An open-source tool for the transcription of paper-spreadsheet data: Code and supplemental materials available online: Https://github.com/deskool/images to spreadsheets

An Open-Source Tool For The Transcription of Paper-Spreadsheet Data

Code and supplemental materials available online: https://github.com/deskool/images_to_spreadsheets

Mohammad M. Ghassemi *, Willow Jarvis*, Tuka Alhanai*, Emery N. Brown*, Roger G. Mark*, M. Brandon Westover†

*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 02139
{ghassemi,wjarvis,tuka,enb,rgmark}@mit.edu

†Department of Neurology, Massachusetts General Hospital, Boston, MA, 02114
mwestover@mgh.harvard.edu

***Abstract*—Clinical researchers, historians, educators and field researchers alike still regularly capture data on paper spreadsheets. In the case of health care and education, data will often contain sensitive personal information, further complicating the process of transcribing paper-based archives into digital form. In this work, we describe a tool that utilizes machine learning and crowd intelligence to automatically transcribe images of paper-based spreadsheets into electronic form while protecting sensitive personal information. Our solution consists of four high-level stages: (1) the extraction of cell-level images from the spreadsheet grid, (2) machine recognition of digits within the cells, (3) human transcription of cell contents that the machine was uncertain of and (4) feedback of human transcription results to the machine to improve future classification performance. We test the algorithm on a novel data-set of 135 heterogeneous clinical flow-sheet images collected from the Massachusetts General Hospital (MGH), 2 hand-drawn spreadsheets, one chalk-board drawing, and one printed table. we demonstrate that our algorithm provides a generalized solution for spreadsheet transcription that maintains privacy, is up to 10 times faster and twice as cost effective than existing alternatives. Our work is valuable both as a tool and as a starting point for the development of better algorithms.**

***Keywords*-Software, Image Segmentation, Crowd-Sourcing, Optical Character Recognition, Transcription**

I. Introduction

All data must be digital before it can be processed; but not all data that requires processing is in a usable digital format. While it is rare for data scientists to interact with non-digital data, many clinicians [1], historians [2], educators and field researchers [3] still regularly capture or must work with historical archives of paper-based spreadsheet data. For small datasets, manual transcription of these records is feasible, but as data requirements grow, researchers and professionals are required to invest considerable resources to transcribe their paper-based data into digital form [4], [5].

The medical profession is one example of an industry that continues to utilize paper records, even after immense government investments. In 2009, United States (US) legislation provided nearly 20 billion dollars in government aid to health care providers to subsidize the "meaningful use" of electronic medical record (EHR) systems [6]. As of 2016, nearly 96% of US hospitals now have some form of EMR system. Nearly the same proportion of care providers also still utilize paper records at some point in their care process.

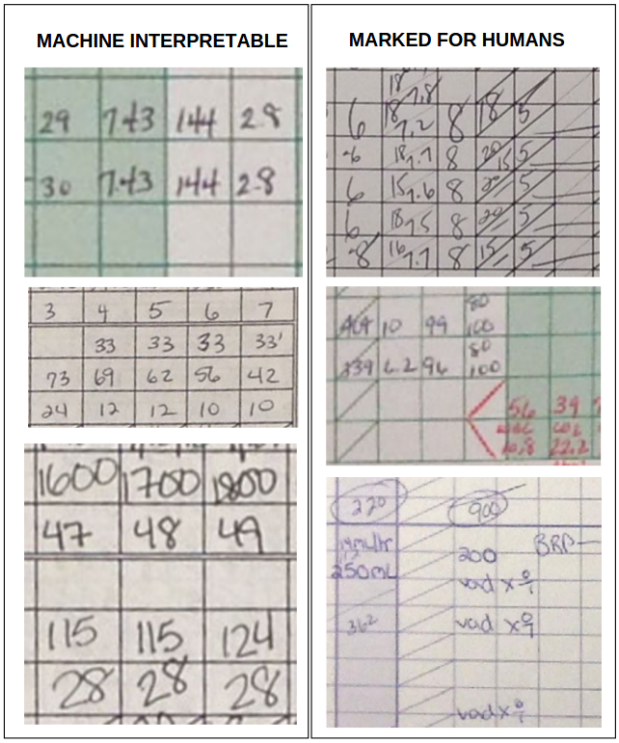

Figure 1. A set of spreadsheet segments from our highly heterogeneous clinical spreadsheet dataset. The left column of images represents cells from real medical spreadsheets which may be machine interpretable, while the right hand column represents a task which is better suited for human transcribers.

