Layout Analysis and Text Column Segmentation for Historical Vietnamese Steles

Anna Scius-Bertrand[1,2], Lars Voegtlin[3], Michele Alberti[3], Andreas Fischer[2,3], and Marc Bui[1]
[1]Ecole Pratique des Hautes Etudes, PSL, Paris, France, {firstname.lastname}@ephe.sorbonne.fr
[2]iCoSys, University of Applied Sciences and Arts Western Switzerland, {firstname.lastname}@hefr.ch
[3]DIVA, University of Fribourg, Switzerland, {firstname.lastname}@unifr.ch

ABSTRACT

Stone engravings in Historical Vietnamese steles allow historians to study the life of common people in the villages. Only recently, a large amount of images of such engravings have become available. For supporting the historians, automatic document analysis systems are needed for reading the ancient Chu Nôm characters that are written in columns from top to bottom. In this paper, we study the problem of layout analysis, which is the first step of automatic reading. Semantic segmentation is applied at pixel-level to find the title, main text, label, and reference number on the page using deep convolutional neural networks. Afterwards, seam carving is used to segment the text columns within the main text. We present baseline results for hundred exemplary pages, discuss error cases, and outline lines of future research.

KEYWORDS

historical Vietnamese steles, document layout analysis, semantic segmentation, text column segmentation, seam carving

ACM Reference Format:
Anna Scius-Bertrand, Lars Voegtlin, Michele Alberti, Andreas Fischer, and Marc Bui. 2019. Layout Analysis and Text Column Segmentation for Historical Vietnamese Steles. In *The 5th International Workshop on Historical Document Imaging and Processing (HIP '19), September 20–21, 2019, Sydney, NSW, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3352631.3352634

1 INTRODUCTION

The documents conveying the history of ancient Vietnam were written by and for the royal court and clerics. The life of the villages is not very present and that of the villagers even less so. A valuable source to fill this gap are reproductions of steles written by the villagers, which have been compiled for a century. These steles were present in the villages. The purpose of the stone engravings was to communicate various information without risking damage by weather. They provide us with information on the economic, cultural and social history of the villagers [1].

Only since a few years this document collection has become available. It consists of about 40,000 images of steles from five centuries [2, 3]. When studied individually, the images do not allow us to generalize about past practices. Hence, the use of quantitative history is essential. This is one of the goals of the Vietnamica project, which recently received an ERC grant.

The steles contain four text categories of interest to historians of this project: the title, main text, label and reference number of the document (see Figure 1). It would be of great help for the historians if a document analysis system could read these contents automatically and thus make the images amenable to searching and browsing.

In order to study the feasibility of document analysis, the authors of [4] have compiled a dataset of hundred exemplary images and annotated them with bounding polygons around the four text categories using a semi-automatic ground truth creation procedure [5]. Such ground truth annotations allow training and assessment of methods for layout analysis, which is an important first step towards automated reading.

In this paper, we investigate the task of layout analysis for historical Vietnamese steles and provide baseline results for semantic segmentation and text column segmentation, respectively. The former aims to classify each pixel of the image as belonging either to one of the four text categories, or to the stone background. We tackle this problem with deep convolutional neural networks, which have demonstrated an excellent performance for image segmentation in recent years [6]. The latter aims to segment pixels belonging to the main text into separate text columns, which contain Chu Nôm characters written from top to bottom and right to left. For this task, we rely on seam carving for finding the optimal cuts [7]. The resulting text columns can then be used as input for handwriting recognition systems.

In the following, we describe the dataset of historical Vietnamese steles and its challenges in Section 2. Afterwards, the methods used for semantic segmentation and text column segmentation are discussed in Sections 3.1 and 3.2, respectively, and experimental results are presented in Section 4. Finally, we draw conclusions and outline future work in Section 5.

2 DATASET OF HISTORICAL VIETNAMESE STELES

The 40,000 steles images have been collected since 1910 and are held in equal parts by the the Ecole Française d'Extrême-Orient and the Hán-Nôm Institute. They originate mainly from northern Vietnam and cover a time period of five centuries, with a large part dating from the 17th-19th century. A stamping procedure has been applied to copy the engravings on paper, which has then been photographed with a digital camera. Example images are shown in Figure 1.

HIP '19, September 20–21, 2019, Sydney, NSW, Australia

ACM ISBN 978-1-4503-7668-6/19/09…$15.00
https://doi.org/10.1145/3352631.3352634

Figure 1: Historical Vietnamese steles.

