Scaling Handwritten Student Assessments with a Document Image Workflow System

Vijay Rowtula, Varun Bhargavan, Mohan Kumar and C.V. Jawahar
IIIT Hyderabad
vijay.rowtula@research.iiit.ac.in, varun@eml.cc, mohankumaar3@gmail.com, jawahar@iiit.ac.in

Abstract

With the increase in the number of students enrolled in the university system, regular assessment of student performance has become challenging. This is specially true in case of summative assessments, where one expects the student to write down an answer on paper, rather than selecting a correct answer from multiple choices.

In this paper, we present a document image workflow system that helps in scaling the handwritten student assessments in a typical university setting. We argue that this improves the efficiency since the book keeping time as well as physical paper movement is minimized. An electronic workflow can make the anonymization easy, alleviating the fear of biases in many cases. Also, parallel and distributed assessment by multiple instructors is straightforward in an electronic workflow system. At the heart of our solution, we have (i) a distributed image capture module with a mobile phone (ii) image processing algorithms that improve the quality and readability (iii) image annotation module that process the evaluations/feedbacks as a separate layer.

Our system also acts as a platform for modern image analysis which can be adapted to the domain of student assessments. This include (i) Handwriting recognition and word spotting [5] (ii) Measure of document similarity [6] (iii) Aesthetic analysis of handwriting [7] (iv) Identity of the writer [4] etc. With the handwriting assessment workflow system, all these recent advances in computer vision can become practical and applicable in evaluating student assessments.

1. Introduction and Related Work

Regular, personal feedbacks are critical to learning. Traditionally, this has been achieved through qualitative/quantitative assessments and through home works. We also had strong tradition of using handwritten assessments that often reflect the student thinking process beyond the final answer. Over time, electronically created and formatted documents have crept into the system which limited the effectiveness of assessment. Managing student assessments consume a significant portion of the effort of a teacher. With the need to scale, modern assessment systems are slowly moving towards solutions that can automate the evaluation process. Examples include multiple choice questions, fill in the blanks, matching two sets and output based computer program evaluation. Personal touch of the assessment process is also disappearing with the penetration of Internet and electronic solutions. We now have a contradicting requirement of scalability and effectiveness.

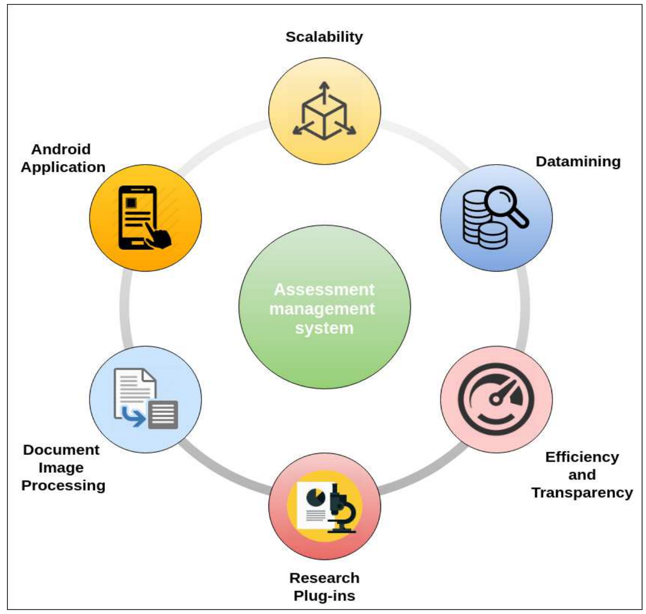

Figure 1. The image highlights some of the key aspects we focused on, to develop our document workflow system. These are the differentiating factors which sets apart our system from existing learning management systems.

This paper makes a contribution in assessment space with a document image workflow system that can bring the advantages of the electronic workflow into the world of physical paper.