The transcription of paper records is a problem without

a cost-effective, open-source solution. Existing approaches involve employing a research assistant, or a professional transcription service. However, transcription services are not always a viable option due to budgetary constraints or data privacy concerns. Data in the educational and health-care domains, for instance, often contain sensitive personal information requiring specialized authorization to share with third parties. These constraints make transcription arduous and costly [1].

A machine learning agent would be a convenient solution, reducing costs and easing task burden while limiting the exposure of sensitive information to third parties. Despite impressive advances in computer vision [7], automated handwriting interpretation [8], and object recognition [9] in recent years, however, there is no fully automated solution to the spreadsheet transcription problem. This may be due to the fact that many of the most effective machine learning algorithms are supervised, requiring large fully-annotated data-sets for training. To our knowledge, such a data-set for spreadsheets is not yet available within in the public domain [1].

The challenges that must be overcome when developing a generalized spreadsheet transcription algorithm are of three varieties. First, there are challenges due to inter- and intra-heterogeneity in the fundamental attributes of the spreadsheets themselves, which often contain varying background colors, cell sizes, table structures, and line orientations (See Figure 1). The second set of challenges is introduced when the physical spreadsheet is captured as a digital photo, where non-uniform lighting and the orientation of the camera can further confound the original image. The third (and perhaps greatest) challenge of transcription results from what is actually written on the spreadsheet. Data that are meant for a single cell may be written outside the cell's border, individual symbols within the cells may or may not overlap (e.g. print vs. cursive handwriting) and different agents may utilize different fonts [10], ink colors, or emphasize a special status of data by circling, drawing arrows, underlining or crossing out entries [11]. Many of these specific challenges, such as handwriting translation or object recognition under non-uniform illumination, are areas of research in their own right. We highlight this to make it clear that the task at hand, while easy for a human agent, presents a non-trivial machine learning challenge [12].

Given the ease of the task for human agents, crowd-sourcing represents one possible solution to the spreadsheet transcription problem. In 2011, Lasecki *et al.* compared multiple crowd-sourcing approaches to transcribe both hand-drawn and digital spreadsheets in real-time [13]. Not surprisingly, the authors found that the efficacy of the crowd depends on both the complexity of the task and the degrees of freedom provided to the worker population (with more degrees of freedom leading to greater errors). That is, a task requiring workers to transcribe an image of a ten-by-ten spreadsheet is more likely to introduce errors than if the same users were asked to transcribe each of the 100 cells in the spreadsheet, one at a time. In 2014, Vaish *et al.* performed a more targeted study comparing human transcription strategies for the digitization of medical records. The authors reported that although crowd-sourced transcription was less expensive than its alternatives, it was also less accurate [11].

Existing for-profit solutions to the spreadsheet-transcription problem utilize a mixed human-machine approach. In 2012, Chen *et al* described a for-profit tool, SHREDDR, to assist in the transcription of paper forms in resource constrained environments [14]. Microtask, and IBM's DataCap are two other for-profit solutions that, like SHREDDR, utilize a combination of machine and crowd intelligence.

As one would expect of for-profit tools, the technical details of the algorithms and the availability of source code is lacking, preventing investigators from independently evaluating or improving upon existing approaches. The disadvantage of these for-profit solutions is not limited to their cost or lack of reproducibility. Existing solutions necessarily cater to the needs of larger organizations with larger transcription needs, thereby reducing the priority of small- to mid-sized transcription projects, typical of the research community. Indeed, one of the solution providers (Microtask) explicitly states on their website that, due to efforts required to configure the tool, they tend not to work on projects with fewer than several thousand documents.

In this paper, we describe an open-source continuously learning transcription tool that utilizes a combination of machine and crowd intelligence to transcribe paper-based spreadsheets into structured digital data. We report the performance of our tool on four distinct experiments: (1) The extraction of cell-level images from the spreadsheet grid (2) machine recognition of digits within the cells, (3) human transcription of cells for which the machine is uncertain and (4) feedback of human transcriptions to the machine to improve future classification performance.

