Automated Page Layout Simplification of *Patrologia Graeca*

Bruce Robertson
Mount Allison University,
Department of Classics
#414 Hart Hall, 63D York St.
Sackville, NB, CANADA
brobertson@mta.ca

Christoph Dalitz
Niederrhein University of
Applied Sciences, Institute for
Pattern Recognition
Reinarzstr. 49, 47805 Krefeld,
GERMANY
christoph.dalitz@hsnr.de

Fabian Schmitt
Niederrhein University of
Applied Sciences, Institute for
Pattern Recognition
Reinarzstr. 49, 47805 Krefeld,
GERMANY

ABSTRACT

This paper illustrates how automated image preprocessing can improve OCR results of an important 19th century series of Greek authors. The roughly 75,000 ancient Greek pages from the volumes published by J.-P. Migne known today as *Patrologia Graeca* have a wholly regular layout, but one that is not easily reduced to a reading order by standard layout-analysis algorithms. We use a Hough transform and k-nearest-neighbor distance rejection to locate and then remove the features that confuse modern layout analysis. This allows us to separate the valuable Greek text from the less useful Latin translation, greatly simplifying later stages of processing, including de-hyphenation, automatic spell-checking and editing. Our implementation is based upon the *Gamera* Python package. We make the source code of our solution freely available.

1. ANCIENT GREEK DIGITAL CORPORA AND OCR

The study of ancient Greek literature from its origins in the eighth century BC to the fall of Byzantium in AD 1453 has benefited for decades from successive efforts to digitize source documents. Yet in today's age of 'linked' and 'big' data, the limitations of these corpora, singularly or combined, impede scholarly research and publication.

The admirably comprehensive Thesaurus Linguae Graecae (hereafter, *TLG*), which made Greek texts available on digital tape as early as 1976, includes the works of all authors listed in the Canon published supplementary to the project. The last print version of this document [1], comprises works roughly to AD 1000. According to the project website, this upper limit in recent years has been extended to the sixteenth century AD. However, the *TLG* corpus is not freely available: at present, the single user is charged a yearly fee of around $100 (USD) and is subject to a license agreement which, among other things, prohibits "the publication of a Greek text from materials supplied by *TLG* if such text does not reflect the addition of significant value by the editor" [9].[1]

In contrast, the Perseus Digital Library (hereafter, *PDL*) provides its ancient Greek texts under a Creative Commons ShareAlike 3.0 license, permitting the desirable forms of reuse that *TLG* forbids. However, the currently downloadable corpus of ancient Greek in *PDL* is far less comprehensive than *TLG*, comprising 399 XML-TEI encoded texts, with the works of the Christian Church Fathers particularly lacking. Indeed, the gap between open Greek texts and the *TLG* has been quantified in the *Open Greek and Latin Project Author Picklist*: it identifies 903 works that are not available in the *PDL* [8].

Since projects like Internet Archive and HathiTrust now make available high-quality page images of public domain texts comprising many ancient Greek authors, optical character recognition presents itself as one practical approach to completing a corpus of desirably licensed texts. In 2013, the Lace repository, comprising over 600 texts with polytonic Greek, was published [12]. Generated using Rigaudon [10], a suite of custom code comprising HOCR manipulation, automatic spell-check and text recognition based on the Gamera Greek OCR package [6], the Lace repository demonstrates that nineteenth century Greek can be OCR'd at the scale and quality necessary to contribute to the permissively licensed Greek corpus in an cost-effective manner.

