NATIONAL ENDOWMENT FOR THE
Humanities

OFFICE OF DIGITAL HUMANITIES

Narrative Section of a Successful Application

The attached document contains the grant narrative and selected portions of a previously funded grant application. It is not intended to serve as a model, but to give you a sense of how a successful application may be crafted. Every successful application is different, and each applicant is urged to prepare a proposal that reflects its unique project and aspirations. Prospective applicants should consult the Office of Digital Humanities program application guidelines at http://www.neh.gov/grants/odh/digital-humanities-start-grants for instructions. Applicants are also strongly encouraged to consult with the NEH Office of Digital Humanities staff well before a grant deadline.

Note: The attachment only contains the grant narrative and selected portions, not the entire funded application. In addition, certain portions may have been redacted to protect the privacy interests of an individual and/or to protect confidential commercial and financial information and/or to protect copyrighted materials.

Project Title:	Image Analysis for Archival Discovery (Aida)
Institution:	University of Nebraska, Lincoln
Project Directors:	Elizabeth Lorang and Leen-Kiat Soh
Grant Program:	Digital Humanities Start-Up Grants, Level 2

1. Table of Contents

2. List of Participants

Principal Investigators

Lorang, Elizabeth, Research Assistant Professor, Digital Humanities Projects Librarian, Center for Digital Research in the Humanities, University of Nebraska-Lincoln

Soh, Leen-Kiat, Associate Professor, Computer Science and Engineering, University of Nebraska-Lincoln

Advisory Board

Conway, Paul, Associate Professor, School of Information, University of Michigan

Cordell, Ryan, Assistant Professor, Department of English, Northeastern University

Ducey, Mary Ellen, University Archivist, University Libraries, University of Nebraska-Lincoln

Houston, Natalie, Associate Professor, Department of English, University of Houston

McGill, Meredith, Associate Professor, Department of English, Rutgers University

3. Abstract

Images created in the digitization of primary materials contain a wealth of machine-processable information for data mining and large-scale analysis, and this information should be leveraged to connect researchers with the resources they need and to augment interpretation of human culture, as a complement to and extension of text-based approaches. The proposed project, "Image Analysis for Archival Discovery" (Aida), applies image processing and machine learning techniques from computer science to digitized materials to facilitate and promote archival discovery. Beginning with the automatic detection of poetic content in historic newspapers and the creation of a comprehensive catalog of such verse, this project will develop image processing as a methodology for humanities research. In doing so, it will advance work on two fronts: 1) it will contribute to the larger reevaluation of poetry in American literary history; 2) it will pioneer new methods of discovery in archival collections.

Statement of Innovation

Aida draws on established techniques from computer science to explore new questions. It is the first research project to apply image processing and analysis techniques to digitized newspapers to automatically identify poetic content by looking at visual cues (rather than parsing text), and it is unique in broadly conceptualizing image processing as a methodology for archival discovery.

Statement of Humanities Significance

Without corpus-scale identification and analysis of poetic content in U.S. newspapers, the history of poetry in American culture is incomplete and inaccurate. New histories of poetry in the U.S. require new methods of discovery and analysis, which this project provides. Further, image analysis has the potential to deal with a range of forms, genres, and modes, as well as with multi-language corpora, since its techniques do not rely on the recognition or processing of text.

4. Narrative

Enhancing the Humanities Through Innovation

By the most conservative of estimates, several hundred thousand poems appeared in early American and U.S. newspapers from the eighteenth through the early twentieth centuries. One estimate suggests more than 100,000 poems appeared in daily newspapers during the years of the Civil War alone.[1] Counting the snippets of verse that appeared in death notices, advertisements, and articles makes the presence of poetry in historic newspapers even more pervasive. Yet until recently, this rich trove of material has been outside the scope of literary study and merely a footnote in histories of American newspapers. In the last five years, however, scholars have made significant inroads in studying the importance of newspaper verse and the public role of poetry in American culture. Underpinning this scholarship is a growing recognition that the evaluation and history of American poetry should not be based on less than one percent of the poetic record. In addition, this new scholarship values and explores the role of poetry in the daily lives of people, including making sense of what it means to be human and in processing national, social, and individual experiences. To the extent that these new histories depend on traditional methods of archival discovery and analysis, however, they will remain anecdotal—individual narratives extrapolated from a miniscule subset of the whole, with limited means of situating the anecdote as either representative or idiosyncratic. In short, the magnitude of the corpus requires new modes of discovery and analysis.

Proposal Goal. This proposal seeks Level II start-up funding to address the first of these needs, the necessity for new modes of discovery. Specifically, this proposal seeks NEH funding to support the algorithmic processing of nearly 7 million page images from Chronicling America—the portal to all newspapers digitized as part of the National Digital Newspaper Program—in order to identify poetic content in the corpus. Image analysis and data mining of images, rather than of text, has the potential to transform research of digitized newspapers as well as to provide new methods of discovery in digitized archival collections more broadly, including in both print and manuscript collections. Locating relevant materials in digital collections is already often a difficult endeavor and will become increasingly so as more content is digitized and as projects such as the Digital Public Library of America aggregate content from dozens of institutions into a single environment—unless we develop new strategies for connecting researchers to the materials most relevant to them. Image analysis stands to be a crucial addition to both the archivist's and the researcher's toolkits, for its ability to reduce noise in digitized collections and connect users with the materials they need.

Problem Statement and Motivations. A fundamental problem in the reappraisal of newspaper verse remains finding and processing poetic content in newspapers in an efficient manner, which is essential for developing new interpretations, analyses, and literary histories. The primary means of finding this content involves paging through original issues of newspapers, scrolling through reels of microfilm, and browsing digital images to visually scan, by human eye, each page for graphical features that resemble poetry. Dealing with only daily newspapers for a single year, 1860, would require visually scanning nearly half a million newspaper pages.[2] This figure does not take into account the typically

[1] Elizabeth Lorang, "American Poetry and the Daily Newspaper from the Rise of the Penny Press to the New Journalism," Ph.D. dissertation, University of Nebraska-Lincoln, 2010, 123.