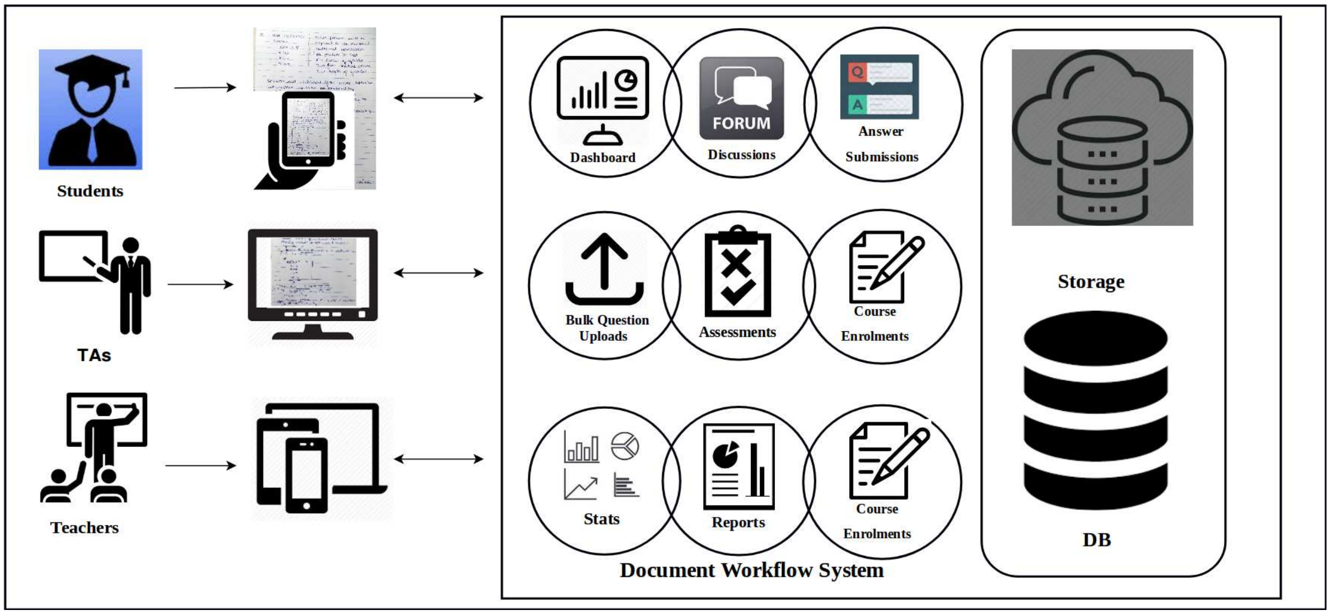

Figure 2. Conceptual Diagram: students, teaching assistants and faculty members have their respective interfaces and roles. Student scans a paper based solution using a dedicated mobile application. Teaching assistants and faculty members have mobile interface where they could annotate the answer sheet with qualitative and quantitative feedback.

With the advent of Web 2.0 and MOOCs, e-learning platforms have gained popularity and have made a profound impact in the field of education. Current virtual learning environments, also called as learning management systems(LMS) typically provide tools for assessment, communication, uploading of content, administration of students, questionnaires, tracking tools, wikis, blogs, chats, forums, etc. over Internet. But they have few drawbacks. Handwritten assessments have been a powerful format to create and evaluate students. It shows the organization of thoughts, original expressions in comparison to the electronically formatted solutions that does not show the fingerprints of a student. It is observed that for handwritten assessments, students do not receive any detailed feedback quickly for it to be helpful enough in their next assessment, because of the time delay involved in distribution, evaluation, entry of grades etc.

In this paper, we present a system that supports several assessment formats with special emphasis on handwritten assessments. The system also provides plug-in support for enhancements to integrate or update further innovations in student assessment space. A conceptual explanation of the system is shown in Figure 2. Students digitize the handwritten document with a mobile phone based interface. Instructors can grade/assess by annotating the images online. This simple yet effective connect between the physical paper world and electronic workflow makes our solution effective and efficient.

Document Image Work Flow Systems: Document images are images with rich textual content. Even in today's world, a large number of documents are generated as handwritten documents. This is specially true when the document/knowledge/expertize is captured conveniently with availability of electronic gadgets. Information extraction from handwritten medical records [9] written in ambulance for doctor's interpretation in hospital, reading postal address [12] to automate the letter sorting are examples where document image work flow helped in scaling the system with minimal human intervention. In such work flow systems, images flow across subjects who can be in different locations. A postal automation module in USA can take help of a person in Asia to recognize the address block and still continue to be efficient. Our work is motivated with the success of these document image workflow systems that were put into practice when the handwriting recognition accuracy was unacceptably low.

The focus of this paper is to demonstrate a scalable paperless grading system for handwritten assessments which allows electronic submission and on-screen grading of the assessments with high transparency between instructors and students. In Section 2, we introduce our document workflow system, its image processing modules and provide a brief overview of system architecture. In Section 3, we describe our experience using the workflow system. We also explain how the recent advances in handwritten document analysis will be integrated into our workflow system, opening up new avenues in research which can impact education.