2. THE *PATROLOGIA GRAECA*

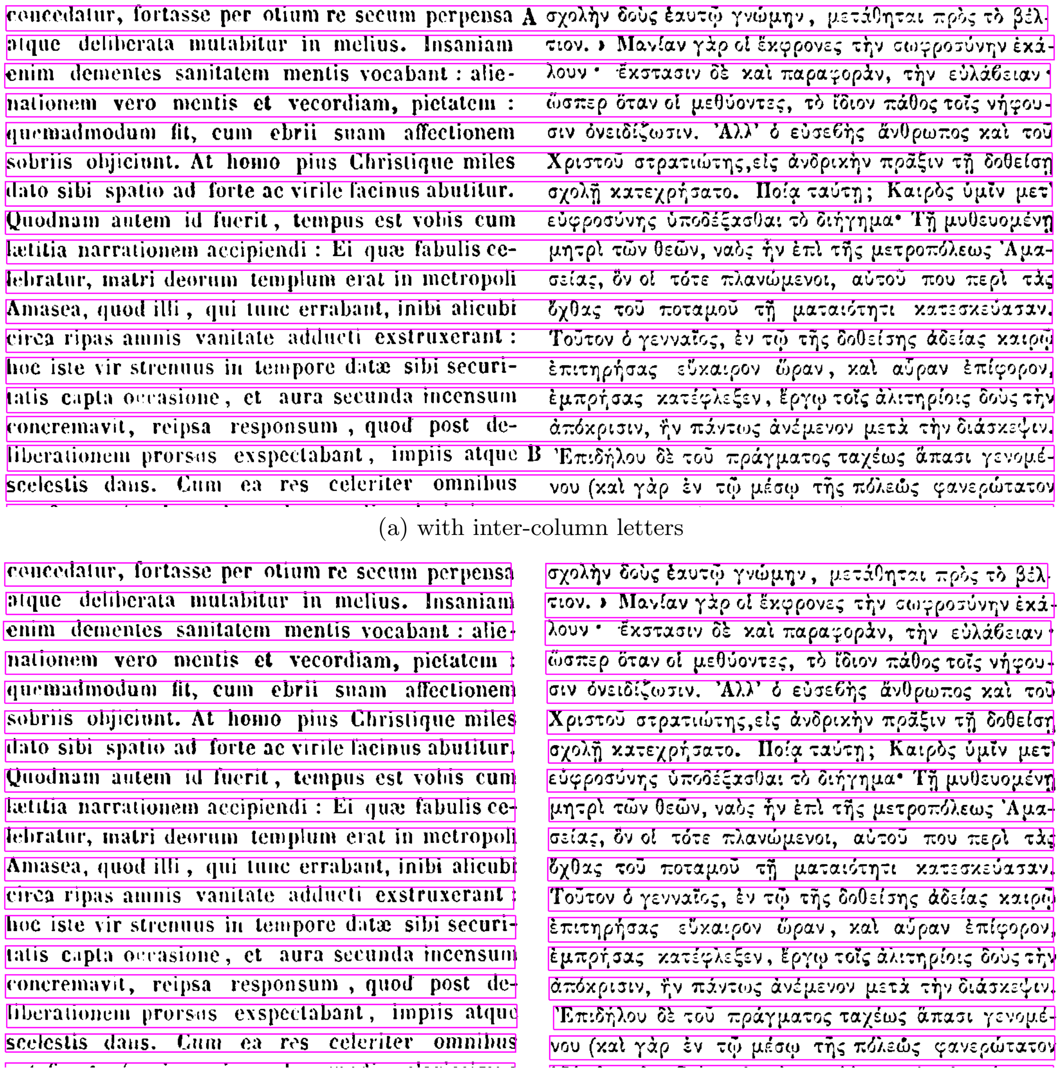
Fortunately, there exists a large series of texts dating from the late nineteenth century whose volumes fill the gap in permissively licensed Greek texts. Between 1857 and 1866, J. P. Migne's Imprimerie Catholique produced the 161 volumes of the *Patrologiae Cursus Completus, Series Graeca* (known as *Patrologia Graeca* hereafter, *PG*) a collection of works in ancient Greek by the Christian Church Fathers, ranging from the Church's beginning to AD 1453. We estimate that 75,000 pages of *PG* contain Greek texts. The vastness of this enterprise as well as the relative obscurity of some of its authors has meant that *PG* remains an frequently cited and quoted source for these authors, even in the cases where more recent editions exist.

PG contains the majority of works lacking in the *PDL*: of the 903 works identified by Franzini as not available in *PDL*,

[1] The 'Abridged Online TLG' allows for search free-of-cost, but may not be downloaded.

DaTECH May 19-20 2014, Madrid, Spain
ACM 978-1-4503-2588-2/14/05.
http://dx.doi.org/10.1145/2595188.2595213

*This paper appeared in the Proceedings of the
First International Conference on Digital Access to Textual Cultural Heritage (DATeCH), pp. 167-172 (2014)*



concedatur, fortasse per otium re secum perpensa A
atque deliberata mutabitur in melius. Insaniam
enim dementes sanitatem mentis vocabant : alie-
nationem vero mentis et vecordiam, pietatem :
quemadmodum fit, cum ebrii suam affectionem
sobriis objiciunt. At homo pius Christique miles
dato sibi spatio ad forte ac virile facinus abutitur.
Quodnam autem id fuerit, tempus est vobis cum
lætitia narrationem accipiendi : Ei quæ fabulis ce-
lebratur, matri deorum templum erat in metropoli
Amasea, quod illi, qui tunc errabant, inibi alicubi
circa ripas amnis vanitate adducti exstruxerant :
hoc iste vir strenuus in tempore datæ sibi securi-
tatis capta occasione, et aura secunda incensum
concremavit, reipsa responsum, quod post de-
liberationem prorsus exspectabant, impiis atque B
scelestis dans. Cum ea res celeriter omnibus

