From Text and Image to Historical Resource: Text-Image Alignment for Digital Humanists

Dominique Stutzmann, Théodore Bluche, Alexei Lavrentiev, Yann Leydier, Christopher Kermorvant

▶ **To cite this version:**

Dominique Stutzmann, Théodore Bluche, Alexei Lavrentiev, Yann Leydier, Christopher Kermorvant. From Text and Image to Historical Resource: Text-Image Alignment for Digital Humanists. DH2015 Global Digital Humanities, University of Western Sydney, Jun 2015, Sydney, Australia. hal-01855337

HAL Id: hal-01855337
https://hal.science/hal-01855337

Submitted on 22 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Stutzmann, Dominique, Théodore Bluche, Alexei Lavrentiev, Yann Leydier, et Christopher Kermorvant. « From Text and Image to Historical Resource: Text-Image Alignment for Digital Humanists », *Digital Humanities 2015*, Sydney, 2015.

Published in html and xml versions as an online abstract for Digital Humanities 2015 (now offline) at the following URLs:

(HTML)
http://www.dh2015.org/abstracts/xml/STUTZMANN_Dominique_From_Text_and_Image_to_Histor/STUTZMANN_Dominique_From_Text_and_Image_to_Historical_R.html

(XML)

http://dh2015.org/abstracts/xml/STUTZMANN_Dominique_From_Text_and_Image_to_Histor/STUTZMANN_Dominique_From_Text_and_Image_to_Historical_Resource__Text_Image_Alignment_for_Digital_Humanists.xml

The XML-archive of the paper is among the among the archive at https://github.com/ADHO/dh2015 at the following URL: https://github.com/ADHO/dh2015/blob/master/xml/STUTZMANN_Dominique_From_Text_and_Image_to_Historical_R.xml

Here: author version without ADCO's automated formatting.

From Text and Image to Historical Resource: Text-Image Alignment for Digital Humanists

Dominique Stutzmann
Institut de Recherche et d'Histoire des Textes (CNRS), France, dominique.stutzmann@irht.cnrs.fr

Théodore Bluche
A2iA, France, tb@a2ia.com

Alexei Lavrentiev
ICAR, Interactions, Corpus, Apprentissages, Représentations (ENS de Lyon – UMR 5191), France, alexei.lavrentev@ens-lyon.fr

Yann Leydier
LIRIS Laboratoire d'Informatique en Image et Systèmes d'information (INSA de Lyon – UMR 5205), France, yann@leydier.info

Christopher Kermorvant
Teklia, France, kermorvant@teklia.com

Introduction

Written texts are both abstract and physical objects: ideas, signs and shapes, whose meanings, graphical systems and social connotations evolve through time. Beyond authorship and writer identification or palaeographical dating of textual witnesses, the materiality of text and the connexion between the ideas and their written instantiations are a matter of cultural history, historic semiology, and history of communication and representations. In the context of large, growing digital libraries of texts and digitized medieval manuscripts, the question of the cultural significance of script and the "dual nature" of texts may at last be addressed.

Several research projects, interfaces and software allow for a closer text-image association during the editing process (TILE[1], T-PEN[2], MOM-CA[3]) and for data visualisation (Mirador[4], DocExplore[5]), with interoperable annotations schemas (SharedCanvas[6]). But in most cases, the finest granularity is at line-level with an alignment being done by hand on small amounts of text (a few pages).

In this paper, we present a new method, derived from Handwritten Text Recognition, to automatically align images of digitized manuscripts with texts from scholarly editions, at the levels of page, column, line, word, and character. During the Oriflamms project funded by the French National Research Agency and Cap Digital[7], it has been successfully applied to two datasets "GRAAL" (130 pages, 13th c. manuscript, online edition at http://catalog.bfm-corpus.org/qgraal_cm) and "FONTENAY" (104 pages, 12th-13th c. charters). Partial results of alignment at word and character level are online: http://oriflamms.a2ialab.com/Charsegm/.