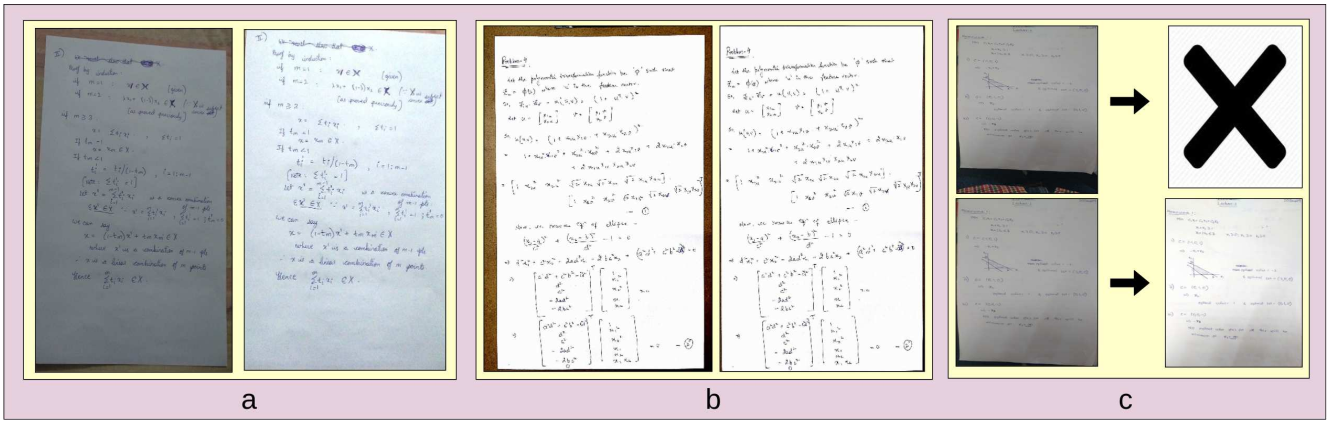

Figure 3. Example of original and processed handwritten assessments before being sent for evaluation. Sets (a), (b) contain pre and post processed handwritten assessments. Set (c) shows assessment rejection due to inconsistencies(top) and a better image was uploaded by student and was processed(bottom). We can notice the border, color and brightness rectification in all three image sets.

2. Assessment Management System

We now start by looking at what can happen in a typical classroom scenario. Faculty member provide questions and students bring their solutions to classroom or submit them at a fixed location. A teaching assistant assigned by the faculty member or the faculty member herself (instructors) grade the assessments and provide quantitative and/or qualitative feedback. Finally the grades are available to students, after a brief discussion between students and instructors about evaluation corrections. In the following sections, we explain how our solution was designed to troubleshoot the pain points faced by instructors and students during the workflow process.

2.1. Design Goals

We started with the following set of goals:

- Make the overall student assessment process efficient by removing paper movement, paper arrangements (eg. sorting pile of papers by student IDs) and additional data entry (manual entry of scores into a database explicitly).
- Bring correction/evaluation electronically as an extra annotation layer. This should enable parallel, distributed and multiple grading of the same student assessment.
- Incorporate a set of computer vision methods required to meet the immediate goal and keep the design open to introduce advanced image recognition modules at a later stage.
- A system that can learn, improve and adapt over time. For example, common errors/feedbacks are mined from the annotations and displayed on novel situations, thus minimizing the effort.

2.2. Document Image Processing

In our assessment evaluation process, student first uploads camera-captured document images using an android application (discussed in section 2.4). It is a known fact that camera-captured images are prone to various degradations such as inadequate lighting, shadows, blur and camera flash at times. Such degradations often lead to difficulties in analysis at subsequent stages of image processing. For example, degradations may result in a significant drop in the performance of Optical Handwriting Recognition (OHR), word spotting and other handwritten document analysis tasks, resulting in unrecoverable information loss.

The degradations introduced can be classified into *(i)Character level - with broken characters, touching, skewed or curved handwriting, (ii)Page level - margin noise, salt-and-pepper, ruled line, warping, curling, skew, blur or translation*. We focused on rectifying page level degradations.