σχολὴν δοὺς ἑαυτῷ γνώμην, μετάθηται πρὸς τὸ βέλ-
τιον. » Μανίαν γὰρ οἱ ἔκφρονες τὴν σωφροσύνην ἐκά-
λουν· ἔκστασιν δὲ καὶ παραφορὰν, τὴν εὐλάβειαν·
ὥσπερ ὅταν οἱ μεθύοντες, τὸ ἴδιον πάθος τοῖς νήφου-
σιν ὀνειδίζωσιν. Ἀλλ' ὁ εὐσεβὴς ἄνθρωπος καὶ τοῦ
Χριστοῦ στρατιώτης, εἰς ἀνδρικὴν πρᾶξιν τῇ δοθείσῃ
σχολῇ κατεχρήσατο. Ποία ταύτη; Καιρὸς ὑμῖν μετ'
εὐφροσύνης ὑποδέξασθαι τὸ διήγημα· Τῇ μυθευομένῃ
μητρὶ τῶν θεῶν, ναὸς ἦν ἐπὶ τῆς μετροπόλεως Ἀμα-
σείας, ὃν οἱ τότε πλανώμενοι, αὐτοῦ που περὶ τὰς
ὄχθας τοῦ ποταμοῦ τῇ ματαιότητι κατεσκεύασαν.
Τοῦτον ὁ γενναῖος, ἐν τῷ τῆς δοθείσης ἀδείας καιρῷ
ἐπιτηρήσας εὔκαιρον ὥραν, καὶ αὔραν ἐπίφορον,
ἐμπρήσας κατέφλεξεν, ἔργῳ τοῖς ἀλιτηρίοις δοὺς τὴν
ἀπόκρισιν, ἣν πάντως ἀνέμενον μετὰ τὴν διάσκεψιν.
Ἐπιδήλου δὲ τοῦ πράγματος ταχέως ἅπασι γενομέ-
νου (καὶ γὰρ ἐν τῷ μέσῳ τῆς πόλεως φανερώτατον

(a) with inter-column letters

concedatur, fortasse per otium re secum perpensa
atque deliberata mutabitur in melius. Insaniam
enim dementes sanitatem mentis vocabant : alie-
nationem vero mentis et vecordiam, pietatem :
quemadmodum fit, cum ebrii suam affectionem
sobriis objiciunt. At homo pius Christique miles
dato sibi spatio ad forte ac virile facinus abutitur.
Quodnam autem id fuerit, tempus est vobis cum
lætitia narrationem accipiendi : Ei quæ fabulis ce-
lebratur, matri deorum templum erat in metropoli
Amasea, quod illi, qui tunc errabant, inibi alicubi
circa ripas amnis vanitate adducti exstruxerant :
hoc iste vir strenuus in tempore datæ sibi securi-
tatis capta occasione, et aura secunda incensum
concremavit, reipsa responsum, quod post de-
liberationem prorsus exspectabant, impiis atque
scelestis dans. Cum ea res celeriter omnibus

σχολὴν δοὺς ἑαυτῷ γνώμην, μετάθηται πρὸς τὸ βέλ-
τιον. » Μανίαν γὰρ οἱ ἔκφρονες τὴν σωφροσύνην ἐκά-
λουν· ἔκστασιν δὲ καὶ παραφορὰν, τὴν εὐλάβειαν·
ὥσπερ ὅταν οἱ μεθύοντες, τὸ ἴδιον πάθος τοῖς νήφου-
σιν ὀνειδίζωσιν. Ἀλλ' ὁ εὐσεβὴς ἄνθρωπος καὶ τοῦ
Χριστοῦ στρατιώτης, εἰς ἀνδρικὴν πρᾶξιν τῇ δοθείσῃ
σχολῇ κατεχρήσατο. Ποία ταύτη; Καιρὸς ὑμῖν μετ'
εὐφροσύνης ὑποδέξασθαι τὸ διήγημα· Τῇ μυθευομένῃ
μητρὶ τῶν θεῶν, ναὸς ἦν ἐπὶ τῆς μετροπόλεως Ἀμα-
σείας, ὃν οἱ τότε πλανώμενοι, αὐτοῦ που περὶ τὰς
ὄχθας τοῦ ποταμοῦ τῇ ματαιότητι κατεσκεύασαν.
Τοῦτον ὁ γενναῖος, ἐν τῷ τῆς δοθείσης ἀδείας καιρῷ
ἐπιτηρήσας εὔκαιρον ὥραν, καὶ αὔραν ἐπίφορον,
ἐμπρήσας κατέφλεξεν, ἔργῳ τοῖς ἀλιτηρίοις δοὺς τὴν
ἀπόκρισιν, ἣν πάντως ἀνέμενον μετὰ τὴν διάσκεψιν.
Ἐπιδήλου δὲ τοῦ πράγματος ταχέως ἅπασι γενομέ-
νου (καὶ γὰρ ἐν τῷ μέσῳ τῆς πόλεως φανερώτατον

(b) without inter-column letters

Figure 1: Effect of the inter-column letters in *PG* on the line segmentation algorithm of OCRopus.