Our contention is that such an alignment method is the only way to gain access to new questions on a large-scale basis and massively transfer the results of traditional Humanities and textual scholarship into Digital Humanities. The automation causes not only a change of scale (larger corpora), but also a change in granularity (page by page or line by line alignment to word and character alignment). It avoids the tedious task of drawing boxes by hand around characters and allows a systematic analysis of the data. It outmatches the preceding attempts made to more closely associate both aspects of text and also opens perspectives on automated transcription.

Methodology

Text-image alignment from existing transcripts is a newly opened, specific challenge. It has relevance for the Humanities, as scholars work with a large set of already transcribed texts from manuscripts (handwritten texts). In other fields, the priority was given to pure (handwritten) text recognition. OCR-free methods have been proposed by our team (Leydier 2014) and others (Hassner 2013) which requires a precise transcript at line level, with either lesser results or no metrics on evaluation. In the present method, we use Handwritten Text Recognition (HTR), which aims to automatically transcribe the image of a handwritten text ("text image") into an electronic text, and implement it as an alignment method.

[1] http://mith.umd.edu/tile/.
[2] http://t-pen.org/.
[3] http://www.mom-ca.uni-koeln.de/mom/home.
[4] http://showcase.iiif.io/viewer/mirador/.
[5] http://www.docexplore.eu/.
[6] http://iiif.io/model/shared-canvas/1.0/index.html.
[7] http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2[CODE]=ANR-12-CORP-0010.

"Forced alignment": from HTR (Handwritten text recognition) to handwritten text segmentation

Most of the current HTR models are based on Hidden Markov Model (HMM) associated with a sliding window approach to segment the input image. These models may produce a precise character segmentation of the "text image" as a by-product. When the transcript of a line image is known, the HMM focuses on the retrieval of characters positions and forces the output to correspond to the actual sequence of characters (so-called "forced alignment"). If only the whole document transcript is available, and not the positions or transcript of the words or lines, we can use line detection methods and map the transcripts to the detected lines. Such methods also relax the annotation effort needed to produce character segmentation.

A method to automatically segment and annotate handwritten words from line images using forced alignments was proposed by (Zimmermann 2002). The problem has then shifted to mapping the whole transcription of historical documents to segmented words or lines (Kornfield 2004). When the word or line segmentation is not known, a global forced alignment of the full transcript is possible as proposed by (Fisher 2011) or at different levels (word, line, sentence, page) as proposed by (Al Azawi 2013). Our model is similar to the last one, but is based on both on A2iA proprietary image processing libraries and on Kaldi, an open source toolkit for speech recognition, which has been has been adapted for the task of text-image alignment. This system adds yet another level of complexity since it also deals with abbreviations and handwritten cursive text with no blank space between the letters, for which the character segmentation cannot easily be found.

Proposed method for handwritten text character segmentation

The character segmentation results of an incremental process (Figure 1): (1°) we convert the image to gray scale and remove black borders, (2°) we apply a text line segmentation algorithm, adapted from (Zahour 2007) to the full page, (3°) we keep the pages with a correct number of detected lines, (4°) we assign the line transcript to the line images and use them to train a first Hidden Markov Models based on Gaussian Mixture Models (GMM-HMM) recognizer, which is (5°) used to align the line transcription with the line images as described in (Bluche, 2014). (6°) Based on this result, we train a new recognizer. This process is repeated until all text lines are correctly transcribed. Afterwards, we train a final text recognizer based on deep neural networks HMMs. This model is trained with a discriminative criterion and yields better transcription results and segmentation accuracy than the standard GMM-HMM (example of forced alignment at word and letter level on Figures 2a and 2b).