[2] In his foundational history of American journalism, Frank Luther Mott estimated the number of U.S. newspapers in existence in 1860 at 3,000, 11 percent of which were dailies (*American Journalism: A History of Newspapers in the United States through 250 Years, 1690 to 1840* [New York: Macmillan, 1942], 216). Most dailies in this period were four pages long.

lengthier weekly newspapers. Certainly no individual in a lifetime could complete a count—to say nothing of a comprehensive bibliography or macro-level analysis—of newspaper verse using this strategy.

While the digitization of historic newspapers has mitigated some issues of access, the main avenues for discovery in these collections are browsing and text-based searching. Browsing for poetic content in such digitized collections follows the strategy outlined above: going image by image through digitized pages and visually scanning the images for the features typical of printed poetry in newspapers. Ironically, web interfaces and variations in Internet connection speeds can make digitally paging through a newspaper a slower process than either scrolling through microfilm or flipping physical pages. Further, the text-based searching enabled by web interfaces is of little utility in the identification of poetic content. For all the power of searchable metadata and full-text transcriptions and the significant ways they have opened up archival material materials, they remain limited to certain kinds of users' needs.

In particular, searching—whether of metadata or full text—presupposes knowing a string to search, which can work only when one is looking for a specific word or phrase and that word or phrase appears in the transcription or accompanying metadata. In the case of poetic content, there is no set of standard words or phrases one could search in order to begin finding this material. Neither the text of the newspaper verse nor its paratextual elements—such as headlines supplied by editors—routinely identify poems or other poetic content. For every poem that appears under a heading such as "Original Poetry" or "Select Poetry," another piece does not. Further, digitization has exacerbated the problems of item-level description already manifest in finding aids to archival collections. The problem is more severe with newspapers and other periodicals, where the heterogeneous and composite nature of the texts means that even item-level description is often not a granular enough form of information.

How, then, might one discover poetic content in digitized historic newspapers? One avenue of approach is to use the electronic text of newspaper issues when available, but to move beyond the searching enabled by web interfaces. With access to the digitized text, one might, for example, use machine learning to train the computer about the linguistic features of newspaper verse, coupled with training about linguistic features of other newspaper content. What are the features of poetic content versus the features of non-poetic content? Following this training, one might process the fully transcribed newspaper texts to find items that share linguistic features of the items known to be poetic content. Elements of poetic diction, including vocabulary (the presence of words such as "o'er," for example) and syntax, could be used to find other pieces likely to be poetry.

A fundamental barrier to this approach, however, is the state of the full-text transcriptions of digitized newspapers. The default method for generating transcriptions of historic newspapers is optical character recognition (OCR). Optical character recognition is the electronic conversion of digital images to text. Despite advances in OCR algorithms for transcription and correction in recent years, newspapers are notoriously difficult materials on which to perform OCR.[3] Perhaps in recognition of this fact, the Library of Congress does not call the electronically generated text of digitized newspapers in Chronicling America transcriptions at all, but identifies them as "OCR interpretations."

Proposed Approach. We believe the *page images themselves rather than machine-translated text hold the most promise for scholarly inquiry with regard to poetic content*. No project has yet applied image processing techniques to this material. The *basis* for our approach is that the appearance of poetic content usually follows certain patterns that can be visually differentiated from other published texts in newspapers. Given a newspaper page, a person can survey or scan the page and figure out quite quickly

[3] See Appendix A for an example of OCR-derived text of historic newspapers.

whether the page contains a poem to a certain degree of accuracy, without having to read or understand the text. The visual cues used in this case include the larger margin within the column where the poetic content resides, the regular patterns of white spacing between stanzas, and so forth.[4] Our project will have the computer do the same visual processing as the human eye and brain when a person moves page-by-page through a newspaper issue looking for poetic content. This image processing approach can also be used as a powerful filter, removing materials from further consideration that do not meet the specified criteria. That is, not only will the process work to identify pages that appear to include poetry, but it will discard those that do not, weeding out much of the noise.

The methodology of the current project has immediate application to a variety of other research questions and broader issues. Significantly, developing image processing as a methodology for humanities research and discovery means we can deal with multi-language corpora, since image analysis need not understand the text of the materials. For poetic content, the approach should be able to find verse in Spanish, German, Cherokee, or Yiddish, for example, just as easily as in English.[5] Within newspapers, image processing and analysis may be used to identify obituaries, birth and marriage announcements, tabular information such as stock reports and sports statistics, and advertisements, among other visually distinct items. Such image processing also has the potential to augment formal studies of early American and U.S. newspapers. Existing formal studies of American newspapers generalize about changes in format, typography, and text-density, as well as about where certain types of materials were likely to appear in a newspaper over time—such as the movement of advertisements from the first page to inner pages of newspapers and the frequency of obituaries and birth and marriage announcements. In facilitating the study of millions of pages, image processing will help to advance—whether through support or through complication—our understanding of the formal aspects of newspapers. These formal qualities of newspapers are typically wrapped up in larger issues relevant to media studies, social and cultural history, and literary studies, among other fields.

Further Rationales and Broader Impacts. The value of image processing and analysis as a tool for humanities research also extends beyond newspapers. The technique can be used both for the identification of poetic content in different types of materials as well as for the identification of other textual and visual forms in digitized collections. For example, this same approach might be used to identify poetic content in manuscript materials such as correspondence, diaries, and scrapbooks. Item-level description is simply not feasible in most collections, and where items are individually described, they are not likely to mention, for example, the presence of lines of poetry in a piece of correspondence. Locating evidence of the readership, use, and circulation of poetry in such archival materials at present is thus tantamount to finding the proverbial needle in the haystack. And yet, such evidence of readership and circulation is a vital part of the story of the lived experience of poetry in American culture. The ability to identify items that include poetic content in archival collections radically changes the research landscape. In addition, the basic methodology can be extended to the identification of other visually distinctive forms in archival collections, such as music (lyrics or musical notation), pictures (whether sketches, photographs, or other illustrations), and maps. Image analysis might conceivably become part of an archive's processing of materials for online presentation, as a way to generate additional metadata, *and* a researcher's tool for more fine-grained processing and analysis.