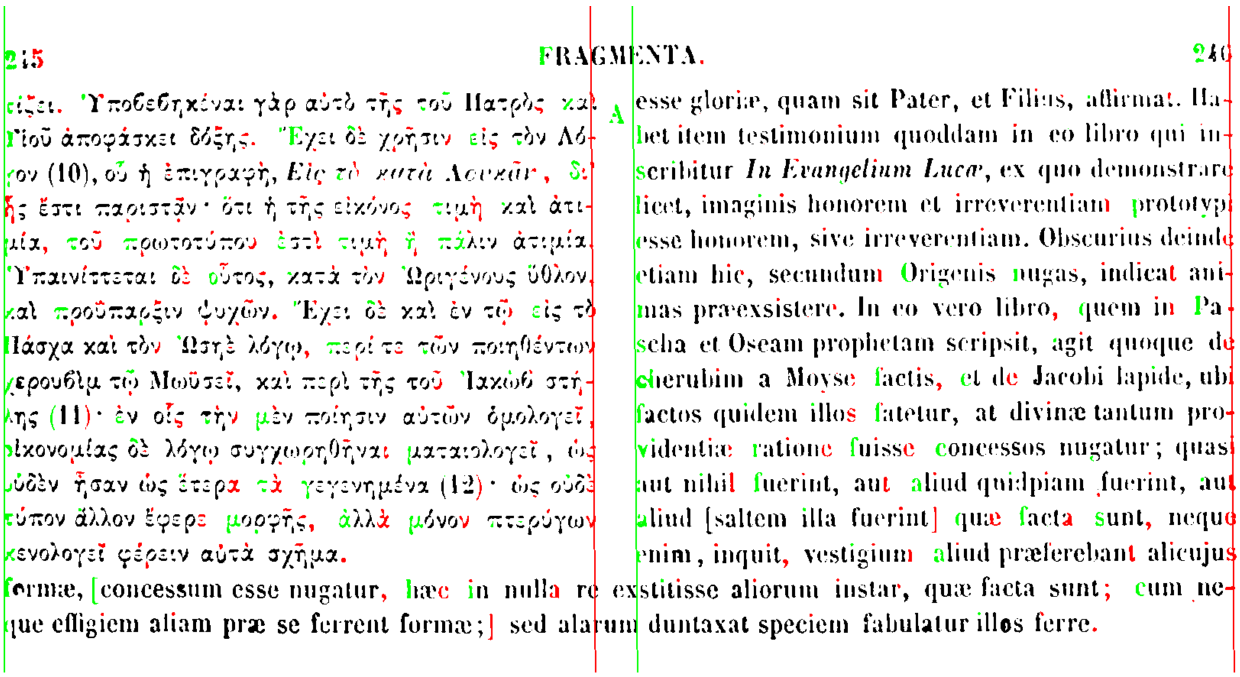
531 are to be found in its pages. That is, 58% of the works remaining to be made available in an open digital library are available in a standard, public domain work through the page images of *PG*. For this reason, an efficient process of digitizing *PG* is a prime desideratum for the overall goal of a comprehensive open digital library of ancient Greek.

This series presents several special challenges, however. Printed cheaply and in poor conditions, the type is small and, in most copies, poorly inked [2]. Despite this, Robertson has trained the OCRopus 0.7 OCR engine [3] to recognize the Greek and Latin of *PG* at around 93 - 95% character accuracy when drawing on 400 ppi images, a resolution typical of Internet Archive and other high-quality image repositories [11].

3. LAYOUT FEATURES OF THE *PG*

However, good OCR does not suffice to extract the Greek from this series or to retain its citation scheme. Migne published his Greek texts in a double-column page, usually with the Greek in one numbered column, either right or left, and a Latin translation in the other column running alongside it. As a citation scheme, he provides column (but not page) numbers and regularly spaced inter-column letters from 'A' to 'D'. These latter letters, illustrated in Figure 1(a), defy the layout analysis of most OCR engines. At very least, they cause the horizontally adjacent lines in the two columns to be read as one large line, joining the right-hand line to the corresponding one in the left column and confusing the reading order of the right-hand column. At other times these letters cause the OCR engine to join all adjacent pairs of column lines, combining the Latin and Greek together as shown in Figure 5, taken from columns 743-4 of *PG* vol. 46.

This is not a mere inconvenience for which automatic post-processing reliably can compensate. Once these letters are embedded in the stream of OCR data, it is difficult to identify them through Natural Language Processing, as they might be recognized as a separate word, as the beginning letter of a following Latin word or as the last letter in a pre-



245 FRAGMENTA. 246

κίζει. Ὑποβεβηκέναι γὰρ αὐτὸ τῆς τοῦ Πατρὸς καὶ Υἱοῦ ἀποφάσκει δόξης. Ἔχει δὲ χρῆσιν εἰς τὸν Λόγον (10), οὗ ἡ ἐπιγραφή, *Εἰς τὸ κατὰ Λουκᾶν*, δι' ἧς ἔστι παριστᾶν· ὅτι ἡ τῆς εἰκόνος τιμὴ καὶ ἀτιμία, τοῦ πρωτοτύπου ἐστὶ τιμὴ ἢ πάλιν ἀτιμία. Ὑπαινίττεται δὲ οὗτος, κατὰ τὸν Ὠριγένους ὕθλον, καὶ προϋπαρξιν ψυχῶν. Ἔχει δὲ καὶ ἐν τῷ εἰς τὸ Πάσχα καὶ τὸν Ὠσηὲ λόγῳ, περί τε τῶν ποιηθέντων χερουβὶμ τῷ Μωϋσεῖ, καὶ περὶ τῆς τοῦ Ἰακὼβ στήλης (11)· ἐν οἷς τὴν μὲν ποίησιν αὐτῶν ὁμολογεῖ, οἰκονομίας δὲ λόγῳ συγχωρηθῆναι ματαιολογεῖ, ὡς οὐδὲν ἦσαν ὡς ἕτερα τὰ γεγενημένα (12)· ὡς οὐδὲ τύπον ἄλλον ἔφερε μορφῆς, ἀλλὰ μόνον πτερύγων κενολογεῖ φέρειν αὐτὰ σχῆμα.