Evaluation of alignment accuracy (GRAAL, FONTENAY)

Two methods were applied to evaluate the automatic alignment at word level. First, a tabular view is used to validate/reject each occurrence of a word, evaluate average accuracy, and spot problematic lines. Then a complete validation is performed by a palaeographer line by line. The accuracy is computed by the distance in pixel between the automatic segmentation and the ground-truth word boundaries (distribution on Figure 3 and 4). In GRAAL, 95%, resp. 89%, of left, resp. right, boundaries are correct with a 10px (0,84mm) tolerance. In FONTENAY, 91,8%, resp. 89,4% with a 30px (3,58mm) tolerance and 85%, resp. 74,4% with a 15px (1,79mm) tolerance, less than half of the average character width (23px i.e. 1,94mm in GRAAL and 45px i.e. 5,37mm in FONTENAY): this is a great achievement.

The alignment was also performed at a character level on GRAAL and FONTENAY. The immense number of characters makes the validation difficult. Samples show a very high accuracy rate for complex

graphic structures (e.g. 100% for st-ligature), but further tools are needed to measure the accuracy. The results of evaluation partly depend on the validator's skills in reading ancient scripts.

Evaluation of transcription accuracy (GRAAL)

The HTR-based system was also tested for recognition and evaluated according to the word and character error rate criterion (WER/CER) by splitting the corpus into a training set (101 pages) and a test set (29 pages). For the recognition, we used a lexicon of 7,005 unique words and 4gram statistical language model estimated on the train set. The standard GMM-HMM achieved 23.0% WER and 6.9% CER, whereas the hybrid deep neural network HMM achieved 19.0% WER and 6.4% CER, that is more than 80% of words and 93% of characters are accurately recognised.

Granularity and scalability

This method unlocks a new level of granularity and allows to model different letterforms ("allographs", e.g. s/ſ/S). In GRAAL, palaeographers provided the analysis of the graphical chain and the system had to choose between identified possible solutions (Figure 5). In FONTENAY, the system had to create several character models without any previous knowledge, resulting in allograph clustering.

Even at this fine level of granularity, this system is scalable to large corpora. GRAAL includes 130 pages, 10'700 lines, 114'268 words and more than 400'300 characters; FONTENAY includes 104 pages, 1'341 text lines, 22'276 words and more than 99'900 characters. In comparison, the historical databases "Saint-Gall" and "Parzival" comprise 60 pages, 1'410 text lines, 11'597 words, resp. 47 pages, 4'477 text lines and 23'478 words. The 4-year DigiPal project produced a database encompassing 61'372 manually annotated images of letters, without text transcriptions.

The system is furthermore robust (book hands and diplomatic scripts) and the data format is fully TEI compliant.

Human in the loop: Evaluation, Verifiability and Ergonomics

In interdisciplinary research, Humanities and Computer Sciences scholars must articulate their respective systems of proof and uncover underpinnings and pre-assumptions, in order to produce efficient systems that present data in a way that scholars on all sides can understand, evaluate, and trust (Stutzmann 2014). During this research, we observed many times that the so-called ground-truth was not 100% accurate or did not correspond to what the system was expected to produce (e.g. transposed words are edited in the order one should read them, while the alignment can only match the words in their order on the line), so that we had hesitations and explored new paths. This is a challenge for future developments: large resources will all comprise inaccuracies or not automatable information. Likewise, some inconsistencies in evaluation may appear. Increase of interactivity in software tools is a solution, not only to overcome shortcomings of strictly automatic approaches, but also to correct the ground-truth and improve the tools and the data models. Therefore this project also developed a user-friendly software (Leydier 2014). Agile development and interoperability concern software creation, but also corpus enhancement, to use Humanist, Computer-Scientist and Machine competences at their maximum. The alignment was performed with 3 person-months. This is obviously less than by manually drawing boxes around words (or even letters). Our tools and system open a large-scale, standardized, interoperable approach of historical scripts. The human in the loop is part of an interdisciplinary work and process avoiding tautological approaches and allowing better results, user-friendly tools and a better understanding on all sides.

References

M. Al Azawi, M. Liwicki, and T. M. Breuel (2013). WFST-based ground truth alignment for difficult historical documents with text modification and layout variations. *Document Recognition and Retrieval*.

T. Bluche, B. Moysset, and C. Kermorvant (2014). Automatic line segmentation and ground-truth alignment of handwritten documents. *International Conference on Frontiers in Handwriting Recognition*.

DigiPal: Digital Resource and Database of Manuscripts, Palaeography and Diplomatic. London, 2011–14. Available at http://www.digipal.eu/

A. Fischer, V. Frinken, A. Fornés, and H. Bunke (2011). Transcription alignment of Latin manuscripts using Hidden Markov Models. *Proceedings of the Workshop on Historical Document Imaging and Processing*.

T. Hassner, L. Wolf, and N. Dershowitz (2013). OCR-Free Transcript Alignment. *International Conference on Document Analysis and Recognition*.

E.M. Kornfield, R. Manmatha, and J. Allan (2004). Text alignment with handwritten documents. *First International Workshop on Document Image Analysis for Libraries*.

Y. Leydier, V. Eglin, S. Bres, D. Stutzmann (2014), Learning-free text-image alignment for medieval manuscripts. *International Conference on Frontiers in Handwriting Recognition*.

D. Stutzmann, and S. Tarte, Executive Summary, in T. Hassner, R. Sablatnig, D. Stutzmann, and S. Tarte, *Digital Palaeography: New Machines and Old Texts*, (*Dagstuhl Reports*, 4:7 (2014): 112-134, p. 112-114 et 132.

A. Zahour, L. Likforman-Sulem, W. Boussellaa, and B. Taconet (2007). Text Line Segmentation of Historical Arabic Documents. *International Conference on Document Analysis and Recognition*.

M. Zimmermann and H. Bunke (2002). Automatic Segmentation of the IAM Off-line Database for Handwritten English Text. *16th International Conference on Pattern Recognition*.