Text-based browsing, searching, data mining, and other textual analysis of digitized materials

[4] See Appendix B for examples of the visual distinctness of poetic content in newspapers.
[5] As of September 4, 2013, Chronicling America includes digitized newspapers published in English, French, German, Hawaiian, Japanese, and Spanish.

have had profound effects on information access and retrieval, our engagement with information architectures and metadata, and our ability to study certain phenomena at corpus scale. There has been significantly less exploration into corpus-scale analysis using the millions of digital images we are creating as we digitize historic materials. Since these images often have more informational value than the limited metadata associated with them, humanities scholars and library and information professionals must expand our methods of analysis to digital images. Text-based searching, analysis, and data mining will remain powerful and appropriate methodologies for many types of research projects. If we focus exclusively on text, however, we are not leveraging the full power of digitization. As the amount of digitized materials continues to expand, we need new ways of connecting researchers—whether scholars, students, or the general public—with the materials they need.

Environmental Scan

This project draws on and extends recent scholarship in American literary studies and converges in productive ways with innovative research in digital humanities. In particular, this project is informed by the literary scholarship of Meredith McGill, author of *American Literature and the Culture of Reprinting* (2003), *The Traffic in Poems* (2008), and a new book on the circulation of poetry in the antebellum U.S.; Ellen Gruber Garvey, whose 2013 monograph *Writing with Scissors* powerfully demonstrates the importance of poetry, including newspaper verse, in the daily lives of average men and women; Faith Barrett, author of *To Fight Aloud is Very Brave: American Poetry in the Civil War* (2012); Joan Shelley Rubin, author of *Songs of Ourselves: The Uses of Poetry in America* (2007); and Mike Chasar, author of *Everyday Reading: Poetry and Popular Culture in Modern America* (2012). In addition, "Image Analysis for Archival Discovery" dovetails with Ryan Cordell's NEH-funded project, "Uncovering Reprinting Networks in Nineteenth-Century Newspapers" and complements other text-based research on historic newspapers such as Robert K. Nelson's *Mining the* Dispatch and *Mapping Texts*, led by researchers at Stanford University and the University of North Texas. [6]

Image processing and image analysis are proven techniques in computer science, and there is a significant body of literature on techniques such as blurring, dynamic thresholding, and machine learning (see Appendix C for more information on these processes). Various digital humanities projects are turning to these and similar techniques to begin answering a variety of questions and to provide access to materials in new ways. The Bodleian Ballads project, for example, uses image matching technology to find prints made from the same woodblock in the Bodleian Broadsides collection. The Software Studies Initiatives' ImagePlot creates visualizations of image and video collections to explore the way image features cluster or diverge. And the NEH-funded "The Visual Page," led by Natalie Houston, is currently studying the visual features of books of Victorian poetry. In this case, however, the project team began with books already identified as books of poetry.

Ryan Cordell, Natalie Houston, and Meredith McGill, whose work informs this project, have all agreed to serve on the advisory board.

[6] See http://dsl.richmond.edu/dispatch/pages/home and http://mappingtexts.org/index.html. Note that Nelson's topic modeling work, which includes a topic he identifies as "poetry and patriotism," uses the text of newspaper issues that have been hand transcribed. For more on the digitization of the *Dispatch*, see Elizabeth Lorang and Brian Pytlik Zillig, "Electronic Text Analysis and Nineteenth-Century Newspapers: TokenX and the *Richmond Daily Dispatch*," *Texas Studies in Literature and Language* 54.3 (2012): 303–323.

History and Duration of the Project

This project has its origins in Elizabeth Lorang's dissertation, "American Poetry and the Daily Newspaper from the Rise of the Penny Press to the New Journalism" (2010). A database of poetic content cataloged by Lorang for her dissertation and subsequent research is available at http://tinyurl.com/lizlorang. Lorang approached Leen-Kiat Soh in the Spring 2013 to enlist his expertise in image processing (which includes research on processing of satellite imagery) to improve the discovery of poetic content in newspapers.

During the 2013-2014 academic year, Lorang and Soh are working with two undergraduate students through the University of Nebraska-Lincoln's Undergraduate Creative Activities & Research Experiences (UCARE) program. The students are preparing image snippets for training sets, beginning the process of developing algorithms for describing image characteristics, training classifiers, and analyzing very preliminary results. The students will perform their own research in image processing, historic newspapers, and nineteenth-century poetry. UCARE provides a stipend total of $2,400 to students.

Following the start-up phase, Lorang and Soh will continue to develop the project based on feedback from the advisory board and in consultation with other humanities scholars and information professionals. Lorang and Soh will pursue internal and external support for future versions of the project, including departmental funding, internal grants, private grants, and other federal funding.

Work Plan

We request Level II Start-Up funds to support an 18-month project (June 2014-November 2015). The majority of the project work will take place in summer 2014 and summer 2015. A detailed description of the steps to be completed in the image processing work is included in Appendix C.

Summer 2014

- Prepare additional training set images.
- Process initial data sets to extract/derive features from image data.
- Develop algorithms for describing the image characteristics.
- Train classifier to recognize poetic content.
- Analyze preliminary results and revise algorithms to achieve higher accuracy rates.
- Write program to interact with Chronicling America application programming interface.
- Process subset of Chronicling America images.
- Review results and revise algorithms.
- Present short paper and/or poster session on preliminary work and results at Digital Humanities 2014 conference (pending submission and peer review; PIs will propose in Fall 2013).

Academic Year 2014-2015

- Present preliminary results to advisory board members and solicit input for future development.
- Seek input from humanities specialists and librarians, archivists, and information professionals outside the advisory board. Consult via email, web conferencing, and at conferences.
- Identify internal and external funding possibilities for future project development.
- Form at least one external partnership with an institution willing to test methods on digitized archival materials.

Summer 2015

- Process all images from Chronicling America (by this time likely to be more than 7 million images), breaking this work into subsets as necessary to continue to modify algorithms.
- Evaluate and analyze results of image processing.

- Communicate results with advisory board.
- Begin research work to tie image zone information for poetic content to OCR text and issue, page, and item-level metadata.
- Process digital image content aggregated by the Digital Public Library of America (DPLA), accessing images via the DPLA API, if other project work concludes ahead of schedule.
- Prepare white paper for submission to NEH.
- Prepare and submit at least one publication to an appropriate peer-reviewed journal.