Capturing handwritten assessments: Though the students in traditional learning management systems have the comfort of submitting the handwritten assessments from any location, the assessments still have to be compressed (zipped) and uploaded to a server. Instructors will have to download the file and then evaluate the handwritten or other file based assessments. For handwritten assessments, our workflow solution includes an android application which is used by students to take pictures of the assessments and upload them to server immediately. This can be very helpful in scenarios such as a surprise or spot assessment in class room. The android application tries to qualify the images based on the visual aesthetics of the uploaded handwritten document image. We used methods described in [7, 14], which uses a set of local character level features and global page level features to arrive at a quality score. The android application will reject the images with lower than a permissible score on distortions as seen in Figure 3. In such cases, student has to re-upload a proper image of his handwritten assessment. Legible images are finally uploaded to server with the consent of student.

Dewraping camera-captured images: Compared to scanners, mobile cameras offer convenient, flexible, portable, and non-contact image capture, which enable better throughput in a document workflow management system. However, camera-captured documents may also suffer from distortions caused by non-planar document shape and perspective projection, which can lead to failure of current OCR/OHR technologies. The images were rectified based on the method explained in [11]. These methods share a similar hierarchical problem decomposition: (i) Split the text into lines. (ii) Find a warp or coordinate transformation that makes the lines parallel and horizontal. Though the cited methods were modeled for printed text, we observed that same methods worked well for camera-captured handwritten document images. A sample of dewrapped images can be seen in Figure 4.

Rule line removal from handwritten assessments: Some of the students submit there assessments in rule lined pages, as shown in Figure 4. Rule lines - both horizontal and vertical, should be removed to ensure better analysis at subsequent stages of image processing. We adapted methods described in [2] which uses rule line detection using Horizontal Projection Profile (HPP) and Hough Lines (HL). The steps involved are: (i) De-skew the image using method described in earlier section (ii) Extract the location of horizontal lines using combination of HPP and HL (iii) Remove the lines from the deskewed version of original document image and (iv) Reconstitute the missing pixels. Image (b) in Figure 4 shows original camera-captured document image and its rectified version.

Annotation of images: Our solution allows on-screen evaluation of uploaded handwritten assessments. The instructor can highlight, annotate and comment on document images. These annotations are saved separately along with its image coordinate details. Since these annotations are immediately available to the students, they can immediately start a discussion with the instructors. The keywords from questions, assessment image and discussions together form a rich set of evaluation annotations for an assessment platform, which can be mined for patterns and reused while evaluating a similar assessment of other students.

Though these are experimental features, they demonstrate the extensibility of our document workflow platform in handwritten assessment space.

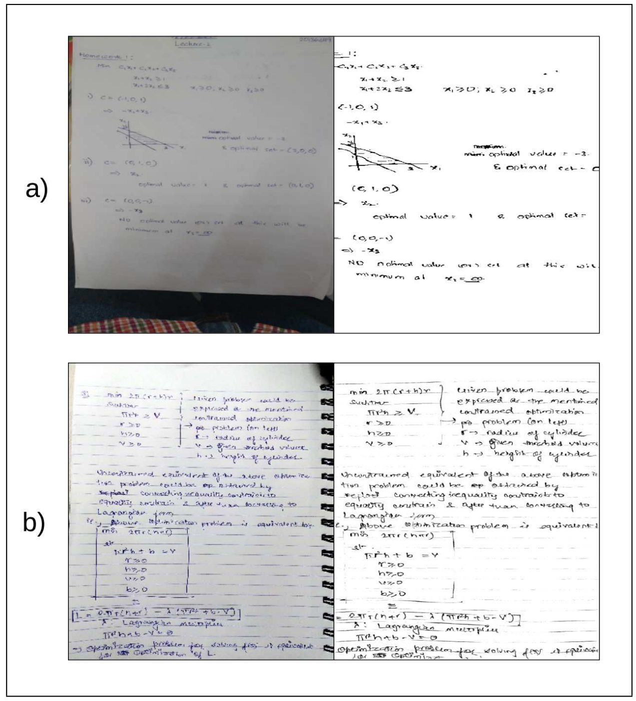

Figure 4. Sample Document Images rectified using Image processing. First row (a) has original image and dewarped document image free from distortions (shadows and bends). The second row (b) has original image with rule lines / bad illumination and dewarped document image free from distortions.