Figures

Figure 1: HTR-alignment

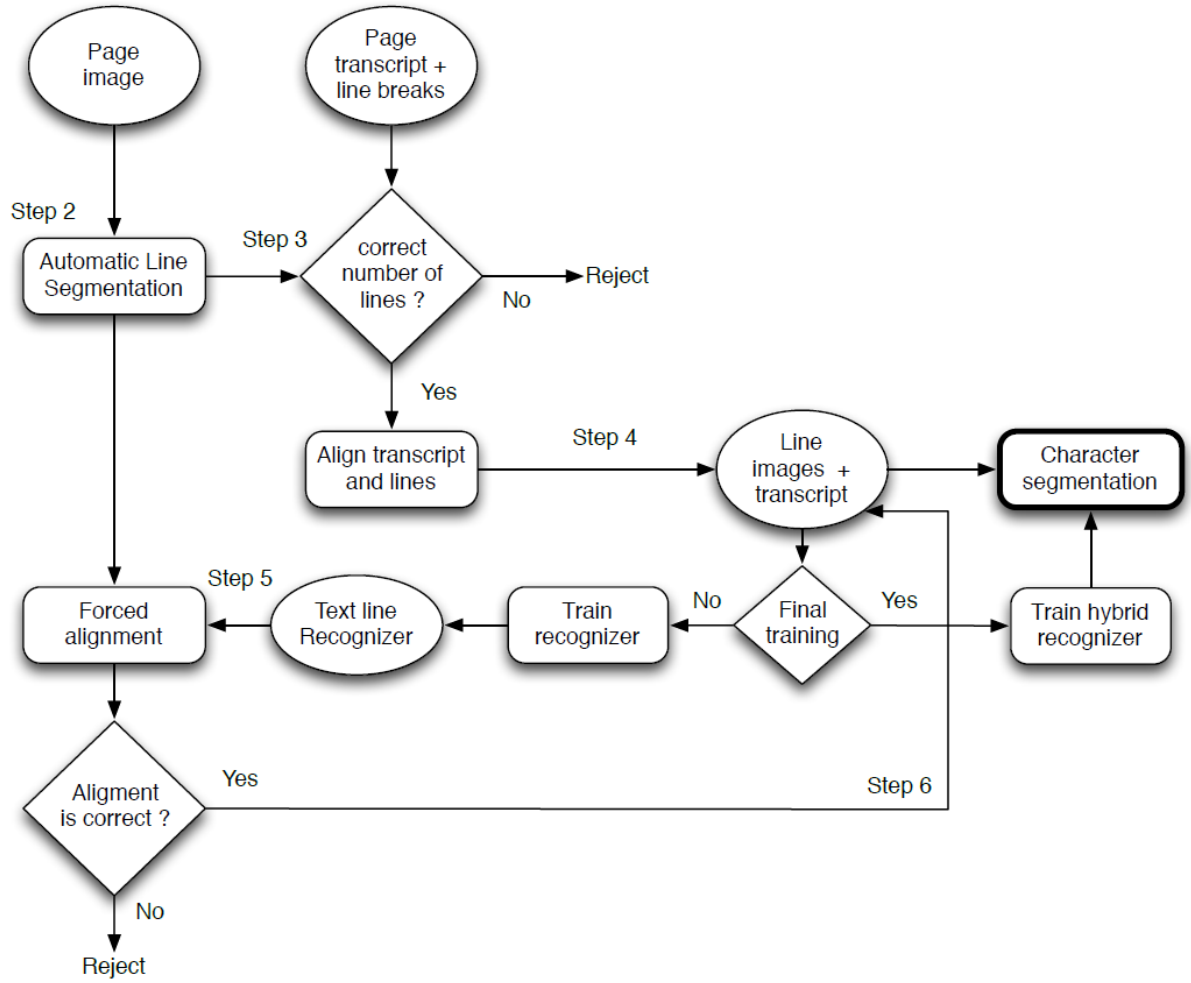

Figure 2: Forced alignment

Figure 2a: Forced alignment at word level

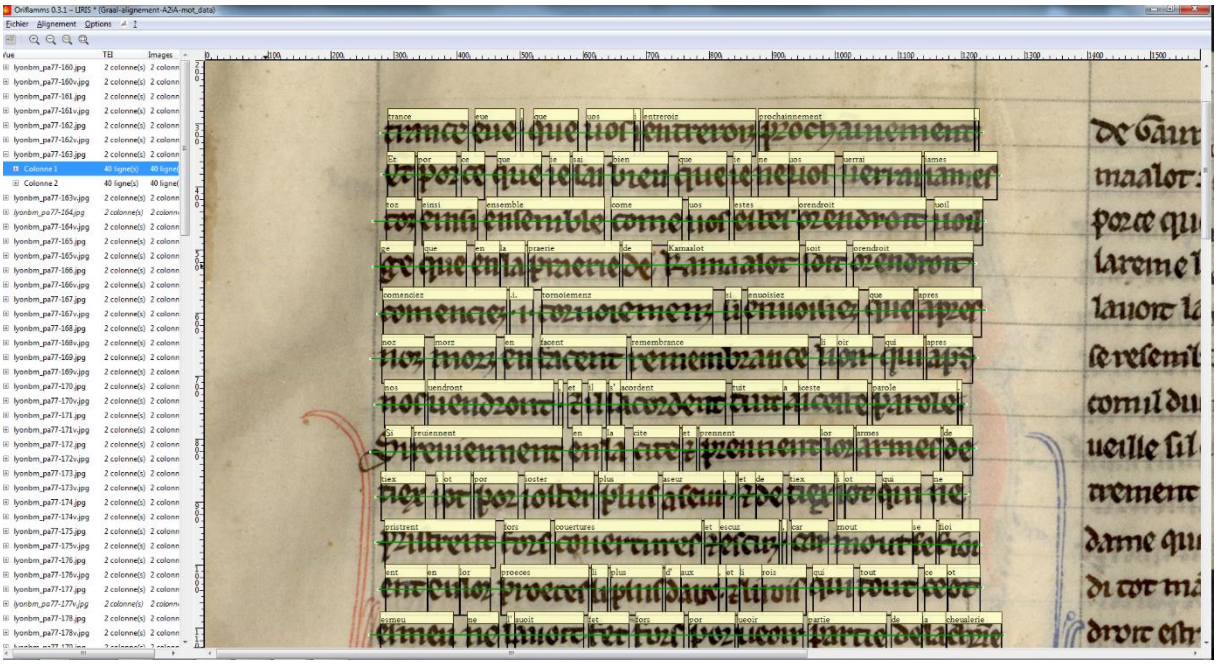