This work will build on existing algorithms developed for image processing of other types of materials. We have developed an initial prototype in Java and thus will continue to develop our code in that programming language. Further, we will make use of the open source data mining and machine learning programs (also Java-based) available online on the Weka website (http://www.cs.waikato.ac.nz/ml/weka/), a popular and highly regarded repository by computer scientists especially in the area of Artificial Intelligence. We will also make use of an interactive image manipulation tool called XV for the X Window system to perform some rudimentary tasks such as displaying and format conversions.

Staff

This project is jointly led by Dr. Elizabeth Lorang, Digital Humanities Projects Librarian in the Center for Digital Research in the Humanities at the University of Nebraska-Lincoln (UNL), and Dr. Leen-Kiat Soh, Associate Professor of Computer Science and Engineering at UNL. Lorang will commit a minimum of two months' work to the project over the 18-month period, and Soh will commit a minimum of one month's work to the project over the 18-month period. Lorang and Soh will be joined by two undergraduate research assistants, one working in the humanities and one in computer science.

The project has an advisory board comprised of humanities faculty and information professionals with expertise and experience in American literature and nineteenth-century poetry; digital humanities; humanities analysis at corpus scale; image and text quality in large-scale digitization programs; and the challenges and opportunities digital information creates for libraries and archives. The advisory board will serve in a consulting role. The board will meet virtually as needed over the course of the project and will be especially central to shaping future areas of development.

Final Product and Dissemination

This work plan will result in a program for retrieving image content from Chronicling America via the site API; source code for processing newspaper images to identify content based on feature recognition; and a white paper for NEH assessing this methodology for historic newspaper research and research in digitized collections more broadly. We will also create a catalog of all pages from Chronicling America that include poetic content. To the extent that we are able, we will provide even more granular information, such as tying particular zones of an image identified as poetic content to the OCRed text. Project team members will pursue appropriate conference and publication venues for disseminating this work, as well as work with university communications to prepare a press release at the time a grant is awarded and at the project's completion.

All source code created for the project will be made freely available through a GitHub repository. Research reports, the white paper, and other publications resulting from this work will be made freely available via the UNL's institutional repository, Digital Commons, and project image data and metadata will be freely available via UNL's Data Repository. See the Data Management Plan for more information.

6. Biographies

Principal Investigators

Elizabeth Lorang is Digital Humanities Projects Librarian and Research Assistant Professor in the Center for Digital Research in the Humanities at the University of Nebraska-Lincoln. She holds a Ph.D. in English and is completing a graduate degree in library science. She co-directs *Civil War Washington* (civilwardc.org) and serves as senior associate editor of *The Walt Whitman Archive* (whitmanarchive.org). With R. J. Weir, she recently completed an electronic scholarly edition of poems published in two newspapers during the Civil War, "'Will not these days be by thy poets sung': Poems of the *Anglo-African* and *National Anti-Slavery Standard*, 1863-1864," which is available at scholarlyediting.org. Her work has appeared or is forthcoming in *Literature and Journalism: Inspirations, Intersections, and Inventions from Ben Franklin to Stephen Colbert* (Palgrave 2013); *Scholarly Editing*; *Texas Studies in Literature and Language*; *Victorian Periodicals Review*; *Walt Whitman Quarterly Review*; the *New York Times'* Civil War blog, *Disunion*; and at civilwardc.org and whitmanarchive.org. More information, including a complete cv, is available at elizabethlorang.com.

Leen-Kiat Soh, Associate Professor of Computer Science and Engineering at the University of Nebraska-Lincoln, has conducted research in multiagent systems, intelligent systems, and image processing, with main applications in computer-aided education systems, multiagent simulations, and intelligent assistive systems. In image processing, Soh has built several tools, including ARKTOS, which classifies Arctic sea ice types based on analyzing satellite imagery. He has also applied data mining and machine learning techniques to imagery, geographical, and textual datasets. Soh has published 150 journal and conference papers on his research and applications. His research has primarily been supported with National Science Foundation funding. He is a member of AAAI, ACM, and IEEE. Soh received his Ph.D. in Electrical Engineering with Honors from the University of Kansas.

Advisory Board

"Image Analysis for Archival Discovery" has an advisory board comprised of humanities faculty and information professionals with expertise and experience in American literature and nineteenth-century poetry; digital humanities; humanities analysis at corpus scale; image and text quality in large-scale digitization programs; and the challenges and opportunities digital information creates for libraries and archives. The advisory board will serve in a consulting role. The board will meet virtually as needed over the course of the project and will be especially central to shaping future areas of development.

Conway, Paul, Associate Professor, School of Information, University of Michigan
Cordell, Ryan, Assistant Professor, Department of English, Northeastern University
Ducey, Mary Ellen, University Archivist, University Libraries, University of Nebraska-Lincoln
Houston, Natalie, Associate Professor, Department of English, University of Houston
McGill, Meredith, Associate Professor, Department of English, Rutgers University

7. Data Management Plan

Data to be Generated

Type of Data	When Shared?	Under What Condition?
Open source computer code associated with tool, including algorithms and interface.	At conclusion of the start-up project, when initial testing has been completed.	Code will be freely available on GitHub.
Processed imagery datasets (original input, intermediate results, and final output) and metadata available as test data.	At conclusion of the start-up project, when initial testing has been completed.	Data will be made freely available from the project website and will be deposited and maintained by UNL's Data Repository.
Multimedia progress reports.	At the time of their writing, throughout the duration of the project.	The progress reports will be freely available on the project website and via UNL's institutional repository.
White paper.	After the project has been completed.	The white paper will be freely available on the project website and via UNL's institutional repository.
Final report to NEH.	At the conclusion of the project.	Dissemination of the final report will be the responsibility of the NEH; it will also be made available on the project website and via UNL's institutional repository.

* The proposed project website will be hosted by the Center for Digital Research in the Humanities (CDRH) at http://cdrh.unl.edu.