A

esse gloriæ, quam sit Pater, et Filius, affirmat. Habet item testimonium quoddam in eo libro qui inscribitur *In Evangelium Lucæ*, ex quo demonstrare licet, imaginis honorem et irreverentiam prototypi esse honorem, sive irreverentiam. Obscurius deinde etiam hic, secundum Origenis nugas, indicat animas præexsistere. In eo vero libro, quem in Pascha et Oseam prophetam scripsit, agit quoque de cherubim a Moyse factis, et de Jacobi lapide, ubi factos quidem illos fatetur, at divinæ tantum providentiæ ratione fuisse concessos nugatur; quasi aut nihil fuerint, aut aliud quidpiam fuerint, aut aliud [saltem illa fuerint] quæ facta sunt, neque enim, inquit, vestigium aliud præferebant alicujus formæ, [concessum esse nugatur, hæc in nulla re exstitisse aliorum instar, quæ facta sunt; cum neque effigiem aliam præ se ferrent formæ;] sed alarum duntaxat speciem fabulatur illos ferre.

Figure 2: The lines found by the Hough transform. The green lines intersect CCs with space to the left, the red lines intersect CCs with space to the right.

ceding Greek one. The 'A' is frequently misidentified as an upper case Greek alpha, adding to the uncertainty; the other letters also are sometimes mis-recognized, either as Greek or Latin characters. Moreover, because the columns in *PG* are narrow, many hyphenated forms appear. The high-quality results that can be obtained using automatic spell correction depend first on de-hyphenating these forms. Yet the inter-column letters make this process much more difficult because they can be misidentified as the first letter of the second half of a hyphenated form or cause the hyphen to not appear at the end of a line.

In contrast, with the letters identified and erased, Ocropus' automatic column detection often cleanly segements the page, as illustrated in Figure 1(b). Even in the cases where other complex layout features cause part of a page with erased inter-column letters to be rendered with joined column lines, the task of algorithmically separating these would become much easier and the output much less prone to error. Subsequently, de-hyphenation would be more accurate, causing the overall results to improve.

Finally, we recall that the inter-column letters are not merely an OCR nuisance: Migne provided them as a citation scheme. Since Ocropus' HOCR-formatted output includes the bounding boxes of all recognized lines, if we store the page location of the identified letters, we can align these with the text bounding boxes and thereby automatically reconstruct Migne's citation system in our digital text. (The other component of the citation system, the column numbers, are easily reconstructed because they are printed in sequence, two to a page.)

It is important that any automatic process be highly reliable, neither falsely identifying any other characters (and therefore erasing them from the OCR output) nor failing to identify the inter-column characters, since following processes rely on their being omitted from output. Naive image analysis approaches therefore do not suffice. For instance, although the absolute horizontal position of these characters might appear to be a useful criterion because the characters appear to be printed in the horizontal center region of the page, a misaligned (though still useful) page scan can easily shift this center region to one side or the other, making that approach impractical. We also found that horizontal and/or vertical space around the characters is not a wholly reliable criterion.

4. INTER-COLUMN LETTER REMOVAL

Our successful algorithm for removing the inter column letters 'A' to 'D' works in two steps, which are described in detail in the following subsections:

1. Locating the inter-column region with two Hough transforms.
2. Removing all characters in the inter-column region that are "not too different" from the letters 'A' to 'D'.

The second step is necessary because the text from one column occasionally extends into the other column to fill out otherwise empty space, as can be seen in Figure 2. Simply removing all characters in the inter-column region would therefore also remove actual content.

The algorithm starts with binary images that have been rotation corrected with Gamera's built in function *rotation_angle_projections*. To make all thresholds in the subsequent steps independent from font size and scan resolution, two statistical values are estimated from the list of all connected components[2] (CCs) of the page: their median width *mwidth* and their median height *mheight*.

[2] A *connected component* is a contiguous black region.

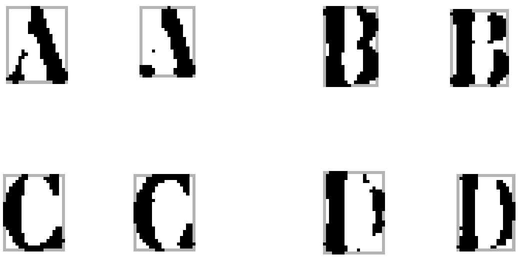


Figure 3: Due to poor print quality, letters are often broken into several connected components. The bounding boxes resulting from the Gamera function *bbox_merging* are shown in grey.