Our tool was designed to interpret highly heterogeneous spreadsheet images. It makes no assumptions about uniformity in image lighting, cell sizes, or the properties of the row and column lines in the spreadsheet. Development of this tool was motivated by a research project performed by the authors which required manual transcription of a large medical flow-sheet archive [15]. We thus designed the tool to reduce the time and cost of transcription, while respecting patient data privacy concerns. To aid others in the community and encourage further development of the tool, we have made our Matlab implementation of the algorithm publicly available online: *https://github.com/deskool/images_to_spreadsheets*. To our knowledge, this is the first publicly-available tool that solves the transcription task using a combination of human and

machine intelligence.

II. Methods

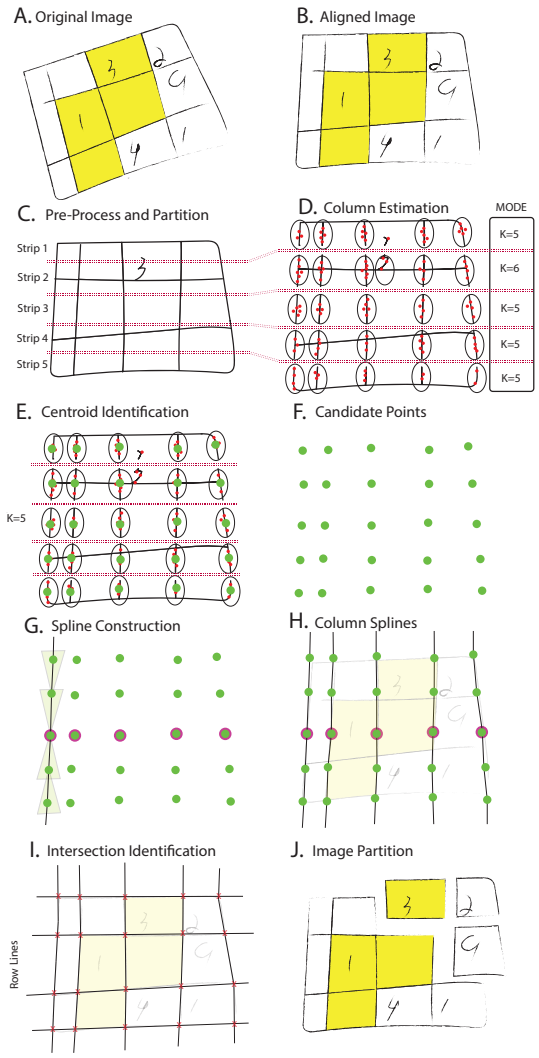

Figure 2. An overview of the unsupervised cell-extraction algorithm. (A) Example of a hand-drawn spreadsheet which we may wish to transcribe. (B) The spreadsheet after manual alignment of the image. (C) Black and white image following preprocessing with image strip lines. (D) Estimation of column number using k-means clustering on Hough transform vertical line peaks (illustrated as red dots). (E) Extraction of candidate points using k-mean centroids with k set according to the estimate. (F) A set of extracted column point candidates. (G) Column splines through the combination of candidate points. (H) Completed column splines. (I) Identification of intersection points between row and column splines. (J) Partitioning of the original image into individual cells. A highly detailed description of this approach may be found in the Supplemental Materials.

A. Data

The performance of our tool was evaluated on a total of 139 spreadsheet images containing 35,930 distinct cells of data: 135 clinical flow-sheet images collected from three intensive care units at the Massachusetts General Hospital (MGH), two hand-drawn spreadsheets, one chalk-board drawing, and one printed table. For validation purposes, we manually extracted the number of rows, columns, and cells from each of the spreadsheets. This study was conducted under a protocol approved by the MGH institutional review board.