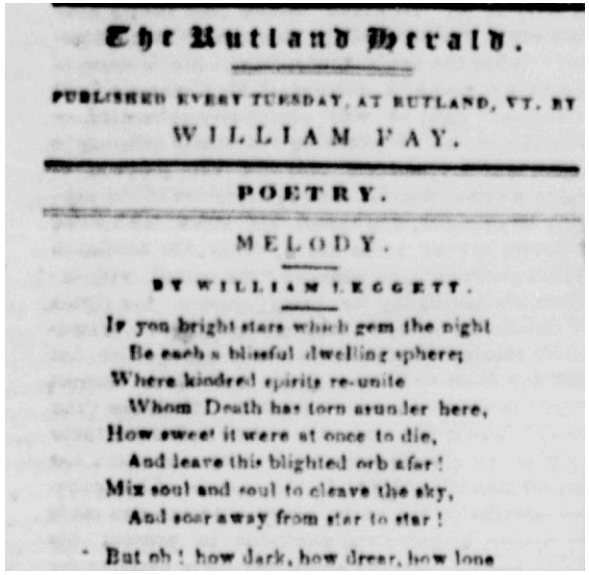
Period of Data Retention
Data and formal reports will be publically available within 1 year of project completion. Data will be retained for a minimum of 5 years beyond the completion of the start-up phase.

Data Formats and Dissemination
Computer code (Java) will be freely available via GitHub, a publically accessible code repository. Our imagery datasets (in PGM and JPEG formats) and corresponding metadata will be deposited and maintained in the University of Nebraska-Lincoln's Data Repository (https://dataregistry.unl.edu/). Reports will be made available in PDF format and deposited in the University of Nebraska-Lincoln's institutional repository, DigitalCommons@UNL (http://digitalcommons.unl.edu/). The project website will serve as a portal to all project data.

Data Management and Maintenance
All computer code will be stored and made available in GitHub, where it will be maintained by project team members. Image data and metadata will be deposited and stored in UNL's Data Repository. The Data Repository is managed by the University of Nebraska-Lincoln Libraries and Information Services and is "a secure site for storage of data collections that are no longer actively in use [allowing] the researcher to stably retain data for future use and/or sharing with other interested parties." The University Libraries will be responsible for long-term data management and maintenance. All reports and publications will be deposited in the University of

Nebraska-Lincoln's Institutional Repository, which is administered by the University Libraries. The institutional repository is hosted by bepress, which has "a multi-tiered disaster recovery plan utilizing fail-over servers and regular on-site and off-site backups" and supports LOCKSS, an Open Archival Information System-compliant preservation solution. The University Libraries will be responsible for the long-term management and maintenance of these assets.

Appendix A: OCR Interpretation in Chronicling America



The Rutland Herald.

PUBLISHED EVERY TUESDAY, AT RUTLAND, VT. BY

WILLIAM FAY.

POETRY.

MELODY.

BY WILLIAM LEGGETT.

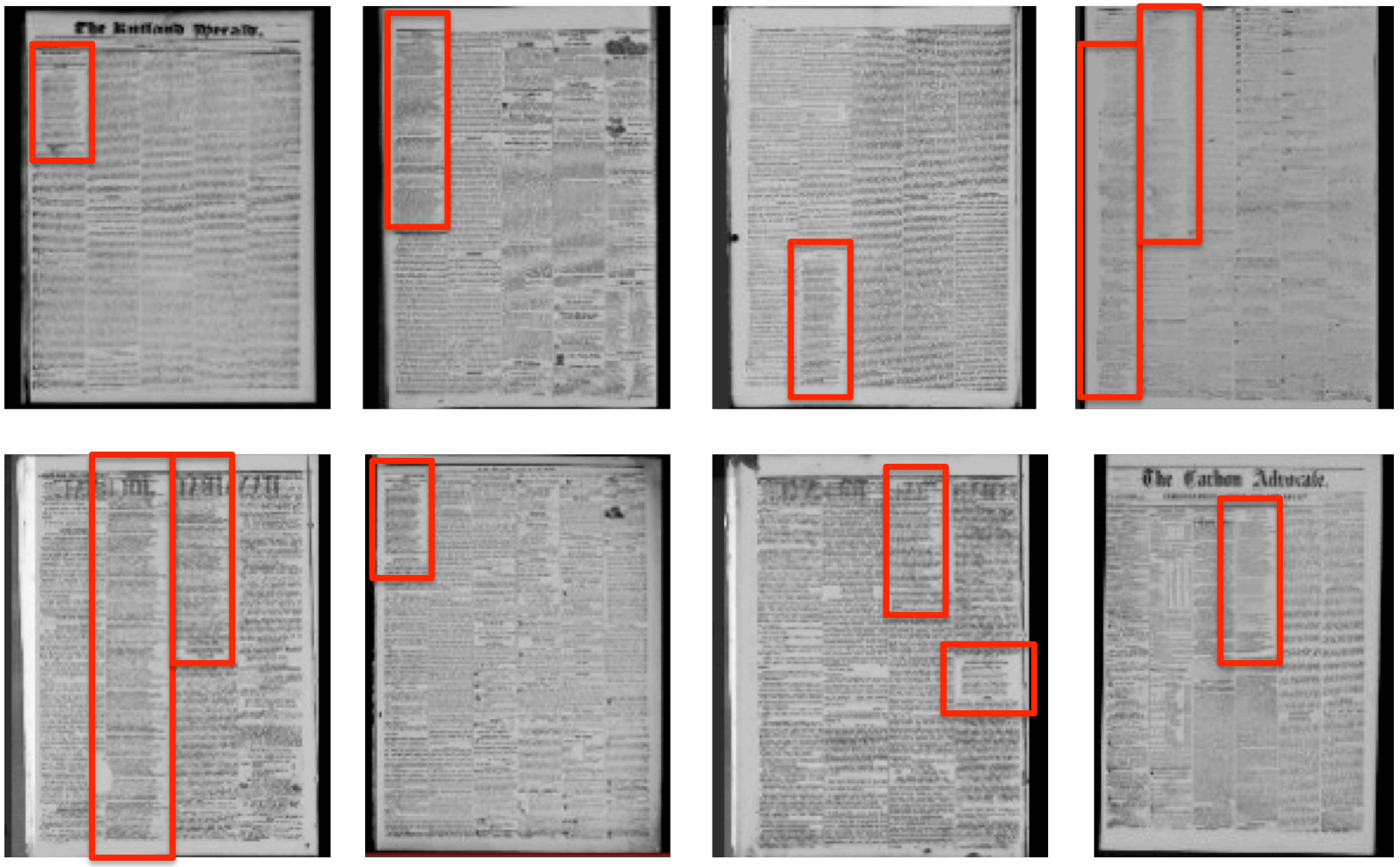
Ir you bright stars which gem the night
Be each a blissful dwelling sphere;
Where kindred spirits re-unite
Whom Death has torn asunder here,
How sweet it were at once to die,
And leave this blighted orb afar!
Mix soul and soul to cleave the sky,
And soar away from star to star!
But oh! how dark, how drear, how lone