A total of 100 exemplary images have been randomly selected and annotated with ground truth for layout analysis and text column segmentation [4]. Following a semi-automatic procedure for ground truth creation [5], polygons have first been drawn around the four text categories, i.e. title, main text, label and reference number. Afterwards, the manually selected text regions have been processed by dynamic programming for suggesting separating seams to the human user, who has corrected them in a graphical user interface. The resulting ground truth consists of polygon boundaries around the layout elements.

Challenges for document image analysis include degradation due to fissures, impacts, and erosion of the stone, which may lead to partly or completely missing characters. Layout analysis has to cope with variations of the frame ornaments around the main text, titles that appear black or white in the image, and variations in the location of the labels and reference numbers. Finally, text column segmentation is rendered difficult by irregular column lengths, gaps, variations in the character size, and occasional splitting of main text columns into smaller sub-columns. For more details, we refer the reader to [4].

In this paper, we use an alternative, pixel-accurate format of the ground truth that has been suggested for recent document image analysis competitions [8, 9]. First, we create a binarized version of the page images for separating the engravings from the stone background. For this purpose, a Difference of Gaussians filter is used to locally enhance the foreground, before applying a global threshold [5]. As we also have layout elements that appear white on the image rather than black, we create a second version of the binarization derived from page images with inverted colors. Afterwards, we label all foreground pixels within the bounding polygons with particular colors to indicate the type of the layout element. The blue channel is used to mark *background* with $0x1$, *title* with $0x2$, *label* with $0x4$, *main text* with $0x8$, and *reference number* with $0x16$. An example of the pixel-accurate ground truth is illustrated in Figure 2. The four layout elements are colored in red and the foreground pixels in black.

The resulting ground truth is illustrated in Figure 2. Note that due to the binarization, it still contains a considerable amount of noise. Nevertheless, it provides us with a more detailed derivative of the ground truth when compared with the bounding polygons.

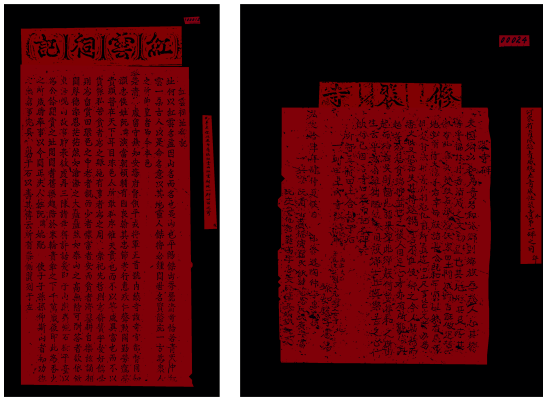

Figure 2: Pixel-level ground truth.

3 LAYOUT ANALYSIS METHODS

We perform automatic document layout analysis in two steps as illustrated in Figure 3. First, we go from the RGB domain (Figure 3a) to the pixel-label domain (Figure 3b) by performing semantic segmentation. Afterwards, we perform seam carving on the main text pixels in order to segment the individual text columns (Figure 3c). The final result consists in tight bounding polygons around the layout elements.

In the following, we describe the methods used for semantic segmentation and text column segmentation in more detail.

3.1 Semantic Segmentation

Semantic segmentation takes an RGB image of the steles as input and returns an image of the same dimension, in which each pixel is labeled with a particular color that corresponds to one of the four text categories, or the stone background.

(a) Original image (b) Semantic segmentation (c) Text column segmentation

Figure 3: Layout analysis for images of historical Vietnamese steles based on semantic segmentation and text column segmentation. The four text categories are *title* (highlighted with orange color), *main text* (white), *label* (purple), and *reference number* (light-green). In (c) each text-column is highlighted in a different color.

The use of semantic segmentation for historical document images has been proposed recently by Stewart and Barrett in [10]. In this work, we follow a similar approach using a Fully Convolutional Neural Network (FCN) to perform this task. Specifically, we employ a vanilla Tiramisu network (FC-DenseNet103) introduced by Jégouet al. [6]. This network leverages dense blocks introduced by Huang et al. [11], combined with an advanced version of skip connections, which have demonstrated state-of-the-art performance on several natural image datasets as shown by Long et al. in [12]. The network architecture is illustrated in Figure 4.

Since the input size of the images is too large to fit into a regular GPU, we perform training by selecting 300 random crops of size 192×192 from each page and define this as an epoch. Network training is conducted with regular Stochastic Gradient Descent (SGD).