B. Algorithm

The algorithm for automated transcription consists of four high-level stages: (1) Unsupervised extraction of cell-level images from the spreadsheet grid (2) supervised machine recognition of digits within the cells using a Multinomial Support Vector Machine (MSVM), (3) human transcription of cells that the MSVM was uncertain of via Amazon's Mechacnical Turk workers and (4) feedback of human transcriptions to the MSVM to improve future classification performance. We *strongly encourage* the reader to download our Matlab implementation of the algorithm, which is publicly available, and has been annotated to reflect each of the steps outlined in this paper. *The core methodological contribution of this work is in the unsupervised cell-extraction algorithm*, which we illustrate in Figure 2. Grid Line identification and training of the MSVM classifier are described in full detail in the Supplemental Materials.

C. Experimental Approach

To evaluate the performance of our tool, we present four experiments that investigate the algorithm's performance, speed, and cost at each of its four stages: (1) Extraction of cell-level images from the spreadsheet grid, (2) machine recognition of digits within the cells, (3) human transcription of cells that the machine was uncertain of and (4) feedback of human transcriptions to the machine to improve future classification performance. We then compare the results of our algorithm against a human research assistant, a machine-aided research assistant, the crowd-alone and a machine-aided crowd.

1) Experiment 1: We investigate the performance of our unsupervised cell extraction algorithm on the 139 collected spreadsheets. Specifically, we report the performance of our algorithm in row and column spline identification and grid-line extraction using the parameter settings described in the Supplemental Materials.

2) Experiment 2: We investigate the performance of the crowd workers on the transcription task. To evaluate the effectiveness of this approach, we compare it against (1) the estimated cost of a certified research assistant to perform machine-aided transcription where *only the contents* of the cell images generated by our cell extraction algorithm are transcribed, and (2) the estimated cost of a certified research assistant to fully transcribe the original spreadsheet images into CSV files. We empirically illustrate that transcription of full spreadsheets will take longer, per cell, as it requires the

research assistant to (1) identify the contents and location of the cell in an image (e.g. '2' in row three, column three) (2) identify the corresponding location in the CSV file (3) transcribe the entry and (4) identify the next cell in the image to be transcribed.

3) Experiment 3: We investigate the classification performance of the MSVM on the subset of extracted cell images which contained digits or blanks. We report a lower-bound on the proportion of correctly classified digits, the total proportion of correctly transcribed cells, and evaluate the performance of the classifier for various MSVM confidence thresholds. We also report the performance of the MSVM on a held-out portion of the MNIST data-set. For more information on the MSVM is provided in the Supplemental Materials

4) Experiment 4: We retrain the MSVM algorithm using 80% of the data transcribed by the Mechanical Turk workers and the corresponding digits extracted from the images. We test our updated MSVM on the remaining 20% of the human annotated data, and report the improvement in classifier performance compared to the MSVM trained in Experiment 3.

III. Results

A. Experiment 1: Spline Identification and Grid Extraction

The average number of cells in the 139 collected spreadsheets was 258.48 cells, with a standard deviation of 155.03 cells. The smallest flow-sheet contained 10 cells (5 rows, by 2 columns), while the largest flow-sheet contained 884 cells (26 rows by 34 columns). Using the default cell extraction parameters shown in the Supplemental Materials, our approach was able to estimate the location of 97% of all row and column splines. In Figure 3 we show the results of row and column spline estimation on two very different spreadsheets from our data-set. Despite large differences in the structure of the flow-sheets, the algorithm was able to extract the row and column lines correctly.

B. Experiment 2: Human Transcription

Amazon Mechanical Turk workers were employed to transcribe the cell-images extracted from the spreadsheets. According to the crowd's transcriptions, 61% of cells were blank, 19.4% contained only digits, 10% contained only written text, while the renaming 9.6% contained a mixture of digits and text. The total number of characters across all cells was 26,484, of which 3,838 characters (15%) were numerical digits. We found that 26% of the total, or 66% of non-blanks contained contents that were touching, or were outside of the cell borders.

There were an average of 11 workers performing the transcription task in parallel at any given time. The total time for the transcription task was 196 minuets (3.26 hours). The average time from acceptance of the task to submission of the transcribed cell image was 7.24 seconds with a total

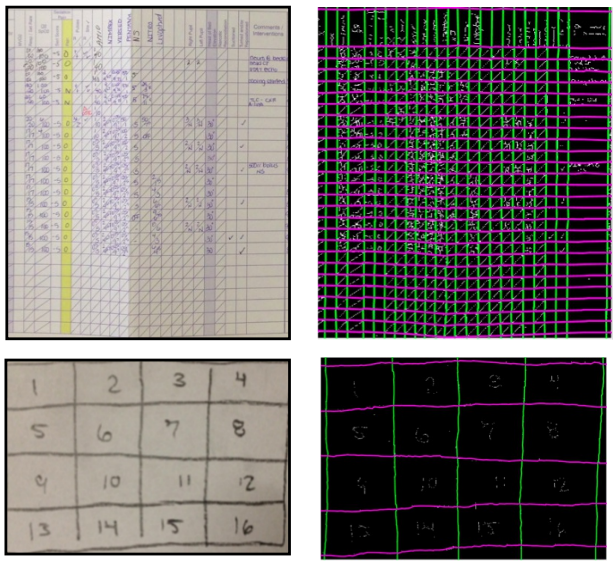

Figure 3. Examples of row and column spline identification for two images. Original images include a medical flowsheet with non-uniform illumination and background (top left), and a hand-drawn spreadsheet with non-ideal row and column lines (bottom left). The images on the right side of the figure show the corresponding row splines (in purple) and column splines (in green) for both spreadsheets.

reward of $0.01 per cell (an equivalent of $4.97 per hour of labor). The total cost of the transcription was $349.10, $174.55 of which went towards to the cost of human workers, *with an equivalent amount spent on Amazon service Fees*.

To evaluate the effectiveness of this approach, we compared it against two alternatives: (1) The estimated cost of a clinical research assistant to transcribe *only the cell contents* of the spreadsheets generated by our algorithm and (2) the estimated cost of a clinical research assistant to transcribe the *original* spreadsheet images into CSV files. The results of this cost comparison are shown in Table I, and are also described below. We note here that the Amazon fees may be reduced by up to 80% by clustering multiple cell images within a single classification task. Hence, what we will report below represents a lower bound on potential savings.

1) Full Transcription: The median hourly wage of a clinical research assistant in the United States is $22.15 [16]. We estimated the average rate of cell transcription for a certified research assistant by manually transcribing a full 10x10 spreadsheet image into a corresponding CSV file. The total time for spreadsheet transcription was 12.5 minutes. The same 10x10 spreadsheet was also submitted to Amazon's Mechanical Turk service. The crowd-based solution was less than half the price ($349.10) and 6.7 times faster than the research assistant.

2) Machine-Aided Transcription: Next, we estimated the rate of cell transcription for a certified research assistant

M - Number of Cells in Task N - Number of Crowd Workers D - Number of MSVM Transcribed Cells C - Machine Confidence	**Individual Cost/Hour**	**Individual Cells/Hour**	**Individual Cost/Cell**	**Total Cost/Task**	**Total Hours/Task**
Research Assistant	$22.15	480	$0.046	$0.046M	$\frac{M}{480}$
Machine-Aided Research Assistant	$22.15	1034	$0.022	$0.022M	$\frac{M}{1034}$
Crowd	N/A	320	$0.02	$0.02M	$\frac{M}{320N}$
Machine-Aided Crowd	N/A	497	$0.02	$0.02M	$\frac{M}{497N}$
Machine-Aided Crowd + Machine Classification	N/A	497	$0.02	$0.02(M-D)	$\frac{M-D}{497N}$
Example Scenario: M = 35,000 Cells, N = 10 Workers,C = 95%, D = 3,500					
Research Assistant	$22.15	480	$0.046	$1,610	72.92
Machine-Aided Research Assistant	$22.15	1034	$0.022	$770	33.84
Crowd	N/A	320	$0.02	$700	10.94
Machine-Aided Crowd	N/A	497	$0.02	$700	7.05
Machine-Aided Crowd + Machine Classification	N/A	497	$0.02	$630	6.30

Table I
A COMPARISON OF INDIVIDUAL AND TASK LEVEL TRANSCRIPTION COSTS AND TIME REQUIREMENTS FOR A RESEARCH ASSISTANT WITH AND WITHOUT MACHINE-AID, THE CROWD WITH AND WITHOUT MACHINE-AID, AND OUR PROPOSED APPROACH. M DESCRIBES THE NUMBER OF CELLS IN THE TASK, N DESCRIBES THE NUMBER OF CROWD WORKERS, D DESCRIBES THE NUMBER OF CELLS TRANSCRIBED BY THE MSVM AND R DESCRIBES THE RATE OF MSVM CLASSIFICATION.

to perform transcription of a 10x10 spreadsheet from the machine extracted cell images. That is, the cells of the spreadsheets were presented one at a time, and the placement of the transcribed contents was performed automatically by our solution. Using this approach, the total time for image transcription was 5.8 minutes. Hence, the machine-aided research assistant was able to complete the transcription at nearly twice the rate (and half the cost) of a research assistant without machine-aid. Assuming an average of 11 workers, as we observed in our experiment, the machine-aided crowd is less than half the price and 11.4 times faster than the research assistant without machine aid. Compared to the machine-aided research assistant, the crowd-based approach is still 5.16 times faster and remains the more cost effective option (at least 7% cheaper) for our transcription task.

C. Experiment 3: Machine Classification

To begin, we tested the classification performance of the MSVM on a held-out portion of the MNIST data-set. Our accuracy on the testing set was above 96%. This performance is within 3% of the best reported "Deep Learning" OCR models on the MNIST dataset in the literature [17], which is sufficient for the purpose of our tool. Next, we tested the overall classification performance of the MSVM on the 12,078 numeric and blank cells from our data. In total, our MSVM was able to correctly transcribe 66% of the numeric cells without any error. The algorithm was also able to correctly identify 70% of the blank cells correctly.

As we described in the Methods section, a realistic implementation of this algorithm would only utilize the machine learning algorithm to classify cell contents above a given confidence threshold, and would submit all other cells to the crowd for human annotation. In the top two plots of Figure 4, we show the results of our MSVM's cell and digit classification performance as a function of the selected confidence threshold. As we expect, the performance of the algorithm improves on the subset of cells and digits for which its confidence is high (solid lines). However, this improvement in performance is not without cost. As the confidence threshold of the algorithm is increased, the proportion of the total data which the algorithm will classify decreases (dashed lines). At a confidence threshold of 99%, cell contents are correctly classified 90.4% of the time. However, only 5.6% of the total numeric cells qualify for machine transcription at this confidence level. Hence, a lower bound for savings using this approach is 5.6%.

D. Experiment 4: Human-Machine Feedback

In an attempt to further improve the performance from Experiment 3, we randomly selected 80% of the human transcribed spreadsheets, and included the extracted cell images in the training set of a revised MSVM. We compared the improvement in performance at various confidence levels between the original and revised MSVMs on the remaining 20% of the spreadsheet images.

In the bottom two plots of Figure 4 we illustrate the *improvement* in our algorithm's accuracy on the held-out testing set after re-training. Importantly, the revised MSVM algorithm exhibited a general trend of improved cell and digit classification performance at the higher confidence levels. Near the 80% confidence threshold, for instance, the cell classification algorithm exhibited nearly 10% improvement in accuracy. With these improvements, the lower bound for savings using this approach this approach would discount the total cost of the human transcriptions task by an additional 5% (for a grand total of 10%). With more data, the improvement in classification performance would only continue to improve, enabling even higher cost savings.

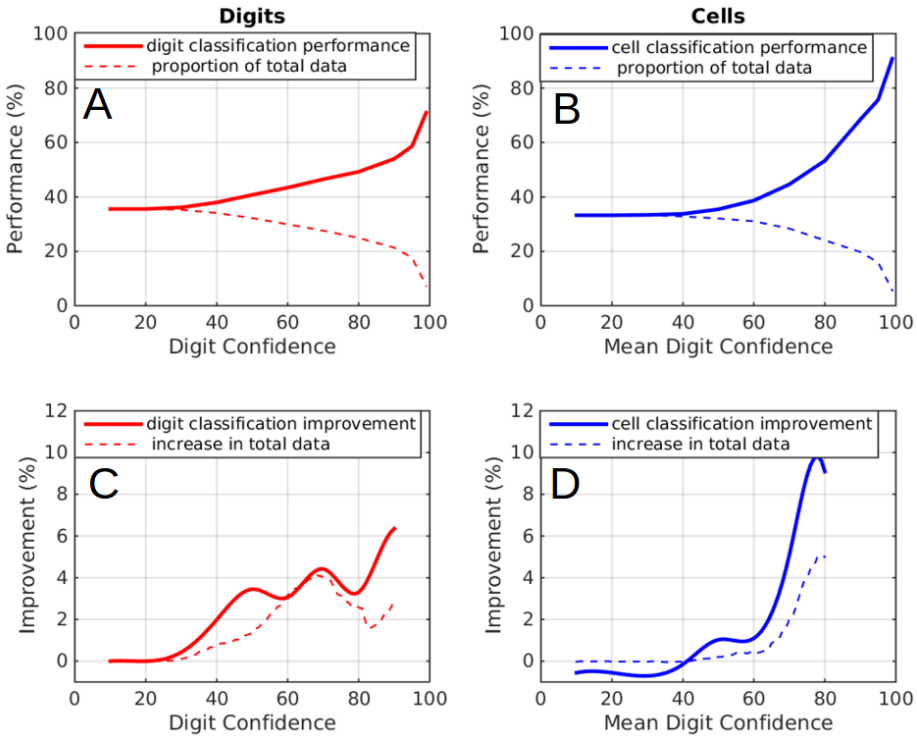

Figure 4. The accuracy of the Multiclass Support Vector Machine in digit and cell level classification and the improvement in accuracy on a held-out test set after retraining the algorithm with 50% of the collected data. (A) shows the digit level classification, (B) shows cell level classification, (C) shows improvements in digit classification after feedback of human annotations (D) shows improvements in cell classification after feedback of human annotations. The "mean digit confidence" refers to the mean confidence of the MSVM across all individual digits within the cell.

IV. Discussion

In this work, we described and tested an open-source tool that utilizes a combination of machine-learning and crowd-sourcing to transcribe paper-based spreadsheets into electronic form. The data we utilized in this work was selected to test the performance of the algorithm on a highly heterogeneous real-world data-set which included medical spreadsheets, hand-drawn tables, and chalk-board-drawings.

In Experiment 1, we found that the default parameters of the cell extraction algorithm correctly inferred 97% of the grid lines. We note here that in the few failure cases, the spreadsheet images were found to have had at least one significant abnormality which contributed to failure of the algorithm using the default parameters (e.g. a dramatic bend in the page or a human generated strike-through of multiple rows/columns). This implies that the default parameters of our cell extraction algorithm may not be ideally suited for all spreadsheets and should be adjusted to suit the individual user's needs. In a realistic use case scenario, we expect that users will wish to transcribe many copies of similarly formatted paper-spreadsheets and will not need to estimate the number of rows and columns for each document. In such cases, the row and column estimation process could be bypassed entirely and our algorithm can proceed immediately to candidate point estimation and spline construction. Indeed, for such a task, this is our recommended approach.

In Experiment 2, we utilized a set of human workers to annotate the individual contents of the cells extracted from our spreadsheet images. We demonstrated that a crowd-based approach is faster and more cost effective than the research assistant. Our experimental set-up placed one image within each submitted task. While from a cost perspective this set-up was non-ideal, the partitioning of each cell into a separate job effectively anonymizes the spreadsheets during the transcription process. Given that the crowd-based spreadsheet transcription is a function of the number of crowd workers, the transcription rates of the crowd we reported in this study may vary. For tasks that must be accomplished more rapidly, one may wish to increase compensation, which would attract more workers. Future implementations of this work can also reduce the cost of transcription by combining multiple random images within a single classification task. We decided not to cross-check the performance of the human agents, although a manual spot check of a random 10% of the records revealed no errors in the human transcription.

In Experiment 3, we found that while our algorithm exhibited excellent performance on the MNIST testing set (96%), its performance on the real-world medical spreadsheets was relatively less impressive (66%). One contributor to the lower performance may be our image segmentation

approach, which assumes that digits do not overlap. We acknowledge this as a weakness of our approach and invite other experts in the domain of image segmentation to improve the publicly available algorithm in this regard. Importantly, for the cells with high confidence levels (99%), the accuracy in transcription was above 90%. Future iterations of this approach may also benefit from the utilization of neural networks for the classification task of both the digits and the identification of cell blocks within the image. The latter of these two task, however, may require the development of a fully annotated training set.

In Experiment 4, we fed back the human annotations from 80% of the numerical cell images into our MSVM to improve classification performance. We found improvements in the classification performance of the revised MSVM at higher confidence levels. This improvement in performance translates into cost savings during the transcription task. These results highlight the long-term utility of our approach, which will continuously improve its ability to classify documents as more data is collected. Importantly, the revised MSVM model need not only be trained to recognize digits, but may extend into the classification of other characters as sufficient training data is collected.

REFERENCES

[1] P. Thompson, R. T. Batista-Navarro, G. Kontonatsios, J. Carter, E. Toon, J. McNaught, C. Timmermann, M. Worboys, and S. Ananiadou, "Text mining the history of medicine," *PloS one*, vol. 11, no. 1, p. e0144717, 2016.

[2] I. Hendrickx, M. Düring, K. Zervanou, and A. Van Den Bosch, "Searching and finding strikes in the new york times," in *Proceedings of the 3rd workshop on annotation of corpora for research in the humanities (ACRH-3). The Institute of Information and Communication Technologies, Bulgarian Academy of Scienes, Sofia*, 2013, pp. 25–36.

[3] S. Van Hooland, M. De Wilde, R. Verborgh, T. Steiner, and R. Van de Walle, "Exploring entity recognition and disambiguation for cultural heritage collections," *Digital Scholarship in the Humanities*, vol. 30, no. 2, pp. 262–279, 2015.

[4] M. Jervis and M. Masoodian, "How do people attempt to integrate the management of their paper and electronic documents?" *Aslib Journal of Information Management*, vol. 66, no. 2, pp. 134–155, 2014.

[5] P. Pandey and R. Misra, "Digitization of library materials in academic libraries: Issues and challenges," *Journal of Industrial and Intelligent Information Vol*, vol. 2, no. 2, 2014.

[6] D. Blumenthal and M. Tavenner, "The meaningful use regulation for electronic health records," *New England Journal of Medicine*, vol. 363, no. 6, pp. 501–504, 2010.

[7] J. Malik, P. Arbeláez, J. Carreira, K. Fragkiadaki, R. Girshick, G. Gkioxari, S. Gupta, B. Hariharan, A. Kar, and S. Tulsiani, "The three rs of computer vision: Recognition, reconstruction and reorganization," *Pattern Recognition Letters*, vol. 72, pp. 4–14, 2016.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *Image Processing, IEEE Transactions on*, vol. 24, no. 12, pp. 5706–5722, 2015.

[10] L. Furrer and M. Volk, "Reducing ocr errors in gothic-script documents," in *Proceedings of the RANLP 2011 workshop on Language Technologies for Digital Humanities and Cultural Heritage*, 2011, pp. 97–103.

[11] R. Vaish, S. T. Ishikawa, J. Liu, S. C. Berkey, P. Strong, and J. Davis, "Digitization of health records in rural villages," in *Global Humanitarian Technology Conference (GHTC), 2013 IEEE*. IEEE, 2013, pp. 209–214.

[12] D. Lopresti, "Optical character recognition errors and their effects on natural language processing," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 12, no. 3, pp. 141–151, 2009.

[13] W. S. Lasecki, K. I. Murray, S. White, R. C. Miller, and J. P. Bigham, "Real-time crowd control of existing interfaces," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 2011, pp. 23–32.

[14] K. Chen, A. Kannan, Y. Yano, J. M. Hellerstein, and T. S. Parikh, "Shreddr: pipelined paper digitization for low-resource organizations," in *Proceedings of the 2nd ACM Symposium on Computing for Development*. ACM, 2012, p. 3.

[15] M. M. Ghassemi, E. Amorim, S. B. Pati, R. G. Mark, E. N. Brown, P. L. Purdon, and M. B. Westover, "An enhanced cerebral recovery index for coma prognostication following cardiac arrest," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 534–537.

[16] P. H. Capital, "Clinical research nurse salary (united states)," Online, accessed May 6, 2015.

[17] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1058–1066.