t
J"' " ' ' " mjiu. i. u ami INK'' Ml III.:.",...J
vom'tif. xi.ii.
KI TI,M. TiifNilnv
;v I'm in: it m.
C I) r il u 1 1 an to Strain.
rrauman rvr.rTC;f)T, at hot Lamp, vt.
WILLIAM l'AV.
I'OKT n V.
m n i. (i n v .
T r 1 1 1, u x I. seer. TT.
If xn bright ttr wl.eli zm h n'jSI
I ! a M.miViI ttwclKor. tphtrfi
Whr liindf sl tptrili rr.unil
Whom Dssth ht lorn itunltr hue,
How iwi' M vcr l orw lo die,
AbJ Irr lhl blighted ntbif.r!
Mis tool and u to clear the Vy,
Aol i&arawsy frm i!r In ittr !
HutnS' how Jark.hnw drsr, liow Inn
WnuU fftn tit btUhtett world of Mm,
tt, wo Itrin; through rch mdient on,
W fill to find (lie ..Tr.l of thu !
If llur im mure th list thill Itrifr,
Which 0lli eolj hnd lone tin icvsr.
Ah ! Oirn tliwn tlsri in nvwktry thfne
More hittful, ii ihty thine fur vt r !
Itean.mt be neh hops in frr,
Thil li;hti ths erf, of clmijt th brow,
rrwUimi thrr it hupphr tphtre
Thin lliii blriV world linl.Ii ui now !
Thre it voir which Snrrntr hsr,
Who heavietl weijhi J.ifo't jailing cliiin;
Til llesTrn thai whitpr ri "dry lliy lurt,

"Melody," by William Leggett, published in the *Rutland Herald* on April 12, 1836 and its OCR interpretation available via Chronicling America (http://chroniclingamerica.loc.gov/lccn/sn84022355/1836-04-12/ed-1/seq-1/).

Appendix B: Appearance of Poetic Content in Historic Newspapers

A variety of newspaper pages showing the visual distinctness of poetic content, visible even in this small scale. The following page reproduces the images and highlights the poetic content.



Newspaper images from the previous page, with poetic content highlighted.

Appendix C: Training and Deployment of Image Classifier for Poetic Content in Historic Newspapers

Overview

The image processing component of the project consists of two important phases: training and deployment. During the training phase, the goal is to produce a classifier that is able to classify an image as either a poem or non-poem image, where a poem image means that the image consists of segments of a poem. To produce the classifier, a training dataset has to be prepared and fed into a machine learning-based classifier. In the following, we outline the specific steps for this phase. After the classifier is produced, we will then move to the deployment phase. During this phase, the steps used to prepare the training sets are streamlined to automatically process and classify new images.

This image processing-based approach has three potential advantages. First, it can be used to detect or identify other types of texts such as advertisements, obituaries, birth and marriage announcements, and headlines—as long as they have distinctive visual cues—after re-training the classifier with the corresponding training sets. Second, it can be used to detect poems in other languages, assuming that those poems appear in similar patterns as English poems have, since it does not depend on textual information. Third, it can be used to handle noisy or low-quality digitized newspapers because of its ability in looking at the "big picture" visually.

Training

Figures 1, 2, 3 and 4 show the four stages during the training phase.

The first stage involves manual extraction of snippets of images from digitized newspapers. For our training, we will have three sets of snippets: (1) at least part of a poem appears in the snippet, (2) the snippet contains no part of a poem, and (3) a snippet where visual cues are similar to a poem. For our initial dataset, we have prepared 66 such snippets, 22 for each set. Each snippet is about 200 pixels (number of columns) by 300 pixels (number of rows).

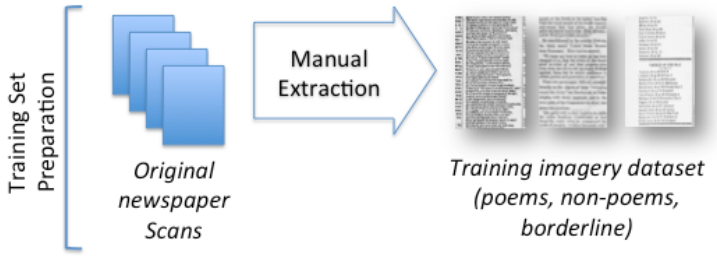


Figure 1. Training phase: preparing training imagery dataset.

For stage 2, we aim to convert each original snippet into a binary image, as shown in Figure 2. Because each snippet is inherently noisy and could be of low quality, we will first perform noise removal. For now, we will perform 3x3 averaging to smooth out noisy pixels, a step known as blurring. Other additional steps to be considered if necessary include histogram equalization (Pizer et al. 1987) to enhance image contrast. Then, to convert the blurred snippet into a binary image—effectively identifying the background pixels from object pixels—so that texts are object pixels, we will use a bi-Gaussian (or bi-normal) curve approximation to automatically obtain the binary segmentation threshold. Co-PI Soh has previously used this technique to successfully segment remote sensing images (Haverkamp et al. 1995; Soh et al. 2004), where images had two distinct classes such as the newspaper snippets.

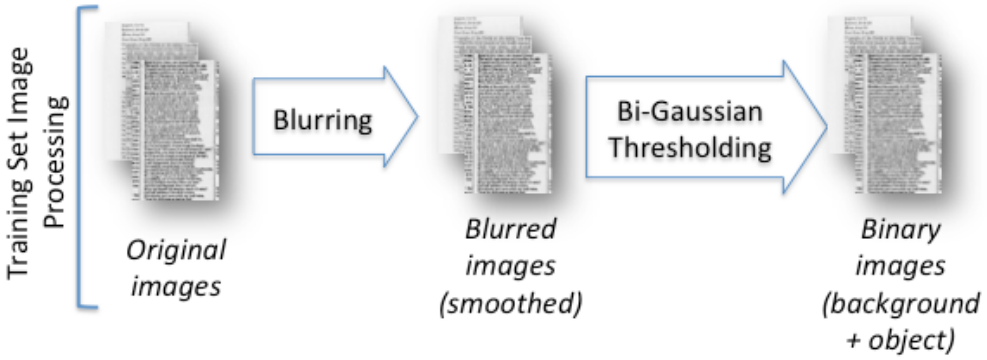


Figure 2. Training phase: converting original images to binary images—background and object pixels.

