Challenges for the Engineering Drawing Lehigh Steel Collection

Bart P Lamiroy, Daniel P Lopresti

Challenges for the Engineering Drawing Lehigh Steel Collection

Bart Lamiroy
Université de Lorraine – LORIA (UMR 7503)
Campus Scientifique – BP 239
54506 Vandœuvre-lès-Nancy CEDEX – FRANCE
Email: bart.lamiroy@loria.fr

Daniel P. Lopresti
Lehigh University – CSE Department
19 Memorial Drive West
Bethlehem, PA 18015 – USA
Email: lopresti@cse.lehigh.edu

***Abstract*—The Lehigh Steel Collection (LSC) is an extremely large, heterogeneous set of documents dating from the 1960's through the 1990's. It was retrieved by Lehigh University after it acquired research facilities from Bethlehem Steel, a now-bankrupt company that was once the second-largest steel producer and the largest shipbuilder in the United States. The documents account for and describe research and development activities that were conducted on site, and consist of a very wide range of technical documentation, handwritten notes and memos, annotated printed documents, etc. This paper addresses only a sub-part of this collection: the approximately 4000 engineering drawings and blueprints that were retrieved.**

The challenge resides essentially in the fact that these documents come in different sizes and shapes, in a wide variety of conservation and degradation stages, and more importantly in bulk, and without ground-truth. Making them available to the research community through digitization is one step the good direction, the question now is what to do with them. This paper tries to lay down some first basic stepping stones for enhancing the documents' meta-data and annotations.

I. INTRODUCTION

The Lehigh Steel Collection (LSC) was first presented at the Tenth IAPR International Workshop on Graphics RECognition (GREC 2013) held at Lehigh University in 2013. At that time the entire collection was still within the abandoned buildings the University just acquired. Subsequent retrieval and some initial curation resulted in approximately 4000 large engineering drawings to be safeguarded, as well as an currently unaccounted number of tens of thousands of office documents [1].

This paper focuses on the technical drawings. These drawings are essentially around the size of the ISO 216 A0 (841 mm × 1,189 mm or 33.1 in × 46.8 in) [4] but still present significant changes in size. Their physical support also varies between plain paper, tracing paper and acetate film. They underwent various level of degradation due to ageing, exposure to sunlight, heat, humidity and moisture, mishandling, tearing and folding, *etc*. Also, some are professionally hand drawn original ink or pencil drawings, others have been mechanically plotted or were photocopied, a few rare blueprints are also present. Samples of the documents are available through the DAE platform[1].

[1] http://dae.cse.lehigh.edu/DAE/?q=browse/dataitem/605893

There is no further metadata available for these documents. In other terms, there is no "ground truth" associated with these drawings and any attempt to use them in a benchmarking or contest effort should inevitably consider either the establishing of the ground truth annotation (*e.g.* using human annotators) or as an alternative, integrate the absence of ground truth in its evaluation protocol [5], [6].

The following sections outline a series of challenges related to the LSC. Section II addresses the topics that need to be addressed before the collection can be used for benchmarking in a more conventional way. Section III then lists a series of possible tasks that can be considered on the documents and that may provide useful insight for the graphics recognition community. Sections IV address the question on how to assess subsequent contest or benchmarking results, given the current and foreseeable absence of ground truth.

II. ESSENTIAL TASKS

Given the fact that the physical collection has not been curated, one has to consider that the documents are in an unordered bulk condition, or at least, that no reliable assumption of their relationship can be made on the mere proximity of two documents in the whole pile.

The documents do fall in different classes, however: they can be construction plans, mechanical assembly plans, electric wiring drawings or 3D representations. Documents also belong to specific setups or identified projects. In some rare cases, duplicate versions of the same document occur in the collection, or various draft versions and the subsequent final version may be present.

A. Scalability of State-of-the-Art Segmentation and Detection Approaches

Given the fact that we are not aware of recent documented graphical document image analysis methods that were explicitly applied on very large documents, one of the first assessments we need to make is that traditional graphic segmentation and detection approaches sufficiently well scale to the LSC proportions. Standard, comparable benchmarking datasets (as those listed below, for instance) usually contain only a few instances of significantly smaller documents.

For instance, the floor plan database reported in [10] contains 42 floor plans of approximately 2000×2000 pixels

Fig. 1. LSC Engineering Drawing Collection *in situ* ... bulk, unordered, of various types and sizes. Each pile contains approx. 2000 drawings.

in bitonal format. The dataset described in [7], contains, in its extended version [2] only 90 documents. [3][2] contains 122 greylevel images ranging from 2000×2000 pixels to 6000×7000 pixels.