2.3. Other Features

Easing assessments: Our system design is focused on the task of decreasing the execution time of student assessments. From the creation of questions to final grading by instructors, our workflow system simplifies the complete process, by moving most of the manual procedures to web application. Students can either upload the handwritten answers using a mobile android application (Figure 5) or upload an answer file using web interface or even directly type in the answer. For code evaluations, students can upload the code to the portal and evaluation is completed online, as explained in section 2.4. Text and image based answers are evaluated on-screen using our portal.

Data mining in e-learning: The application of data mining in e-learning systems is an iterative cycle. The mined knowledge should enter the loop of the system and enhance learning as a whole, and facilitate filtering of mined knowledge for decision making. Our solution uses simple data analysis to observe student's behavior and assist instructors in detecting possible shortcomings to incorporate improvements. It mines the data and creates report on student assessment submission delays, highly performing and under performing teaching assistants, forums harboring negative

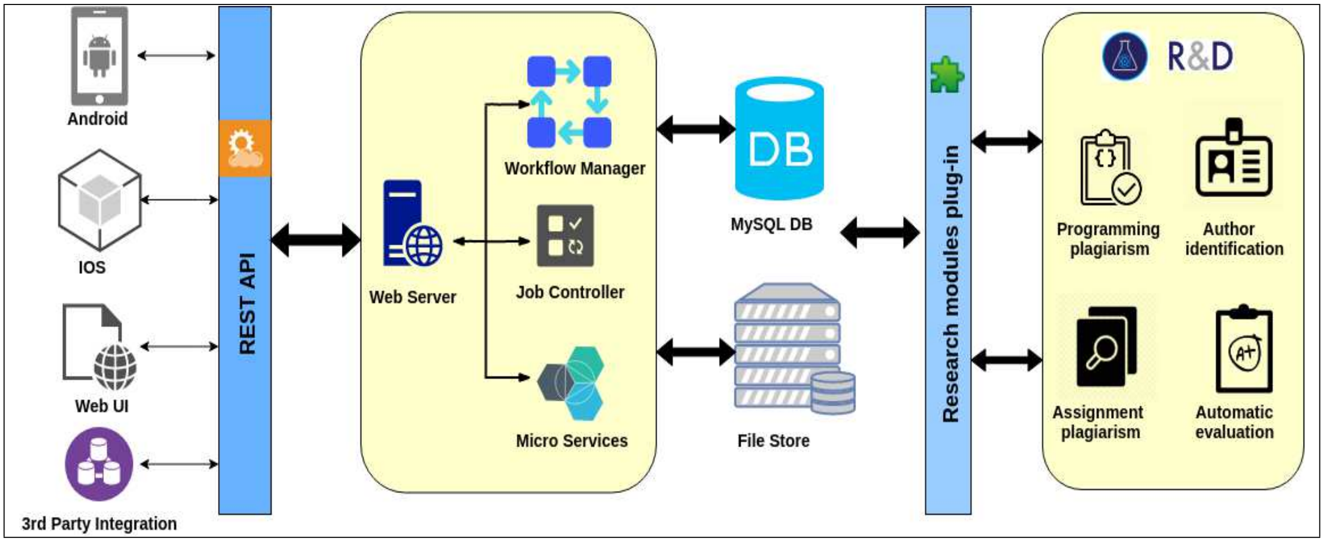

Figure 5. System architecture and workflow of Assessment Management System.

discussions and other similar vital stats. A weekly status update by email is sent to both students and instructors with consolidated stats.

Thus, the system helps in identifying the achievement gaps among students and tutors alike, measures the effectiveness of a course, academic program or learning experience over the course duration.

System transparency: This is implemented by processes such as double blind assessments, peer review of evaluations, discussion forums, dashboards by profile hierarchies and weekly status updates by email. The double blind procedure makes sure of unbiased evaluations and discussion between students and instructors. The queries and discussions on evaluations can be monitored down the work flow hierarchy. Based on roles, the login page has dashboard which summarizes important updates to students and instructors. The performance of students and TAs are mined from databases which are available on teacher dashboards, hence promoting transparency throughout the workflow.

2.4. System Architecture and Implementation

The Assessment Management System architecture was designed with modularity, scalability and extensibility in mind. Fig 5 describes the software architecture of the system and shows the modules therein. Some of the key aspects are discussed next.