After obtaining the binary images, the next task involves representing and extracting visual cues as salient features from them. Presently, we aim to compute the following: (1) the left and right margins (average, standard deviation, and distribution), (2) the white spacing between each pair of adjacent lines of text (average, standard deviation, and distribution), and (3) the cross-section profile from the top to the bottom row of each snippet, for each column of pixels. Some of these will be represented as numbers (e.g., average, standard deviation) and some will be represented as a vector of numbers (distribution, cross-section profile).

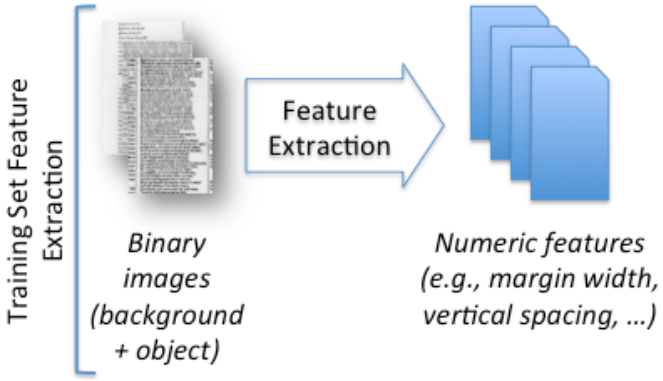


Figure 3. Training phase: extracting salient features—visual cues—from binary images to obtain their numeric representation.

Now, with the imagery data re-represented as numeric data, we will be able to make use of machine learning techniques to train a classifier. For our classification task, we will investigate three widely used and successful machine learning approaches with very different properties. First, artificial neural networks (ANNs) learn a vector of weights on features in the dataset to choose the labels for new data (Witten et al. 2009). ANNs consist of multiple nodes connected to threshold functions or to additional layers of nodes. ANNs are updated iteratively (e.g., using gradient descent) until they correctly predict the labels for the training data. Second, decision trees (for classification) learn a tree data structure to generate the labels for new data (Witten et al. 2009). The decision tree first selects one feature as the root node and adds an edge for every label value. The decision tree continues to add nodes and edges recursively until all the training data has been sorted into groups with similar labels. The leaves are then set to the common label. Third, support vector machines (SVMs) learn a hyperplane to separate the training data such that data on the same side mostly have the same label (Witten et al. 2009). SVMs first use a kernel function to transform all values for the dataset into higher dimensional space where they are linearly separable. Then, the SVM attempts to maximize the distance (i.e., margin) between the training data with different labels.

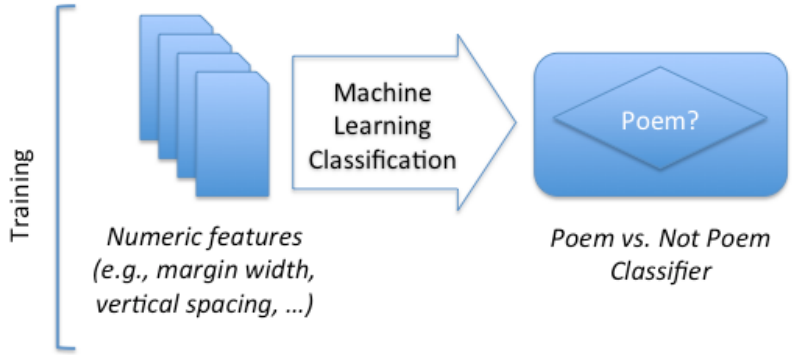


Figure 4. Training phase: training a machine-learning-based classifier to determine whether an image is a poem image or non-poem image based on the extracted numeric representation.

Deployment

After testing and confirmation that the classifier can classify poem vs. non-poem snippets with acceptable accuracy, we will then streamline the above stages and automate them. That is, only the sequence of techniques that results in the best classification accuracy is automated and used. Features that are not helpful will not be extracted, for example, to speed up the process. When deployed, new newspaper scans will be fed into the system as input, as shown in Figure 5. The system then generates a classification decision. Note also that when deployed, it is more important to have high recall—identifying as many poem snippets as possible—than high precision—making as few false identification as possible—since human experts will have a chance to examine and filter out false positive snippets but will likely not have time to review all false negative snippets. Thus, our classifier will be trained with a bias towards high recall, thus willing to sacrifice precision to a certain degree.