As discussed in Section 2, we use all foreground pixels that lie within a bounding polygon of a particular layout element as targets for the neural network. For the challenging stele images, we have to take into account binarization errors, i.e. missing foreground as well as background noise. The neural network is thus trained with a noisy pixel-level ground truth.

3.2 Text Column Segmentation

After semantic segmentation, we further segment the pixels classified as main text into individual text columns using seam carving, a well-established technique for text line segmentation in historical document images [13]. In this work, we use the recently introduced seam carving method proposed by Alberti et al. [7], which has achieved a strong performance on several medieval manuscript datasets. The result of text column segmentation are tight polygons around the foreground pixels of the individual text columns.

The seam carving algorithm aims to find optimal vertical cuts between the text columns based on an energy map. Considering a binary image $x = (x_1, .., x_{n\times m}) \in \{0, 1\}^{n\times m}$ where main text pixels have the value 1, the energy map $E = (e_1, \ldots, e_{n\times m})$ is composed of three parts:

$$E(x) = B(x) + T(B(x)) + S(B(x), T(B(x))) \quad . \tag{1}$$

The first part is the background energy $B(x)$, which is based on the Euclidean distance of a pixel to the centroid of its closest connected component (CC):

$$B_i(x_i) = \frac{1}{\min\limits_{c \in CC} \|l(x_i) - l(c)\|} \quad \forall i = 1, \ldots, n \times m \quad , \tag{2}$$

where $l(.)$ returns the image coordinates. Only CCs with an area larger than γ times the average CC area are taken into account, with respect to the CC size parameter $\gamma > 0$. The background energy is high for pixels close to a CC.

The second part is the text energy $T(B(x))$, which corresponds to $B(x)$ if the pixel is labeled as main text, and zero otherwise, in

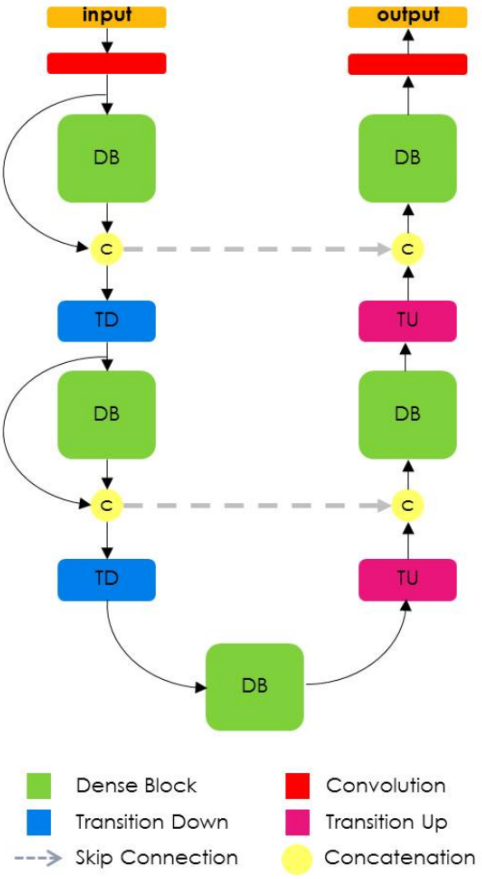

Figure 4: Network architecture of the vanilla Tiramisu network used for semantic segmentation (FC-DenseNet103). Figure taken from [6].

order to emphasize the importance of main text pixels:

$$T_i(B_i(x_i)) = \begin{cases} B_i(x_i) & x_i = 1 \\ 0 & x_i = 0 \end{cases} \quad \forall i = 1, \ldots, n \times m \tag{3}$$

The third part is the smoothing energy $S(B(x), T(B(x)))$, which aims to fill gaps and to remove high-frequency noise. It is computed by means of two consecutive convolutions:

$$S(B(x), T(B(x))) = C_2(C_1(T(B(x)) + B(x))) \tag{4}$$

where C_1 uses a global kernel of size $n \times n$ in the form of a "+" sign of ones and C_2 is a local mean filter of size 32×32.

After computing the energy map, seams are initiated every α vertical pixels from top to bottom as well as from bottom to top, with respect to the density parameter $\alpha > 0$. They are penalized with β for deviating from the vertical axis, where $\beta > 0$ is the third and final user-defined parameter. If two seams overlap in two positions, they are merged into the one with the lower energy.

Finally, the CCs are clustered with respect to the number of seams to their right in order to form text columns and tight polygons are computed around all main text pixels. For more details on the seam carving method, we refer to [7].