The GREC 2011 Symbol Recognition Contest public dataset [13][3] contains 40 bitonal images of various sizes, ranging from 2000×2000 pixels to 5000×3000 pixels separated in two classes: electrical diagrams and floor plans. These are then further duplicated with 3 different levels of artificial noise.

All above cited datasets share a single characteristic: they are based on synthetic data. This allows them to provide accurate ground truth, and have controlled image degradation conditions. On the other hand, the proposed degradations are not representative of the ones observed in the LSC.

Challenge 1: Assess Scalability

> The first challenge to meet is to assess how the existing state-of-the-art standard binarization, segmentation and vectorization algorithms scale to the 5000×8000 pixel color images in the LSC.
> This challenge can consist of comparing specific implementations of identified standard reference algorithms, specifically optimised for low memory footprint and fast execution times.

The following challenges in this section aim to provide an environment to help create initial ground-truth hints and tools for rapidly classifying a large quantities of bulk image data.

B. Duplicate Detection

Some instances of the documents are duplicate drawings. These duplicates can be of different sorts. They can be simple copies on the same physical support. These copies can be photocopied pages, of re-executed drawings. Sometimes they can be of significantly different visual aspect (*e.g.* pen drawing and blueprints of the same document).

[2]http://dag.cvc.uab.es/resources/floorplans

[3]http://iapr-tc10.univ-lr.fr/index.php/contest/previous/symbol-segmentation/2011?id=162

Challenge 2: Detect Duplicates

> Identify duplicate drawings, notwithstanding possible significant physical differences of the drawing support and/or technique.

C. Style Classification

The drawings have a wide variety of execution styles. Some are professionally hand drawn engineering drawings, others are machine plotted or printed. In order for subsequent higher level analysis like text-graphics separation, or the detection of dimension markings to be effective, it is generally useful to have a prior knowledge of drawing style [9], [11].

Challenge 3: Drawing Style Classification

> Classify documents according to their drawing style such that various categories of drawing (machine plotted or printed *vs.* professionally hand drawn *vs.* sketch, ink *vs.* pencil, individual writing styles ...) can be made explicit.

D. Topic Classification

A quick visual analysis of the documents has shown that at least four major topics are covered by the drawings: electrical diagrams, mechanical diagrams, construction diagrams, and 3D representations.

Challenge 4: Drawing Topic/Content Classification

> Classify documents according to their generic topic such that various diagram categories (electrical, mechanical, construction/building, 3D representations ...) can be exhibited.

III. Other Possible Tasks

Once the previously described challenges have been addressed, the LSC will contain sufficient meta-data to try and focus on higher level analyses that can extract structure from the bulk collection. The following possible tasks are ranked in order of probable increased difficulty and complexity.

A. Title Block Detection and Analysis

Most (not all) of the documents contain a standardised title block referencing information related to the drawing (its title, the project it belongs to, an unique identifying label, date, author ...).

Challenge 5: Title Block Analysis

> Robust detection of the title block its subsequent analysis and interpretation are interesting challenges by themselves [8]. They can also significantly contribute to the annotation of the whole collection by contributing to attributing drawings to projects and authors (and thus possibly cross-validating results from Challenges II-C and II-D).

The results from this task can also aid in reordering and sorting the collection, by grouping them by sequence number, or by project.

B. Inter- and Intra Document Link Detection

Technical drawings have been long investigated by the Graphics Recognition community. Conventional processing consists of *Text/Graphics Separation* [12], *Segmentation of Dimension Marks and Lines* [16] and *Parts Detection*.

Challenge 6: Link Detection

> Building upon traditional graphics recognition and segmentation algorithms, extend the concepts developed in [14] to apply in the global, uncontrolled framework defined by the LSC, exhibiting relations within documents or between documents, based on recurring shapes, indexes, reference numbers, *etc.*

C. Marking and Annotation Identification

Another interesting task is to identify and segment manual annotations on the drawings, as well as color highlights.

IV. Contest Assessment and Benchmarking

Notwithstanding the wide range of possible challenging tasks to be addressed on the LSC, the main hurdle remains that there is no known ground truth to the documents. Therefore no immediate benchmarking or performance analysis can be expected to spawn from tackling any of the above challenges.

It seems counter-productive and a waste of precious time and resources to try and manually annotate all of the possible data embedded in this document collection.

Targeted, cheap and continuous human annotation or evaluation of algorithms can be achieved efficiently though a Games-with-a-Purpose crowd-sourcing approach [15]. This is a large scale approach comparable to Re-Captchas. Only the task is encoded in an addictive multi-user game.