Scalability: The ease of use for assessments, described in sec 2.1 brings up a new challenge - scalability. Platform is massively scalable due to use of open source technologies such as Django, MySQL and Docker[8]. It is scalable in terms of hosting number of courses, enrolling and managing large number of students, assessments etc. Currently, more than 15 courses were hosted on our document workflow system, with students count varying from 30 to 150 per course. Even the possible bottlenecks for automated code evaluations are addressed using docker containers. A docker container is a virtual sandbox to create and manage resource per user. Pre-defined resources are allocated per user using docker, hence avoiding system downtime due to possible hacking or resource consumption beyond permissible limits. Another possible bottleneck is handwritten assessment evaluations. This is addressed by on-screen evaluation provided by an intuitive user interface to navigate through assessments.

Mobile application: An android application was designed to work with REST API, which also supports assisted image capture and image corrections. This application supports submission of hand-written answers, by allowing the capture of the hand-written document using the camera of the mobile device. This android application is currently being extended for touch screen devices to speed-up assisted evaluation as explained in section 3.2.

Code evaluation module: Code evaluation module supports automated evaluation of programing assessments. It supports accepting source code/code snippets as answer submissions and evaluation of those submissions in secure and contained environments. It uses various sandbox and container technologies to run these codes in a safe environment and supports popular programming languages like C, C++, Python, Java, etc. Instructors can customize evaluations by adding custom code snippets during creation of programming questions.

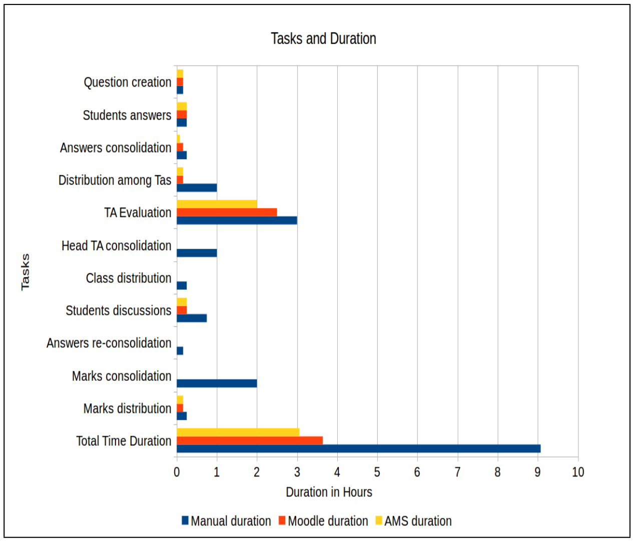

Figure 6. Graph shows the effectiveness of our document workflow system compared to manual and Moodle based student evaluations in handwritten assessments space

Research plug-in: Various top research papers in handwritten and programming assessment space are evaluated and converted into research modules. These modules are first evaluated on smaller test sets and are finally plugged into the system. We have focused specifically on handwriting and programming space to assist the evaluators dealing with courses containing handwriting and programming assessments. Various in-house research projects are also integrated into the system. The research modules are discussed in detail in section 3.2.

Peer review module: Our document workflow system can support peer review of answer submissions to enhance or replace evaluation by a dedicated evaluator. The anonymity which this system can provide increases the reliability of the peer-review process as a whole. The time required for distribution and collection of the submissions, which makes up the bulk of the time wasted during a regular peer-review process, is saved by using such a system. This makes peer-review a feasible option even for assessment evaluation.

Third party integrations: Our document workflow system provides a set of robust REST APIs (web-services) that provides an easy usability and extensibility of the platform. The APIs can be used to integrate our application to any third party systems and websites. The advantages of such an integration is many-fold. It is possible to use the research modules of our system in 3rd party applications and websites. Any on-line teaching platform will be able to integrate our document workflow system, as an extension to manage their assessments.

3. Experience and Discussions

3.1. Experiences

We tested our document workflow system in the real world for 15 university courses. A total of 101 assessments were posed on the platform so far. The assessments contain 607 questions out of which 540 are handwriting based. The total number of student answers is 29300 out of which, the total number of handwritten answers is 20200. We receive feedback from the tutors and students after every course for improvements. The feedback is based on the 6 different aspects of usage of the document workflow system - *interactivity, tutor support, peer support, user-friendly, time management and insights*. Student can also report bugs and enhancement requests. The feedback so far indicated that all students experienced an optimal learning environment and most often suggested improvements in peer-support and interactivity.