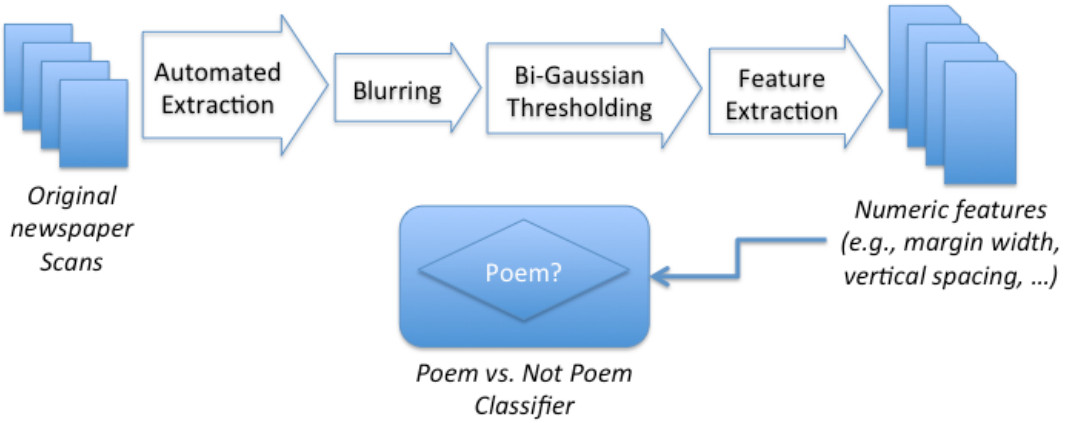
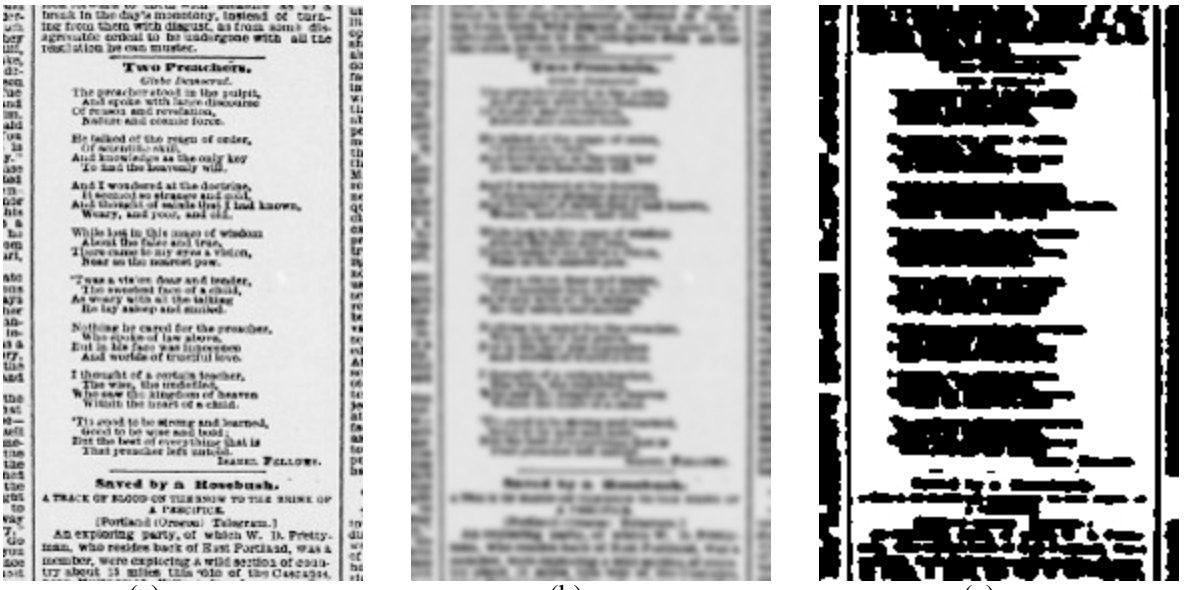
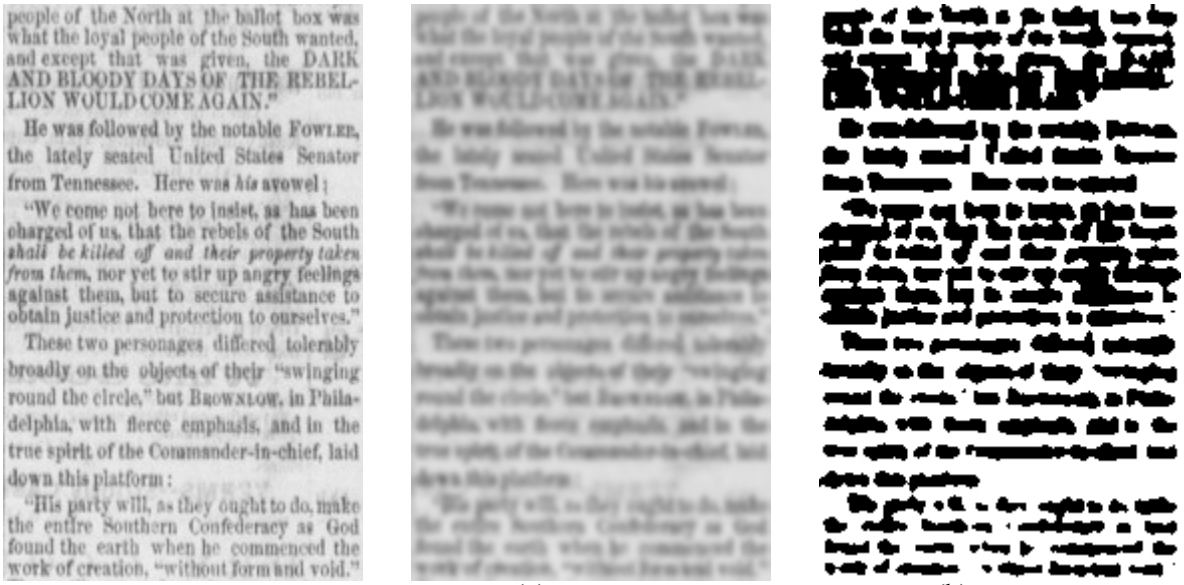


Figure 5. Deployment phase: all processing steps are automated to ultimately represent newspaper snippets in numeric features to be fed into the trained classifier.



(a) (b) (c)

Figure 6. An image segment that has poetic content: (a) original segment, (b) blurred segment using a 2x3 averaging filter, and (c) the binary image automatically generated using bi-Gaussian approximation. Clear margins and also periodic patterns can be seen on the binary image.



(a) (b)

(c)

Figure 7. An image segment that does *not* have poetic content: (a) original segment, (b) blurred segment using a 2x3 averaging filter, and (c) the binary image automatically generated using bi-Gaussian approximation.

References

Haverkamp, D., L.-K. Soh, and C. Tsatsoulis (1995). A Comprehensive, Automated Approach to Determining Sea Ice Thickness from SAR Data, *IEEE Transactions on Geoscience and Remote Sensing*, 33(1): 46-57.

Pizer, S. M., E. P. Ambum, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld (1989). Adaptive Histogram Equalization and Its Variations, *Computer Vision, Graphics, and Image Processing*, 39(3): 355-368.

Soh, L.-K., C. Tsatsoulis, D. Gineris, and C. Bertoia (2004). ARKTOS: An Intelligent System for Satellite Sea Ice Images, *IEEE Transactions on Geoscience and Remote Sensing*, 42(1):229-248.

Witten, I., E. Frank, and M. Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier.