The Art of Teaching Computers:The SIMSSA Optical Music Recognition Workflow System

Ichiro Fujinaga
Music Research, Schulich School of Music
McGill University
Montreal, Canada
ichiro.fujinaga@mcgill.ca

Gabriel Vigliensoni
Music Research, Schulich School of Music
McGill University
Monteal, Canada
gabriel.vigliensonimartin@mcgill.ca

***Abstract*—In many machine learning systems, it would be effective to create a pedagogical environment where both the machines and the humans can incrementally learn to solve problems through interaction and adaptation. We are designing an optical music recognition workflow system within the SIMSSA (Single Interface for Music Score Searching and Analysis) project, where human operators/teachers can intervene to correct and teach the system at certain stages in the optical music recognition process so that both parties can learn from the errors and, consequently, the overall performance is increased progressively as more music scores are processed. In this environment, the humans are learning how to teach the machine more effectively.**

Keywords—optical music recognition, machine learning, machine pedagogy

I. Introduction

In the context of developing an optical music recognition (OMR) software system, we became aware of the importance of human involvement in the entire workflow. Our research is motivated by admitting that it will be difficult and practically impossible to foresee and recognize every possible music notations in the world from the start. We, therefore, need an adaptive learning system that can be taught by people who are familiar with music notation but not necessarily with machine learning methodologies.

One of the strengths of current learning machines lies in their ability to recognize complex patterns, provided that there is a large amount of labeled training data (ground truth). In cases where massive ground-truth datasets are not readily available, one solution is to incrementally and interactively train an adaptive system, with gradual exposure of new data. We argue that in these supervised adaptive learning environments, it is important to study how humans impart their knowledge to the machine, the different teaching methods to achieve the desired performance, and how humans acquire these effective pedagogical strategies.

A. Machine pedagogy

Here, we propose the idea of pedagogy for learning machines as the study of the methods and activities of teaching machines. This pedagogy is about creating an environment where humans can learn the art of how to teach machines running learning algorithms in an incremental learning process in rapid training-inference-correction-feedback iterative cycles.

What we observed in this environment is that the human teacher's abilities to teach the student (machine) improves as the teacher spends more time with the student. This is analogous to human-to-human pedagogy, especially in one-on-one situations, such as music lessons. In other words, the human teacher's pedagogy is gradually modified by observing the responses of the student.

Interactive machine learning [1] is a potentially powerful technique for enabling end-users (e.g., music scholars) interaction with machine learning. Teachers are rewarded with the results they seek quicker (in our case, transcribed music), if they are effective.

The tasks for humans are to find out what is hard for the machine and choose the most effective sets of training data to improve the classification of these hard areas of the problem space. Put it another way, it is not necessary to provide lots of training data for things that the machine can already easily solve [2].

We are proposing to exploit human skills and knowledge to teach machines to optimize their performance. In order to achieve this, we first need to understand how humans interact with a machine-learning component and then we need to build a clever workflow to take advantages of the intelligence of the human and the ability to perform fast calculations of the computer.

An early example of involving humans in improving a learning system iteratively ("human-in-the-loop") is a learning framework by Widmer [3], who developed an interactive environment for creating music counterpoint.

The impact of human intervention in the context of supervised machine learning workflows has been also empirically studied. For example, Fails and Olsen [1] built a system for creating image classifiers and proposed the concept of interactive machine learning for those environments where human teachers evaluate a model created by a learning machine, then edit the training data, and retrain the model according to their expert judgment to improve the performance of the system in the given task.

More recently, Bieger, Thórisson, and Steunebrink [4] proposed a conceptual framework for teaching intelligent systems. They identified the constituent elements of that framework and stated that the interaction between teachers (humans) and learners (machines) has the goal of teaching the learning system to gain knowledge about something or about a specific task. As a pedagogical strategy, we hypothesize that by knowing the learner, and how the learner reacts to correction and new input, teachers can adapt their teaching tactics to improve the pedagogy.

In the following sections, we will provide a general background to OMR and SIMSSA, followed by our particular implementation of OMR and its pedagogical environments.

This research is, in part, supported by the Social Sciences and Humanities Research Council of Canada, Compute Canada, le Fonds de recherche du Québec – Société et culture, and McGill University.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

II. Background

A. Optical Music Recognition

Although OMR research began in the late 1960s the development of this technology—automatically transcribing music notation from a digital image—has been slow (for recent reviews, see [5] and [6]). Although several commercial and open-source OMR software have been available since the mid-1990s, most of them are designed to be used by individuals for small-scale recognition tasks, and for Common Western Music Notation (post 18th century), although there have been some efforts to recognize other types of music notation such as for the lute [7] and for earlier Western music [8].

The automatic recognition of music is considered much more complex than recognition of text via optical character recognition (OCR), as staff lines, notes, lyrics, all need to be identified; multiple voices need to be aligned in polyphonic music; and there are many styles of music notation.

It should be noted that although OCR is quite successful for processing modern documents, for historical multilingual documents, a recent report has shown that the problem is far from solved [10]; despite various efforts (e.g., IMPACT, which includes industry partners such as IBM and ABBYY Production [9]).

In general, there are several steps to consider in building a complete OMR system as shown in Figure 1. The three major steps are: image preprocessing, music symbol recognition, and music notation reconstruction. Within each of these steps, there are several subtasks that may be implemented depending on a particular implementation.

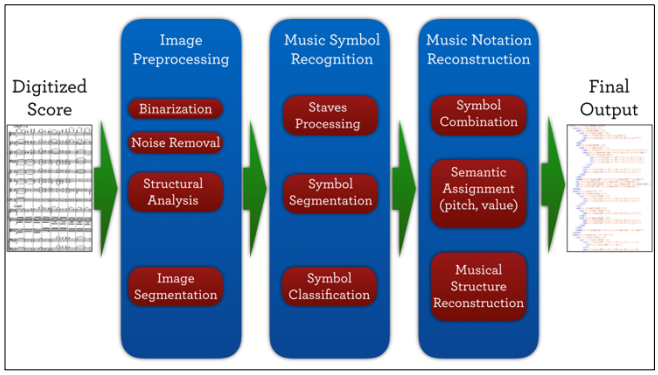

Fig. 1. A workflow of a complete OMR system.

B. SIMSSA

The SIMSSA (Single Interface for Music Score Searching and Analysis: https://simssa.ca) project is a large government-funded multi-year research endeavor involving over 30 researchers, 22 institutional partners, most of them music libraries, and, at any given time, more than 20 students [11]. Our main goal is to make musical scores searchable online the way we can now search text (e.g., Google Books). In order to achieve this, we are developing an OMR system that can be used by almost anyone with some knowledge of music notation. Because of the large volume of musical scores to be processed, we will be relying on the interested people in the music community, such as musicians, music scholars, and students, to develop this online music library.

Our general strategy is not to create a monolithic software that can read any type of music notation from any period in history, but to create an adaptive system where humans can easily and effectively teach the system to deal with specific music notation in manuscripts with different conditions of preservation, whether they are fire-damaged orchestral scores printed in the early 19th century or chant music in neume notation copied by monks on stretched goat skins (parchments) from the 12th century.

The original digital images of the music will be provided by our partners including many of the great music libraries of the world, such as the Bibliothèque nationale de France, the Bayerische Staatsbibliothek, and the British Library. But rather than keeping copies of the images locally, we are using the technology offered by the International Image Interoperability Framework (IIIF) [12] Image API. This allows for the display of high-resolution images served directly from the institutions having the rights to these images. Thus, IIIF enables the access to images of scores from numerous different institutions through one website. These images are hosted at the home institution, thus avoiding the huge cost of storing the images at our site. Only the encoded music (text files) and their indices are maintained so that the searching is possible. As a result, users will be able to search those images and see the exact place on the page of the original digital images.

By offering an easy and free access to a large collection of music scores and manuscripts, with sophisticated search and analytical tools, we are hoping that SIMSSA will transform the ways in which people access cultural music heritage in an unprecedented manner.

III. SIMSSA OMR Workflow

The goal of OMR, in general, is to read and extract the content from digitized images of music documents and encode it into a machine-readable format, so that the music can be searched, analyzed, heard, transformed, and viewed in other notation systems. Despite more than 50 years of research, it remains to be a difficult task.

Because OMR results are never perfect, human intervention is inevitably required to correct the errors generated by the automated process. We exploit this situation by involving more "humans-in-the-loop" and offering them interfaces to teach an incremental, interactive, adaptive machine learning system.

Furthermore, we take advantages of the adaptive system by keeping the model that it has learned, say for a particular manuscript or a set of scores published by a printer, and use it as a starting point for subsequent OMR sessions for a similar manuscript or from another contemporary publisher in the same city [13]. We have previously experimented with this idea by building book-adaptive OMR models for music from microfilms [14]. The experiments showed that human editing costs were substantially reduced and that the approach was especially well suited to handle the various degradation levels of music documents from typographic prints. Similar approaches have been attempted in the text document analysis research [15].

Our entire OMR workflow for neume notation is depicted in Figure 2. This process is divided into several stages: Digitized music scores images are the input to the system. In the Document Segmentation stage (orange boxes in the figure), the Layout Analysis stage, the images are segmented into different layers, such as the background, text, staff lines,

and music symbols in the Layout Analysis step. Immediately after, in the layout correction step, humans intervene to correct the errors and the image with its layers is sent back for another attempt at the analysis. This is repeated until the human is satisfied with the results. In the Symbol Classification stage (purple boxes in the figure), music symbols are recognized and labeled. Again, the results are corrected and sent back to the classifier until satisfactory results are obtained. In the last stage, Music Reconstruction (blue boxes), the system heuristically determines the pitches of the notes and encodes them into the MEI file format [16]. The file is then displayed using Verovio music engraver [17] and edited via the Neon.js neume notation editor [18] so that the final check of the encoded music can be performed. Finally, the MEI file is merged with the necessary metadata (in this case, from Cantus Database[1]) and presented to the user using the Cantus Ultimus Interface,[2] which uses the Diva.js document image viewer [19] that takes advantage of IIIF.

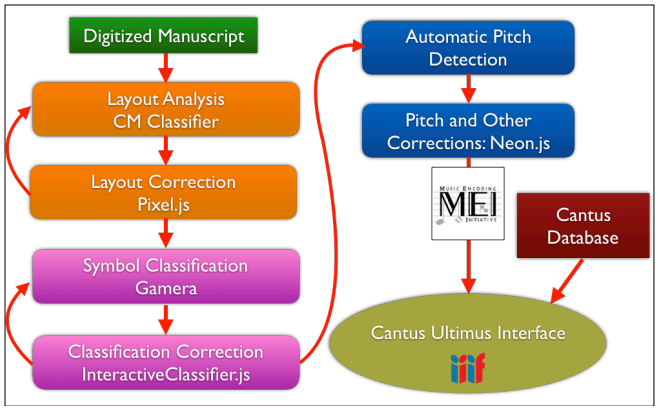

Fig. 2. SIMSSA Neume OMR Workflow.

A. Rodan OMR Workflow Management System

To coordinate all the different components of the workflow, we use Rodan, which is a distributed and collaborative workflow management system [20]. It is specifically designed to allow large-scale OMR processing of entire manuscripts of music. Different stages of a workflow, called jobs, can be scheduled in advance. These schedules can be applied to different target documents in parallel. Rodan allows multiple users from different locations to perform OMR simultaneously.

Each job can be specified to be interactive or non-interactive. Since, at certain stages, our workflow requires a human operator to teach the learning algorithms, we need to be able to create interactive checkpoints, where the system stops a process and waits for user input. Rodan is also responsible for housekeeping tasks such as the file management and indexing of the transcribed MEI files to be used by a search engine.

In the next section, we present the two interactive environments we have developed for teaching the machine how to perform tasks in the first two stages of the OMR workflow.

IV. Pedagogical Environments

Currently, there are two pedagogical environments where humans interact with the machine-learning algorithms with appropriate user interfaces. These are the Layout Analysis and the Symbol Classification stages in our OMR workflow.

A. Layout Analysis

The first stage in the OMR workflow is the document layout analysis. In this stage, all pixels of the music score image are classified into one of the pre-defined layers (e.g., musical notes, lyrics, staff lines, ornamental letters, etc.), so that the input image is segmented into parts. Most approaches for layout analysis have been developed using heuristic techniques that rely on specific characteristics of the documents, and so these methods usually have not generalized well to music documents of a different type or era. For high scalability, we are taking a machine learning-based approach for layout analysis. Since we need training data as examples for creating a model to recognize the different layers within an image, and creating ground truth from scratch is onerous and expensive, we have tested a few approaches for teaching the computer to perform the image preprocessing.

Most image preprocessing techniques (based on heuristic or machine learning approaches) output a non-negligible amount of misclassified pixels, and so we developed Pixel.js (see Figure 3), an open source, web-based, pixel-level classification application to correct the output of image segmentation processes [21]. We use this tool interactively with a selectional autoencoder- based classifier [22], to create ground-truth data incrementally. This type of interactive learning is well-motivated in many modern machine learning problems, where unlabeled data may be abundant or easily obtained, but labels are difficult, time-consuming, or expensive to obtain.

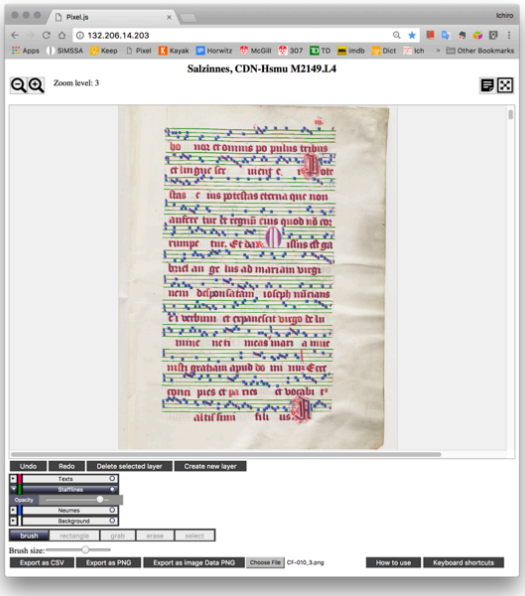

Fig. 3. Pixel.js

[1] http://cantus.uwaterloo.ca
[2] https://cantus.simssa.ca/manuscripts/

Our system is similar to the Crayons system [1], which allowed users with no machine-learning background to train pixel classifiers by iteratively marking pixels as foreground or background through brushstrokes on an image. After each user interaction, the system responded with an updated image segmentation for further review and corrective input. used in an incremental learning fashion [23]. We start by preprocessing a small number of pages (typically two or three) with a pre-existing model, usually with a model learned in pages of similar characteristics. Then, we correct the coarse errors in the output of the previous stage with a pixel-level editor. In this step, we only spend the amount of time required to correct the major errors in order to have a reasonable set of corrected data. Finally, we iterate over the two previous steps until desired performance is achieved.

In our approach for image segmentation, the output of a learning system is used by a human teacher to further inform the system about the performance of the task. As a result, we are implementing an incremental and adaptive workflow based on tactics and strategies by which human teachers modify their actions depending on the outcome of a task given to learning machines. Preliminary implementations of these pedagogical strategies and actions have permitted us to considerably reduce the amount of effort when creating ground truth for image preprocessing for OMR by 40 percent. Importantly, we have not only obtained similar performance than using ground truth created from scratch, but we have also achieved higher user satisfaction [24]. We are currently increasing the iteration rate between training, correction, and retraining to see if even better results can be obtained.

Once the Layout Analysis stage has been completed to the user's satisfaction, the system outputs a number of image files per original score image, where each file contains a layer representing a different type of musical information. For example, these layers may contain notes, staff lines, lyrics, annotations, or ornamental letters.

B. *Symbol Classification*

Our application for the second stage of the OMR workflow, Music Symbol Classification, is called Interactive Classifier (IC) (see Figure 4), which is a web-based version of the Gamera classifier [25]. In this stage, the pixels of the layer containing musical symbols, are automatically grouped into glyphs by a method based on connected component analysis. The results are displayed to the human teacher, who manually labels the classes of a number of musical glyphs. IC then extract a set of features for describing each of the glyphs and classify the data based on the k-nearest-neighbor classifier. The results are again displayed allowing the human to correct any misclassified glyphs. This process is repeated until the teacher is satisfied with the results.

The IC is purposely designed so that it can learn from scratch (*tabula rasa*); without knowing anything about the symbols it would classify. This allows IC to learn practically any set of symbols—musical or otherwise—as long as each symbol can be reasonably segmented. On the other hand, if the IC was previously used to train a class of symbols from a certain notation system (e.g., Saint Gall neume notation), the resultant model can be used as a starting point for classifying the symbols for a set of different manuscript using a similar notation system (e.g., Old Hispanic neume notation).

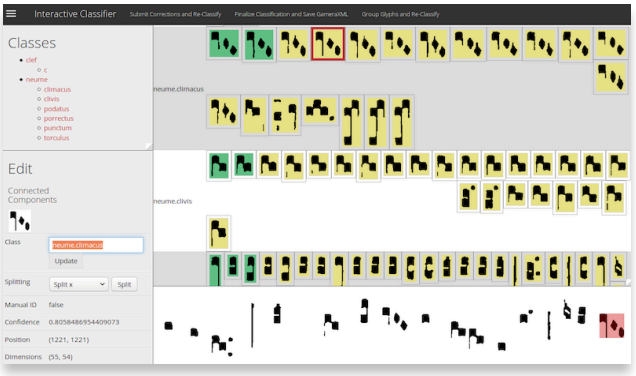

Fig. 4. Interactive Classifier (IC).

It should be noted that it is not necessary to correct all the errors made by the system. By correcting a strategic subset of the errors, the machine can dramatically increase its classification accuracy. This is what makes the IC environment pedagogically interesting: How well the machine learns depends on how well the human teaches it. In fact, the human, through interaction, can gradually learn how to teach the machine better. Furthermore, human teachers do not need to know the intricacies of machine learning or need to be a domain expert because, for humans, these are simple visual tasks.

V. Conclusions

One of the main goals of the SIMSSA project is to develop a large-scale OMR system that relies on community input for the creation of a large online music library. In this paper, we explained the project, in general, and in particular, how human users are intimately involved in teaching the system to perform well.

There are many open research questions arising for this work that need further investigation. One of the most critical is the need for user studies; not only to increase the effectiveness of the OMR process but, at the same time, improve the user experience and the community participation.

One of the known challenges is to increase the transparencies of the system so that it becomes more apparent how the teaching at an early stage, for example, at the document layout stage, affects the final transcribed output. Another challenge is to determine how different types of users, such as music scholars and amateur musicians, or people with different personalities affect the effectiveness of teaching. This idea is inspired by studies such as those by Hayes and Reik [27], who have studied the effects of the personalities of teachers in teaching robots. Finally, we need more studies to fully understand the potential of multiple end users interacting with machine learning systems [26]. As we open up our system to a wider audience, studying how people interact with it should stimulate further research.

We hope that by providing an attractive and rewarding pedagogical environment, which is both enjoyable yet productive, the system would offer a better user experience and attract more people to participate in this global effort in the music content creation.

Acknowledgment

We would like to acknowledge all the researchers and students who worked on this project and the anonymous reviewers with their valuable comments.

REFERENCES

[1] J. A. Fails and D. R. Olsen Jr., "Interactive machine learning," in *Proceedings of the 8th International Conference on Intelligent User Interfaces,* Miami, FL, pp. 39–45, 2003.

[2] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proceedings of the 17th International Conference on Machine Learning,* Stanford, CA, 2000.

[3] G. Widmer, "A tight integration of deductive and inductive learning," in *Proceedings of the 6th International Workshop on Machine Learning,* pp. 11–13, Ithaca, NY, 1989.

[4] I. Bieger, I, K. R. Thórisson, and B. R. Steunebrink, "The pedagogical pentagon: A conceptual framework for artificial pedagogy," in *International Conference on Artificial General Intelligence* (Lecture Notes in Computer Science, vol. 10414), T. Everitt, B. Goertzel, and A. Potapov, eds, Cham, Switzerland: Springer, 2017, pp. 212–222.

[5] I. Fujinaga, A. Hankinson, and L. Pugin, "Automatic score extraction with optical music recognition," in *Current Research in Systematic Musicology,* Rolf Bader, ed. Heidelberg: Springer, 2018.

[6] G. A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. S. Marcal, C. Guedes, and J. S. Cardoso, "Optical music recognition: State-of-the-art and open issues." *International Journal of Multimedia Information Retrieval* (March 2012), pp. 1–18, 2012.

[7] C. Dalitz and T. Karsten. "Using the Gamera framework for building a lute tablature recognition system," in *Proceedings of the 6th International Conference on Music Information Retrieval,* London UK, pp. 478–481, 2005.

[8] K. Helsen, J. Bain, I. Fujinaga, A. Hankinson, and D. Lacoste. "Optical music recognition and manuscript chant sources." *Early Music,* vol. 42 no. 4, pp. 555–558, 2014.

[9] H. Balk and L. Ploeger, "IMPACT: Working together to address the challenges involving mass digitization of historical printed text," *OCLC Systems & Services,* vol. 25, no. 4, pp. 233–248, 2009.

[10] D. Smith and R. Cordell, "A research agenda for historical and multilingual optical character recognition." Technical Report, deposited at Northeastern University Library's Digital Repository Service. 2018. Accessed on: March 1, 2019. [Online]. Available at: https://ocr.northeastern.edu/report/

[11] I. Fujinaga and A. Hankinson, "Single Interface for Music Score Searching and Analysis (SIMSSA)," in *Proceedings of the International Conference on Technologies for Music Notation and Representation*, Paris, France, pp. 109–115, 2015.

[12] S. Snydman, R. Sanderson, and T. Cramer, "The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images," *Archiving Conference 1,* pp. 16–21, 2015.

[13] I. Fujinaga, "Adaptive optical music recognition," Ph.D. Dissertation, McGill University, Montréal, QC, 1996.

[14] L. Pugin, J. A. Burgoyne, D. Eck, and I. Fujinaga, "Book-adaptive and book-dependent models to accelerate digitization of early music," in *Proceedings of the NIPS Workshop on Music, Brain, and Cognition,* Whistler, BC, pp. 1–8, 2007.

[15] A. Stoffel, D. Spretke, H. Kinnemann, and D. A. Keim, "Enhancing document structure analysis using visual analytics," in *Proceedings of the ACM Symposium on Applied Computing,* New York, NY, pp. 8–12, 2010.

[16] P. Roland, "The Music Encoding Initiative (MEI)," in *Proceedings of the First International IEEE Conference on Musical Applications Using XML,* Milan, Italy, pp. 55–59, 2002.

[17] L. Pugin, R. Zitellini, and P. Roland, "Verovio: A library for engraving MEI music notation into SVG," in *Proceedings of the International Society for Music Information Retrieval Conference,* Taiwan, Taipei, pp. 107–112, 2014.

[18] G. Burlet, A. Porter, A. Hankinson, and I. Fujinaga, "Neon.js: Neume Editor Online," in *Proceedings of the International Society for Music Information Retrieval Conference,* Porto, Portugal, pp. 121–126, 2012.

[19] A. Hankinson, W. Liu, L. Pugin, and I. Fujinaga, "Diva.js: A continuous document image viewing interface," *Code4lib Journal,* 2011. [Online.] Available at: http://journal.code4lib.org/articles/5418

[20] A. Hankinson, "Optical music recognition infrastructure for large-scale music document analysis," Ph.D. Dissertation, McGill University, Montréal, QC, 2015.

[21] Z. Saleh, Ke. Zhang, J. Calvo-Zaragoza, G. Vigliensoni, and I. Fujinaga. "Pixel.js: Web-based pixel classification correction platform from ground truth creation," in *Proceedings of the 12th IAPR International Workshop on Graphics Recognition*. Kyoto, Japan, vol. 2, pp. 39–40, 2017.

[22] A.-J. Gallego and J. Calvo-Zaragoza. "Staff-line removal with selectional auto-encoders." *Expert Systems with Applications,* vol. 89, pp. 138–148, 2017.

[23] R. Polikar, L. Upda, S. S. Upda, and V. Honavar, "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE Trans. on Systems, Man, and Cybernetics—Part C: Applications and Reviews,* vol. 31, pp. 497–508, November, 2001.

[24] J. Calvo-Zaragoza, K. Zhang, Z. Saleh, G. Vigliensoni, and I. Fujinaga, "Music document layout analysis through machine learning and human feedback," in *Proceedings of 12th IAPR International Workshop on Graphics Recognition.* Kyoto, Japan, vol. 2, pp. 23–24, 2017.

[25] M. Droettboom, K. MacMillan, and I. Fujinaga, "The Gamera framework for building custom recognition systems," in *Proceedings of the 2003 Symposium on Document Image Understanding Technologies,* Greenbelt, MD, pp. 275–286, 2003.

[26] E. Law, K. Z. Gajos, A. Wiggins, M. L. Gray, and A. Williams, "Crowdsourcing as a tool for research: Implications of uncertainty," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing,* Portland, OR, pp. 1544–1561, 2017.

[27] C. Hayes and L.D. Riek, "Robot errors and human teachers: The effects of personality and patience during learning tasks," *RSS Workshop on Negative Results in Robotics*, Freiburg im Breisgau, Germany, 2015.