital automated restoration of manuscripts and antique printed books,” EVA - Electronic Imaging and the Visual Arts, 2005.
2. J. V. Richardson Jr., “Bookworms: The Most Common Insect Pests of Paper in Archives, Libraries, and Museums.”
3. A. Pritchard, “A history of Infusoria, including Desmidiaceae and Diatomaceae,” British and foreign. Ed. IV. 968. London, 1861.
4. W. Smith, “A synopsis of the British Diatomaceae; with remarks on their structure, function and distribution; and instructions for collecting and preserving specimens,” vol. 1 pp. [V]-XXXIII, pp. 1-89, 31 pls. London: John van Voorst, 1853.
5. W. West, G S.. West, “A Monograph of the British Desmidiaceae,” Vols. I–V. Ray Society, London, 1904–1922.
6. C. R. Dod, R. P. Dod, “Dod’s Peerage, Baronetage and Knightage of Great Britain and Ireland for 1915,” London: Simpkin, Marshall, Hamilton, Kent and co. ltd, 1915.
7. J. B. Burke, “Book of Orders of Knighthood and Decorations of Honour of all Nations,” London: Hurst and Blackett, pp. 46-47, 1858.
8. B. Gatos, I. Pratikakis, and S. J. Perantonis, “An adaptive binarisation technique for low quality historical documents,” Proc. of Int. Work. on Document Analysis Sys., pp. 102–13.
9. E. Kavallieratou and E. Stamatatos, “Adaptive binarization of historical document images,” Proc. 18th International Conf. of Pattern Recognition, pp. 742–745.
10. H. J. Wolfson and I. Rigoutsos, “Geometric Hashing: An Overview,” IEEE Comp’ Science and Engineering, 4(4), pp. 10-21, 1997.
11. X. Bai, X. Yang, L. J. Latecki, W. Liu, and Z. Tu, “Learning context sensitive shape similarity by graph transduction,” IEEE TPAMI, 2009.
12. E. J. Keogh, L. Wei, X. Xi, M. Vlachos, S. Lee, and P. Protopapas, “Supporting exact indexing of arbitrarily rotated shapes and periodic time series under Euclidean and warping distance measures,” VLDB J. 18(3), 611-630, 2009.
13. P. V. C. Hough, “Method and mean for recognizing complex pattern,” USA patent 3069654, 1966.

References

14. Q. Zhu, X. Wang, E. Keogh, and S. H. Lee, “Augmenting the Generalized Hough Transform to Enable the Mining of Petroglyphs,” SIGKDD, 2009.
15. R. O. Duda and P. E. Hart, “Use of the Hough transform to detect lines and curves in pictures,” Comm. ACM 15(1), pp.11-15, 1972.
16. D. H. Ballard, “Generalizing the Hough transform to detect arbitrary shapes,” Pattern Recognition 13, 1981, pp. 111-122.
17. M. Tompa and J. Buhler, “Finding motifs using random projections,” In proceedings of the 5th Int. Conference on Computational Molecular Biology. pp 67-74, 2001.
18. T. Koch-Grunberg, “Südamerikanische Felszeichnungen” (South American petroglyphs), Berlin, E. Wasmuth A.-G, 1907.
19. A. Fornés, J. Lladós, and G. Sanchez, “Old Handwritten Musical Symbol Classification by a Dynamic Time Warping Based Method. in Graphics Recognition: Recent Advances and New Opportunities,” Lecture Notes in Computer Science, vol. 5046, pp. 51-60, 2008.
20. G. Sanchez, E. Valveny, J. Llados, J. M. Romeu, and N. Lozano, “A platform to extract knowledge from graphic documents. application to an architectural sketch understanding scenario,” Document Analysis Systems VI, Vol. 3163, pp. 389 -400, 2004.
21. J. Mas, G. Sanchez, and J. Llados, “An Incremental Parser to Recognize Diagram Symbols and Gestures represented by Adjacency Grammars,” Graphics Recognition: Ten Year Review. Lecture Notes in Computer Science, vol. 3926, pp. 252-263, 2006.
22. K. B. Schroeder et al., “Haplotypic Background of a Private Allele at High Frequency,” the Americas, Molecular Biology and Evolution, 26 (5), pp. 995-1016, 2009.
23. G. A. Smith, and W. G. Turner, “Indian Rock Art of Southern California with Selected Petroglyph Catalog,” San Bernardino County, Museum Association, 1975.
24. C. Davenport, “British Heraldry,” London Methuen, 1912.
25. C. Davenport, “English heraldic book-stamps, figured and described,” London : Archibald Constable and co. ltd, 1909.
26. X. Xi, E. J. Keogh, L. Wei, and A. Mafra-Neto, “Finding Motifs in a Database of Shapes,” Prof. of Siam Conf. Data Mining, 2007.

Thank you for your attention

QUESTION?

Supplementary