Table 1: Mean intersection over union (mean IU) results on the test set for semantic segmentation. The mean IU is indicated per text category, as well as macro averaged (M) and micro averaged (μ).

	mean IU		class-wise				
	M	μ	bg	title	label	text	reference
FC-DenseNet103	0.88	0.98	0.99	0.84	0.86	0.98	0.76

Table 2: Mean intersection over union (mean IU) results on the test set for text column segmentation.

	semantic segmentation	ground truth
Fischer et al. [5] (2010)	0.63	0.82
Alberti et al. [7] (2019)	0.84	0.90

4 EXPERIMENTAL EVALUATION

We have evaluated the layout analysis methods on the dataset of historical Vietnamese steles [4], in order to establish baseline results for semantic segmentation and text column segmentation, respectively, when applying methods from the current state of the art.

4.1 Setup

The 100 pages in the dataset are split randomly into three disjoint sets for training (50% of the pages), validation (20%), and testing (30%). The training set provides learning samples for the neural network, the validation set is used for parameter optimization, and the test set is used to evaluate the final performance.

For semantic segmentation, we used the deep learning framework DeepDIVA [14, 15], in order to ensure reproducibility of our experiments. The network is trained from scratch for 300 epochs using a momentum of 0.9 and an initial learning rate of 0.001 with the adapting policy of decaying it by a factor of 10 at epochs 100 and 200.

For text column segmentation, we have optimized α over the range [10, 50], β over [3000, 13000], and γ over [0.1, 0.5] using the hyper-parameter optimization framework SigOpt [16].

The performance is measured in terms of mean intersection over union (mean IU)[1] [17], for both semantic segmentation and text column segmentation.

4.2 Results

Table 1 presents the results for semantic segmentation. The mean IU is computed for each text category separately and then averaged. Macro averaging (M) attributes the same weight to all categories, while micro averaging (μ) weights the categories according to their number of pixels in the ground truth. That is, micro averaging

[1] Also referred to as Jaccard index.

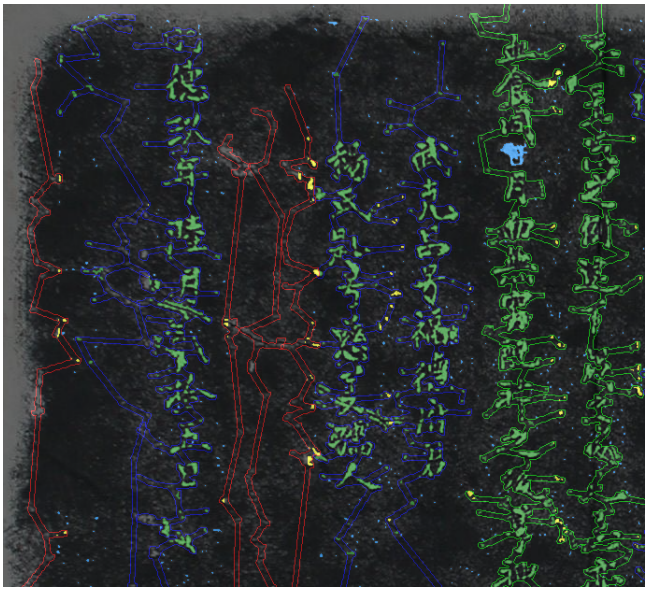

Figure 5: Typical error cases for semantic segmentation and text column segmentation. Missing columns are highlighted in blue and additional columns in red.

assigns higher weights to the categories background and main text, which are more frequent than the others.

The deep convolutional neural networks have a strong performance for distinguishing the main text from the background with a mean IU of 0.98 for the main text (the optimal mean IU is 1.0). Their performance drops for the less frequent layout elements, which may be due to a lack of learning samples. Nevertheless, as illustrated in Figure 3b, the networks also achieve a considerable quality for these elements.

Table 2 shows the results for text column segmentation. The mean IU is computed both with respect to the ground truth and the automatic result of semantic segmentation, in order to evaluate the impact of errors during semantic segmentation. Furthermore, we compare the seam carving method with the dynamic programming algorithm that has been used during ground truth creation [5]. The former significantly outperforms the latter, suggesting that it may be beneficial to use it for ground truth creation in the future. The best fully automatic result is a mean IU of 0.84 for extracting text columns from historical Vietnamese steles.

4.3 Discussion

Figure 5 illustrates typical error cases for the layout analysis methods. Green polygons are correctly segmented text columns, blue polygons are missed columns, and red polygons are additional columns that are inserted even though there is no main text. Both errors are linked to failures of the semantic segmentation method, which lost too many main text pixels (leading to missed columns) or added too many stone background pixels (leading to additional columns).