Figure 2b: Forced alignment at character level

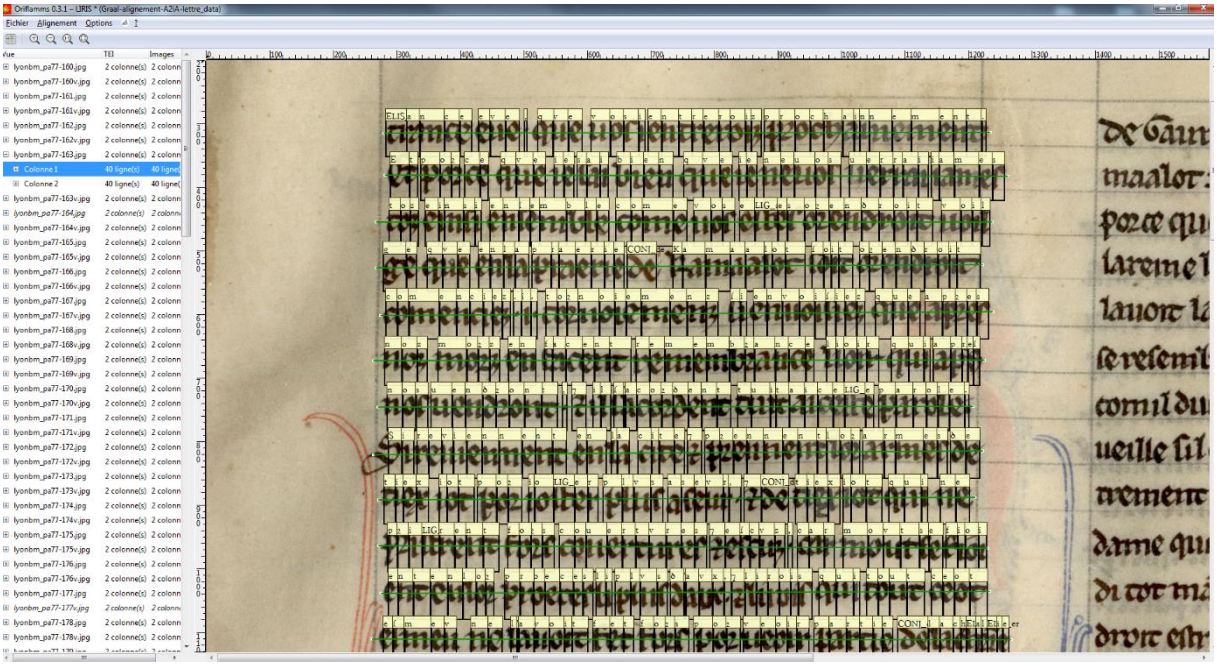

Figure 3: GRAAL alignment accuracy (number of occurrences / correction on left or right boundaries, in pixels)