4.1 Locating the inter-column region

Our approach is based on the observation that the inter-column space is limited on the right side by characters with some space to the left, and on the left side by characters with some space to the right. In Figure 2, all CCs with white space to the left greater than 4 × *mwidth* are highlighted in green, and those with white space to the right greater than 4 × *mwidth* are highlighted in red.

These criteria also randomly match other CCs that are not on the column borders, but when we search for approximately vertical straight lines that intersect *as many of these CCs as possible*, these will be the column borders. An algorithm that solves this problem is the *Hough transform* [4]. The Hough transform counts, for all possible lines, how many points from a given point set they intersect. Figure 2 shows both for the green and red CCs the two lines that intersect most of them. The inter-column region is the region between the two lines in the middle, i.e. between the left red line and the right green line.

Obviously, this algorithm does not work when the left column is not aligned, but has a ragged right, as happens, e.g., when the text is written in verses. In this case, the left red line has a random location. This is less of a problem however, because in these cases the space between the two columns is typically so large that the lines of the two columns are not merged by OCRopus' page layout analysis and the letters are therefore never appended to the words in the columns. In other words: the original problem is absent in these cases, and failure to remove the inter-column letter does not hinder the OCR of such pages.

4.2 Identifying the letters 'A' to 'D'

The removal of the letters in the inter-column space is made complicated by two problems:

- Connected components (CCs) do not necessarily correspond to characters, as shown in Figure 3.
- Text from the columns can run into the inter-column space, as can be seen in Figure 2.

We solve the first problem by enlarging the bounding boxes of all CCs in all directions by *mwidth*×2/3 (this value worked best on 53 sample pages) and merging thereafter overlapping bounding boxes. This routine is already provided by the Gamera built-in function *bbox_merging*. As can be seen in Figure 3, this merges the broken characters.

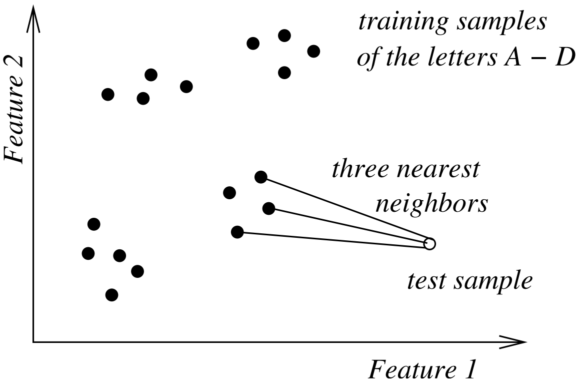


Figure 4: The idea of distance rejection demonstrated in a two dimensional feature space. The test sample is rejected, because its average kNN distance given by Equation (1) with $k = 3$ is too large compared to the training data.

To solve the second problem, we use a classification technique to discriminate between the letters 'A' to 'D' and other letters. To avoid the training of all possible character classes beforehand, as is required by a normal classifier, we utilize a technique known as *distance rejection* [5]. The idea is to train only the letters 'A' to 'D' and to identify everything else due to a "suspiciously large" distance to the training data in feature space (see Figure 4).

The criterion for a "suspiciously large" distance is based on the average distance of a test sample x to its k nearest neighbors $(y_1, \ldots, y_k)$ in feature space:

$$d_{av}(x) = \frac{1}{k}\sum_{i=1}^{k} d(x, y_j) \quad (1)$$

where $d(x, y_j)$ denotes the Euclidean distance between the feature vectors x and y_j. When $d_{av}(x)$ is greater than some threshold, the test sample x is "rejected", which means that it is considered to be too different from the training data to belong to one of the character classes in the training data. We have set the rejection threshold to 1.2 times the maximum d_{av} in the training data. We thus remove all characters in the inter-column region that have a value d_{av} below this threshold.

This approach has the advantage that very little training is required: for a chosen value of k, at least $k + 1$ training samples of each of the letters 'A' to 'D' are strictly necessary, albeit a few more will be useful in general. Moreover, it is easily implemented with the Gamera function *knn.distance_statistics* and setting *knn.confidence_types* to *CONFIDENCE_AVGDISTANCE*. As features we have used the combination of *moments*, *volume64regions*, *nrows*, and *aspect_ratio*, because these performed best in the study [7]. The feature *nrows* is the absolute height of the character, and is thus considerably larger than the other feature values, which are normalized for scale invariance. To avoid that the distance in feature space is dominated by *nrows* too much, we have therefore weighted this feature by 1/*mheight*.

5. RESULTS

To begin, we twice took a random sample of 75 pages from the 11,659 pages (16 volumes) of *PG* available in colour scans at Internet Archive. In both cases, we manually set aside the pages that contained no Greek, leaving one a sample of 57 pages with which to train, and another sample of 53 pages on which we tested the removal of the inter-column letters. We have used 172 inter-column letters ‘A’ to ‘D’ from the training pages as training data for the distance rejection with $k = 3$. Out of 187 inter-column letters on the test pages, only two were not removed. In one case this was due to noise, and in the other case the inter-column letter ‘D’ had strong serifs that were absent in the training data. As the kNN classifier also returns the character class of each removed inter-column letter, we have compared these classes to the true character classes and found not a single error. Perhaps even more importantly, our process also never erroneously removed a CC.