Class room experiment: We have also conducted an experiment to validate the effectiveness of usage of our workflow system for handwritten assessments. As described in Table 1, a set of three questions from *Optimization Methods* course was provided to a class of 127 students with 4 teaching assistants. Students were divided into 3 groups to submit the assessment answers using three channels - manual(paper based), Moodle and our workflow system. We collected stats (time duration in hours) for each task from - assessment creation to marks distribution back to students for all three mentioned channels. The tasks are described below:

- Question creation: Time taken to create assessment question.
- Student answers: Average time taken to answer all assessment questions.
- Answers collection: Approximate time taken to collect the student answers.
- Distribution among TAs: Approximate time taken to distribute student answers among TAs.
- TA Evaluation: Average time taken by TAs to evaluate student answers.

Class Room Experiment	count
No. of students	127
No. of questions	3
No. of instructors	4
Total answers	381

Table 1. Controlled Class Room Experiment details.

- Head TA consolidation: Time taken by Head TA to consolidate student answers from other TAs.
- Class distribution: Time taken by TAs to distribute evaluated student answers back to students.
- Students discussion: Time taken for evaluation discussion among TAs and students.
- Answers re-consolidation: Time taken by TAs consolidate student answers again after evaluation discussions.
- Marks consolidation: Time taken by Head TA to consolidate student marks in spread sheet or a system.
- Marks distribution: Average time taken by TAs to distribute marks to students.
- Total time duration: The total time taken to complete above mentioned 11 tasks sequentially.

Figure 6 shows a graph with time duration in hours for each of the task mentioned above, for channels - manual submission, Moodle submission and submission through our document workflow system. The graph shows (i) duration for each task - which is average time taken per task for all three channels of submission and (ii) total time duration - is the total time taken to assess students using three mentioned channels. We observed that, in general our document workflow system saves time for most tasks as shown in the Figure 6. Our document workflow system also saves considerable time (average assessment time for class) when compared to manual handwritten paper based assessments. This is because few tasks can be skipped while using online assessments. As seen in Figure 6, the system also outperforms Moodle due to ease of use through mobile upload of assessments.

3.2. Discussion - Emerging Research Problems

Handwriting plagiarism: Most universities use online plagiarism detection software to root out Internet plagiarism. The problem of predicting the similarity between two handwritten document images has already been addressed here [6, 13]. Though this is not a completely solved problem, we are trying to find better ways to enhance the ability to detect plagiarism among students. Our preliminary observations indicate that simple word spotting techniques does not suffice and we also need semantic techniques on handwritten text to solve the problem (Figure 7).

Author identification handwritten text: This is to identify documents containing more than one document signature style. A student typically spends several years in college. Hence a single document from student can used as unique fingerprint/signature to identify his handwriting across semesters. Our current module developed using method described in [4] is able to identify the students with decent accuracy but is not perfect. Better and faster methods are required to enhance both accuracy and speed when comparing across thousands of students on college premises.

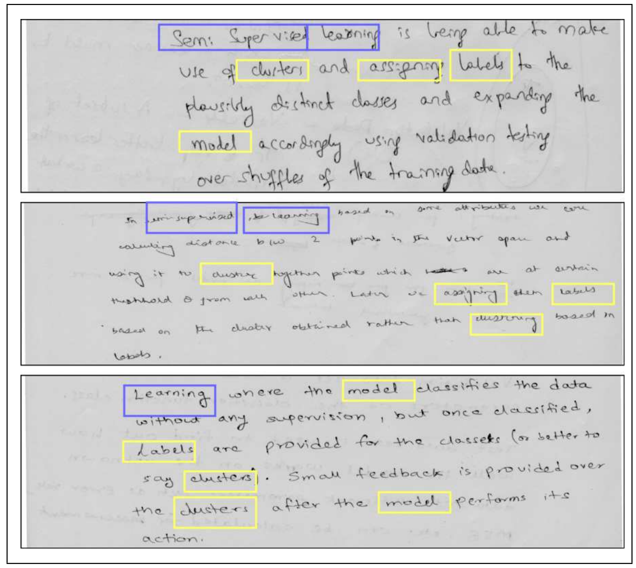

Figure 7. Sample Hand written ML assessment analyzed for plagiarism. Blue and yellow bounding boxes show common and important words using which, a plagiarism score is calculated.

Code plagiarism: Plagiarism is a statement that someone copied code deliberately without attribution While MOSS [1] automatically detects program similarity, it has no way of knowing why codes are similar. Systems like MOSS also use web-services for code comparison which makes them even more slow. It is still up to a human to go and look at the parts of the code that MOSS highlights and make a decision about whether there is plagiarism or not. Though we have integrated a custom code analyzer which uses sequence based models, it is limited to C language and better models are required to scale to large number of students.