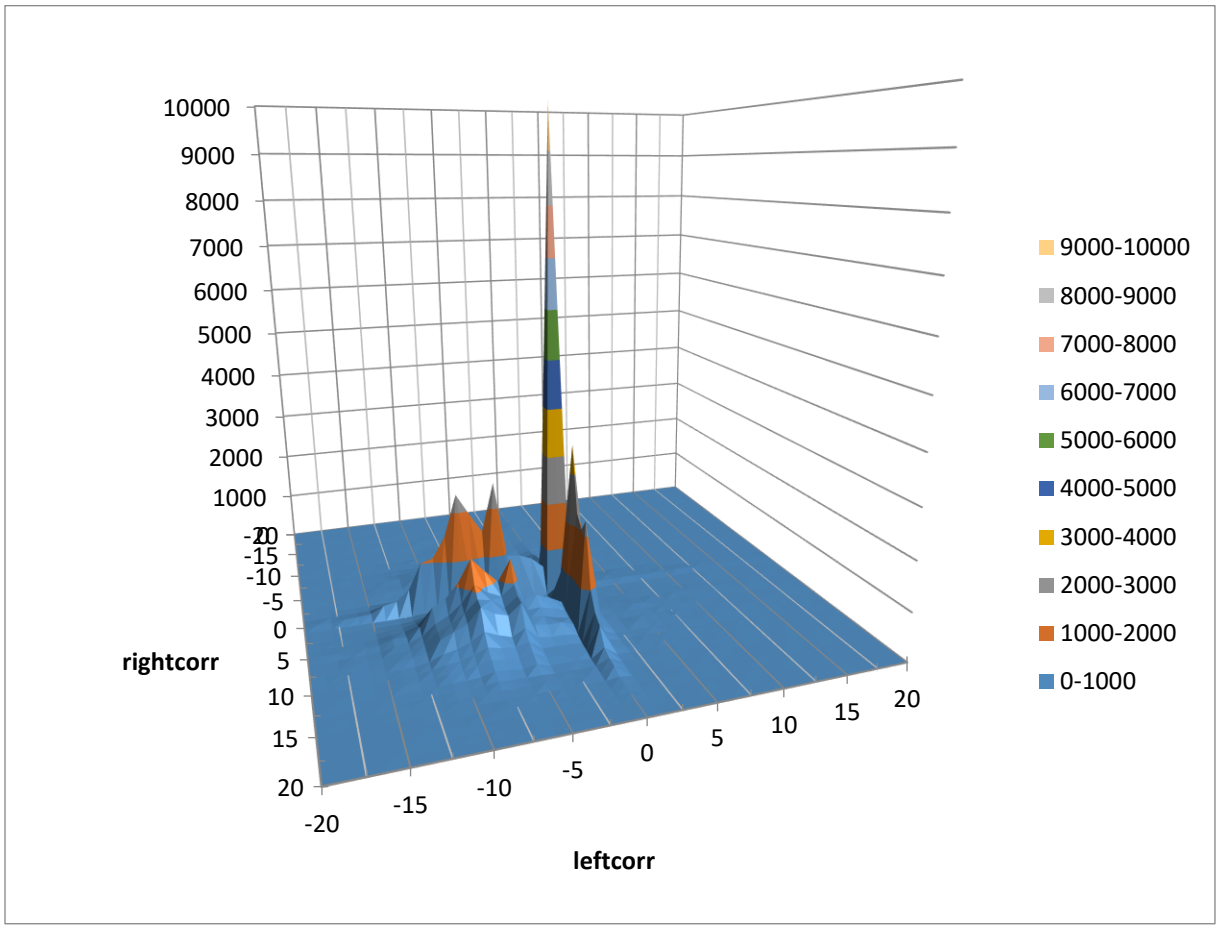

Figure 4: FONTENAY alignment accuracy (number of occurrences / correction on left or right boundaries, in pixels)