As predicted, the removal of the inter-column letters improves the OCR output, often causing the Ocropus line-segmentation algorithm to completely separate the Greek and Latin columns. This is illustrated in the OCR output of Figure 6, generated from the a cleaned image as in Figure 1(b). The entire left-hand Latin column appears before the Greek one. Here and also in cases where other layout features impede the Ocropus column identification, the recognition of hyphenated forms is improved and consequently spell-check is more effective.

6. CONCLUSIONS

We have demonstrated a working and reliable approach to simplifying the OCR of Greek in Migne’s *Patrologia Graeca.* Once we identify and remove the inter-column letters in *PG* through a Hough transform and k-nearest-neighbor distance rejection, the output usually conforms to reading order and in all cases is much more amenable to automatic post-processing. We make the python script for removing the inter column letters freely available on our website[3]. Apart from a working Gamera installation, it does not require any other third party software and runs on the operating systems Linux, MacOS X, and Windows.

The sheer volume of material in *PG* that is both needed by the academic community and that can benefit from this simplification makes an automatic approach to this particular problem especially desirable. Specifically, if a single copy of every Greek page in Migne has 75,000 pages and it takes an operator only one minute on average to identify the inter-column letters on a page, this task would take 1250 hours of skilled labour. Because there are multiple image sets of these volumes available, this labour might need to be repeated two or three times to derive the best results.

In contrast, a skilled operated could vet the colourized output of our program in a second or two. If our sample error rate pertains, no more than 4% of pages would need to be set aside for manual zoning, reducing the labour of zoning an entire run of Migne to around 75 hours, a much more affordable prospect.

PG may be unique in its size and uniformity, but other text series might benefit from a similar approach. References in Arabic numerals often appear in the outer margin of the many available Teubner editions, for example. Locating these would similarly improve de-hyphenation and allow the editor’s reference system to be preserved.

7. ACKNOWLEDGMENTS

The authors would like to thank Federico Boschetti for his advice regarding training Ocropus and Compute Canada for HPC computing resources.

8. REFERENCES

[1] L. Berkowitz, K. Squitier, and W. Johnson. *Thesaurus Linguae Graecae canon of Greek authors and works.* Oxford University Press, New York, 1990.

[2] R. H. Bloch. *God’s plagiarist: being an account of the fabulous industry and irregular commerce of the abbé Migne.* University of Chicago Press, Chicago, 1994.

[3] T. Breuel. The OCRopus open source OCR system. In *Proceedings IS&T/SPIE 20th Annual Symposium 2008*, 2008.

[4] W. Burger and M. J. Burge. *Principles of digital image processing: Core algorithms.* Springer, 2009.

[5] C. Dalitz. Reject options and confidence measures for kNN classifiers. In C. Dalitz, editor, *Document Image Analysis with the Gamera Framework.*, pages 16–38. Shaker Verlag, 2009.

[6] C. Dalitz, M. Droettboom, and I. Fujinaga. The gamera project, 2013. URL: `http://gamera.informatik.hsnr.de/`.

[7] C. Dalitz, G. Michalakis, and C. Pranzas. Optical recognition of psaltic Byzantine chant notation. *International Journal of Document Analysis and Recognition*, 11:143–158, 2008.

[8] G. Franzini. *Open Greek and Latin Project: Author Picklist.* University of Leipzig, Leipzig, 2013.

[9] M. Pantelia. Thesaurus linguae graecae: Single-user network license agreement, 2009. URL: `http://www.tlg.uci.edu/subscriptions/single_license.php`.

[10] B. Robertson. Rigaudon: polytonic Greek OCR engine, 2013. URL: `https://github.com/brobertson/rigaudon`.

[11] B. Robertson. Ciaconna, 2014. URL: `https://github.com/brobertson/ciaconna`.

[12] B. Robertson and F. Boschetti. Lace: Greek OCR, 2013. URL: `http://heml.mta.ca/lace`.