Evaluation of handwritten assessments: The typical engineering homework assessment may involve sketches, formulas with special symbols, as well as calculation steps. The most time efficient way for students to do this work is by hand, on paper. The handwritten assessment of student will be available for further evaluation by instructors, either using on screen evaluation tools or semi/auto evaluation methods which are still research problems as explained below.

Semi-automated evaluation: In a university setting, tutors are required to evaluate several students and thousands of answers at a time. This can be cumbersome and any assistance provided to the instructors which can increase the throughput of evaluations will be a value-add. Clustering based assessment techniques are available for text based assessments [3]. The method first trains a model on similarity metric between student responses, but then go on to use this metric to group responses into clusters and sub clusters. A similar method can be implemented for handwritten evaluations where segmented words can be clustered based on semantic similarity between students response and reference answer given by the instructor. Student responses can be queued from the clusters based on the similarity metric which can increase the throughput of evaluations. We call this semi-automated evaluation of handwritten assessments. Our method can currently detect key phrases in the assessment.

Fully Automated evaluation: Automated evaluation of handwritten assessments can be seen as an extension to the above mentioned method, where assistance was restricted to clustering answers, queuing them and highlighting the keywords in assessments. This can be further enhanced provided that the reference answer is available. A regression model can be trained on a set of semantic word features [10] in visual space, which can predict an evaluation score similar to that of an instructor. The score may not be necessarily accurate but we feel that a nearest score with a confidence metric can boost the throughput of evaluations enormously. We are currently testing the efficiency of this method and it is yet to be integrated into the our document workflow system.

4. Conclusion

Handwriting recognition has not reached a state that can directly help with the scalability of automated evaluations. However, we argue that our work flow system can enhance the efficiency and quality of the assessments without the need of OHR. Our system presented in this paper addresses the need for a tool to computerize the existing handwritten assessments at all levels of our education system. Through this paper we tried to showcase the capabilities of our document workflow system. To summarize, it has useful set of tools which encompass existing technologies for text, code and handwritten assessments, which can enhance the tutors and students experience alike by minimizing the time required for the whole assessment management process. Though the process is not yet perfect, the platform is open for future enhancements not only in text and handwritten work space but also in integrating research output from audio and video space.

References

[1] A. Aiken. Measure of software similarity. *URL http://www. cs. berkeley. edu/-aiken/moss. html*, 1994.

[2] K. Arvind, J. Kumar, and A. Ramakrishnan. Line removal and restoration of handwritten strokes. In *Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on*, volume 3, pages 208–214. IEEE, 2007.

[3] S. Basu, C. Jacobs, and L. Vanderwende. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402, 2013.

[4] D. Bertolini, L. S. Oliveira, E. Justino, and R. Sabourin. Texture-based descriptors for writer identification and verification. *Expert Systems with Applications*, 40(6):2069–2080, 2013.

[5] P. Krishnan, K. Dutta, and C. Jawahar. Word spotting and recognition using deep embedding. In *DAS*, 2018.

[6] P. Krishnan and C. Jawahar. Matching handwritten document images. In *European Conference on Computer Vision*, pages 766–782. Springer, 2016.

[7] A. Majumdar, P. Krishnan, and C. Jawahar. Visual aesthetic analysis for handwritten document images. In *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*, pages 423–428. IEEE, 2016.

[8] D. Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2, 2014.

[9] R. Milewski and V. Govindaraju. Handwriting analysis of pre-hospital care reports. In *CBMS*, 2004.

[10] M. Mohler, R. Bunescu, and R. Mihalcea. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 752–762. Association for Computational Linguistics, 2011.

[11] F. Shafait and T. M. Breuel. Document image dewarping contest. In *2nd Int. Workshop on Camera-Based Document Analysis and Recognition, Curitiba, Brazil*, pages 181–188, 2007.

[12] S. N. Srihari. Handwritten address interpretation: a task of many pattern recognition problems. *International journal of pattern recognition and artificial intelligence*, 14(05):663–674, 2000.

[13] S. Sudholt and G. A. Fink. Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*, pages 277–282. IEEE, 2016.

[14] P. Ye, J. Kumar, L. Kang, and D. Doermann. Real-time no-reference image quality assessment based on filter learning. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 987–994. IEEE, 2013.