Challenge 7: Serious Games

> The main challenge in conceiving a GWAP is to come up with a game play concept that is appealing to the players but efficient for the underlying annotation and document analysis goals.
> This requires both game design and graphic design and document image analysis knowledge.

V. Conclusion and Perspectives

As a conclusion, the best approach to the optimal use of the LSC is to just do things, and sharing them in an as open and as public possible way, such that each of the above challenges can be met by integrating or building upon tools and techniques used for other challenges.

References

[1] Barri Bruno and Daniel Lopresti. The lehigh steel collection: a new open dataset for document recognition research. In *IS&T/SPIE Electronic Imaging*, volume 9021. International Society for Optics and Photonics, 2013.

[2] L.-P. de las Heras, J. Mas, G. Sanchez, and E. Valveny. Wall patch-based segmentation in architectural floorplans. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 1270–1274, Sept 2011.

[3] Lluís-Pere de las Heras, Oriol Terrades Ramos, Sergi Robles, and Gemma Sánchez. Cvc-fp and sgt: a new database for structural floor plan analysis and its groundtruthing tool. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(1):15–30, 2015.

[4] ISO/TC 6. Writing paper and certain classes of printed matter – Trimmed sizes – A and B series, and indication of machine direction. ISO ISO-216:2017, International Organization for Standardization, 2007.

[5] Bart Lamiroy and Tao Sun. Computing Precision and Recall with Missing or Uncertain Ground Truth. In Young-Bin Kwon and Jean-Marc Ogier, editors, *Graphics Recognition. New Trends and Challenges. 9th International Workshop, GREC 2011, Seoul, Korea, September 15-16, 2011, Revised Selected Papers*, volume 7423 of *Lecture Notes in Computer Science*, pages 149–162. Springer, February 2013.

[6] Lamiroy, Bart and Pierrot, Pascal. Statistical Performance Metrics for Use with Imprecise Ground-Truth. In *Eleventh IAPR International Workshop on Graphics Recognition*, August 2015. to be presented.

[7] Sébastien Macé, Hervé Locteau, Ernest Valveny, and Salvatore Tabbone. A System to Detect Rooms in Architectural Floor Plan Images. In *IAPR International Workshop on Document Analysis Systems - DAS 2010*, ACM International Conference Proceedings Series, pages 167–174, Boston, MA, United States, June 2010. ACM.

[8] Laurent Najman, Olivier Gibot, and Stéphane Berche. Indexing technical drawings using title block structure recognition. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 587–591. IEEE, 2001.

[9] Rudolf Pareti and Nicole Vincent. Global discrimination of graphic styles. In Wenyin Liu and Josep Lladós, editors, *Graphics Recognition. Ten Years Review and Future Perspectives*, volume 3926 of *Lecture Notes in Computer Science*, pages 120–130. Springer Berlin Heidelberg, 2006.

[10] Marçal Rusiñol, Agnés Borràs, and Josep Lladós. Relational indexing of vectorial primitives for symbol spotting in line-drawing images. *Pattern Recognition Letters*, 31(3):188 – 201, 2010.

[11] P. Sarkar and G. Nagy. Style consistent classification of isogenous patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(1):88–98, Jan 2005.

[12] Karl Tombre, Salvatore Tabbone, Loïc Pélissier, Bart Lamiroy, and Philippe Dosch. Text/Graphics Separation Revisited. In J. Hu D. Lopresti and R. Kashi, editors, *5th International Workshop on Document Analysis - DAS'02*, volume 2423 of *Lecture Notes in Computer Science*, pages 200–211, Princeton, NJ, USA, 2002. Springer Verlag. Colloque avec actes et comité de lecture. internationale.

[13] Ernest Valveny, Mathieu Delalandre, Romain Raveaux, and Bart Lamiroy. Report on the symbol recognition and spotting contest. In Young-Bin Kwon and Jean-Marc Ogier, editors, *Graphics Recognition. New Trends and Challenges*, volume 7423 of *Lecture Notes in Computer Science*, pages 198–207. Springer Berlin Heidelberg, 2013.

[14] Ernest Valveny and Bart Lamiroy. Scan-to-XML : Automatic Generation of Browsable Technical Documents. In C. Suen R. Kasturi, D. Laurendeau, editor, *Sixteenth International Conference on Pattern Recognition - ICPR 2002*, volume 3, pages 188–192, Québec city, QC,

Canada, 2002. IAPR, IEEE Computer Society. Colloque avec actes et comité de lecture. internationale.

[15] Luis Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.

[16] Laurent Wendling and Salvatore Tabbone. A new way to detect arrows in line drawings. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(7):935–941, 2004.