[3] `http://heml.mta.ca/lace/datech2014/`

```
coneedatur, fortasse per otium re secum perpensa A σχολὴν δοὺς ἑαυτῷ γνώμην, μετάθηται πρὸς τὸ βέΑ.
atque delibberata mutabitur in melius. Insaniam τιον. » Μανίαν γὰρ οἱ ἔκφρονες τὴν σωφροσύνην ἐκά-
enim dementes sanitatem mentis vocabant: alio- λουν· ἔκστασιν δὲ καὶ παραφορὰν, τὴν εὐλάβειαν·
nationem vero mentis et vecordiam, pietatem : ὥσπερ ὅταν οἱ μεθύοντες, τὸ διον πάθος τοῖς νήφου-
quemadmodum fit, cum ebrii suam affectionem  σιν ὀνειδίζωσιν. Ἀλλ΄ ὁ εὐσεβὴς ἄνθρωπος καὶ τοῦ
sobriis objiciunt. At homo pius Christique miles Χριστοῦ στρατιώτης,εἰς ἀνδρικήν πρᾶξιν τῇ δοθείσῃ
dato sibi spatio ad forte ac virile facinus abutitur. σχολῇ κατεχρήσατο. Ποίᾳ ταύτῃ· Καιρὸς ὑμῖν μετ΄
Ouodnam autem id fuerit, tempus est vobis cum εὐφροσύνης ὑποδέξασθαι τὸ διήγημα Τῇ μυθευομένῃ
ætitia narrationem accipiendi: Ei quæ fabulis ce- μητρὶ τῶν θεῶν, ναὸς ἧν ἐπὶ τῆς μετροπόλεως Ἀμα-
lebratur, matri deorum tem plum erat in metropoli σείας, δν οἱ τότε πλανώμενοι, αὐτοῦ που περὶ τὰς
Amasea, quod illi, qui tunc errabant, inibi alicubi ὄχθας τοῦ ποταμοῦ τῇ ματαιότητι κατεσκεύασαν.
cirea ripas amnis vanitate addueti exstruxerant : Τοῦτον ὁ γενναῖος, ἐν τῷ τῆς δοθείσης ἀδείας καιρῷ
hoc iste vir strenuus in tempore datæ sibi securi- ἐπιτηρήσας εὔκαιρον ὥραν, καὶ αὔραν ἐπίφορον,
tatis capta occasione, et aura secunda incensum ἐμπρήσας κατέφλεξεν, ἔργῳ τοῖς ἀλιτηρίοις δοὺςτὴν
concremavt, reipsa responsum, quod post de- ἀπόκρισιν, ἧν πάντως ἀνέμενον μετὰ τὴν διάσκεψιν.
liberationem prorsus evspectabant, impiis atque B ΤΤκιλήλου δὲ τοῦ πράγματος ταχέως ἅπασι γενομέ-
```

Figure 5: OCR results before removal of the inter-column letters.

```
concedatur, fortasse per otium re secum perpensa
atque delllerata mutabitur in melius. Insaniam
enim dementes sanitatem mentis vocabant: alie-
nationem vero mentis et vecordiam, pietatem :
qucmadmodum fit, cum ebrii suam affectionem
sobriis objiciunt. At homo pius Christique miles
dato sibi spatio ad forte ac virile facinus abutitur.
Ouodnam antem id fuerit, tempus est vobis cum
lætitia narrationem accipiendi: Ei quæ fabulis ce-
lehratur, matri deorum tem plum erat in metropoli
Amasea, quod illi, qui tunc errabant, inibi alicubi
cirea ripas amnis vanitate addueti exstruxerant :
hoc iste vir strenuus in tempore datæ sibi securi-
tatis caapta occasione, et aura secunda incensum
concremavR, reipsa responsum, quod post de-
liberationem prorsus exspectabant, impiis atque
. . .
σχολὴν δοὺς ἑαυτῷ γνώμην, μετάθηται πρὸς τὸ 8έΞ.
τιον. » Μανίαν γὰρ οἱ ἔκφρονες τὴν σωφροσύνην ἐκά-
λουν· ἔκστασιν δὲ καὶ παραφορὰν, τὴν εὐλάβειαν·
ὧσπερ ὅταν οἱ μεθύοντες, τὸ ἴδιον πάθοστοῖς νήφου-
σιν ὀνειδίζωσιν. Ἀλλ΄ ὁ εὐσεβὴς ἄνθρωπος καὶ τοῦ
Χριστοῦ στρατιώτης,εἷς ἀνδρικὴν πρᾶξιν τῇ δοθείσῃ
σχολῇ κατεχρήσατο. Ποίᾳ ταύτῃ· Καιρὸς ὑμῖν μετ΄
εὐφροσύνης ὑποδέξασθαι τὸ διήγημα· Τῇ μυθευομένῃ
μητρὶ τῶν θεῶν, ναὸς ἧν ἐπὶ τῆς μετροπόλεως Ἀμα-
σείας, δν οἱ τότε πλανώμενοι, αὐτοῦ που περὶ τὰς
ὄχθας τοῦ ποταμοῦ τῇ ματαιότητι κατεσκεύασαν.
οῦτον ὁ γενναῖος, ἐν τῷ τῆς δοθείσης ἀδείας καιρῷ
ἐπιτηρήσας εὔκαιρον ὥραν, καὶ αὔραν ἐπίφορον,
ἐμπρήσας κατέφλεξεν, ἔργῳ τοῖς ἀλιτηρίοις δοὺςτὴν
ἀπόκρισιν, ἧν πάντως ἀνέμενον μετὰ τὴν διάσκεψιν.
Ἐ᾿Επιάήλου δὲ τοῦ πράγματος ταχέως ἅπασι γενομέ-
```

Figure 6: OCR results after removal of the inter-column letters.