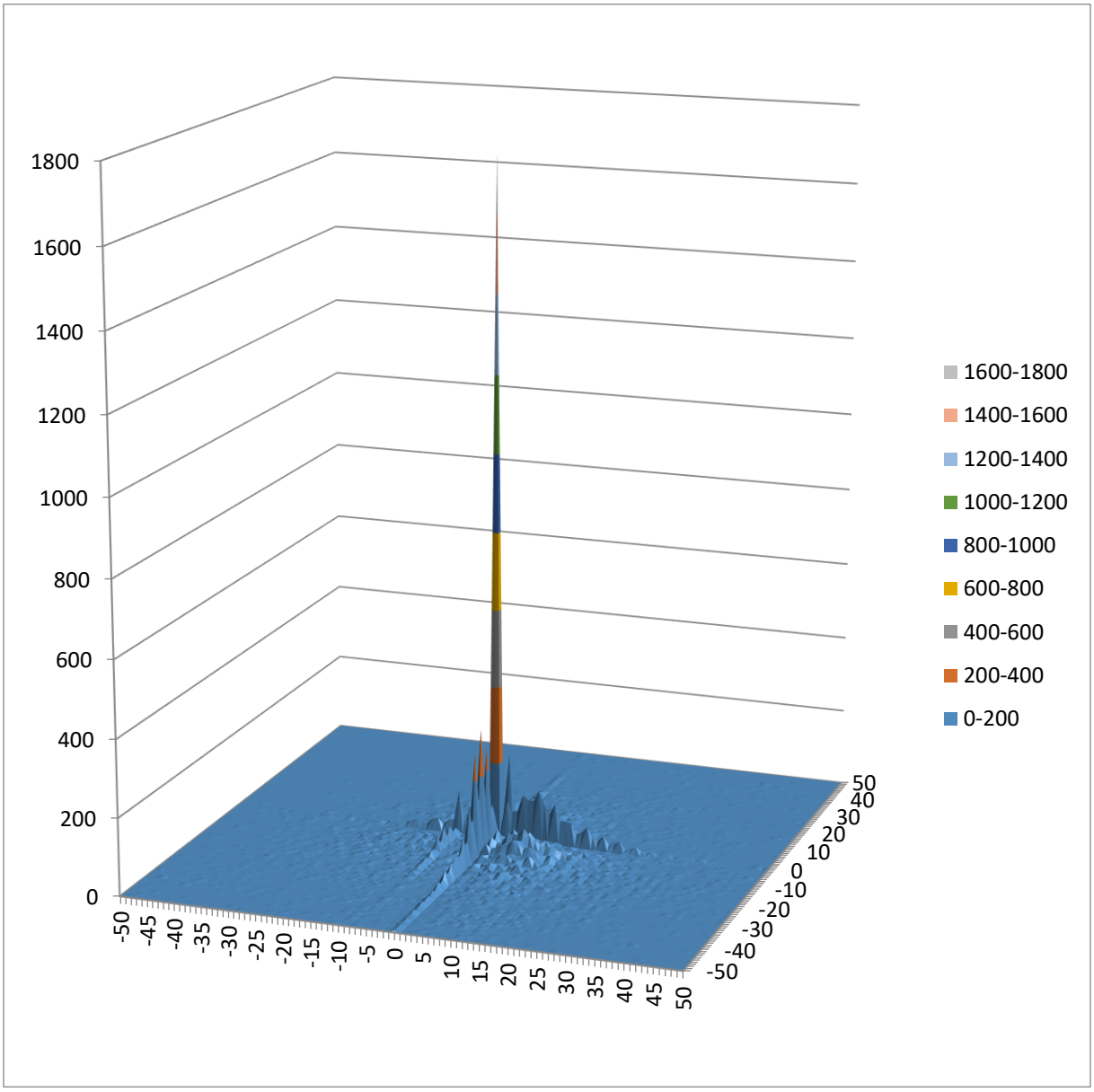

Figure 5: Modelling allographs and graphical connexions