We assume that these failures may be due to the ground truth, which is rather noisy. As stated in Section 2, we base our ground truth on binarization to obtain main text pixels within the bounding polygons of the layout elements. Every binarization error is thus a wrong learning sample for the network. In future work, we envisage to investigate more closely the impact of noise in the ground truth on the final network performance.

5 CONCLUSIONS

Considering the challenges for document image analysis posed by the historical Vietnamese steles, strong baseline results have been achieved for layout analysis when using state-of-the-art methods for semantic segmentation and text column segmentation. The combination of deep convolutional neural networks and seam carving led to a notable mean IU of 0.84 for fully automatic extraction of text columns.

Clearly, there is room for improvement over this baseline result. First of all, the quality of the pixel-level ground truth should be improved. Also, the neural networks are expected to profit from transfer learning as well as data augmentation with synthetic samples, especially in the case of the label and reference number that lack training data. It may also be a rewarding line of research to include prior knowledge about the shape of the Chu Nôm characters into the semantic segmentation process.

The strong results also encourage to investigate the next step towards automatic reading, i.e. handwriting recognition and keyword spotting based on the extracted text columns. Even if the transcription accuracy is far from perfect, we would expect that historians can already profit from the resulting indexation for searching and browsing the comprehensive document collection.

REFERENCES

[1] P. Papin, "Aperçu sur le programme "publication de l'inventaire et du corpus complet des inscriptions sur stèles du viêt-nam"," *Bulletin de l'École française d'Extrême-Orient*, vol. 90, no. 1, pp. 465–472, 2003.

[2] P. Papin, T. K. Manh, and N. V. Nguyên, *Catalogue des inscriptions du Viêt-Nam*. EPHE, EFEO, Institut Han-Nôm, 2007–2012.

[3] ——, *Corpus des inscriptions anciennes du Vietnam*. EPHE, EFEO, Institut Han-Nôm, 2005–2013.

[4] A. Scius-Bertrand, J. Bosom, P. Papin, and M. Bui, "Towards Automated Reading of Historical Vietnamese Steles," in *Proc. 19th Conf. of the International Graphonomics Society (IGS)*, 2019.

[5] A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, and M. Stolz, "Ground truth creation for handwriting recognition in historical documents," in *Proc. 9th Int. Workshop on Document Analysis Systems (DAS)*, 2010, pp. 3–10.

[6] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–19.

[7] M. Alberti, L. Voegtlin, V. Pondenkandath, M. Seuret, R. Ingold, and M. Liwicki, "Labeling, Cutting, Grouping: an Efficient Text Line Segmentation Method for Medieval Manuscripts," in *Proc. 15th Int. Conf. on Document Analysis and Recognition (ICDAR)*, Sydney, Australia, sep 2019.

[8] F. Simistira, M. Bouillon, M. Seuret, M. Wursch, M. Alberti, R. Ingold, and M. Liwicki, "ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Kyoto, Japan: IEEE, nov 2017, pp. 1361–1370.

[9] F. Simistira Liwicki, R. Saini, D. Dobson, J. Morrey, and M. Liwicki, "ICDAR 2019 Historical Document Reading Challenge on Large Structured Chinese Family Records," in *to appear in 15th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, mar 2019.

[10] S. Stewart and B. Barrett, "Document image page segmentation and character recognition as semantic segmentation," in *Proc. 4th Int. Workshop on Historical Document Imaging and Processing*, 2017, pp. 101–106.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[13] A. Asi, R. Saabni, and J. El-Sana, "Text line segmentation for gray scale historical document images," in *Proc. 1st Int. Workshop on Historical Document Imaging and Processing*, 2011, pp. 120–126.

[14] M. Alberti, V. Pondenkandath, M. Würsch, R. Ingold, and M. Liwicki, "DeepDIVA: A Highly-Functional Python Framework for Reproducible Experiments," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Niagara Falls, USA, aug 2018.

[15] M. Alberti, V. Pondenkandath, L. Voeögtlin, M. Würsch, R. Ingold, and M. Liwicki, "Improving Reproducible Deep Learning Workflows with DeepDIVA," in *6th Swiss Conference on Data Science (SDS)*, Bern, Switzerland, jun 2019.

[16] S. Clark and P. Hayes, "SigOpt Web page," http://www.sigopt.com, 2019. [Online]. Available: http://www.sigopt.com

[17] M. Alberti, M. Bouillon, R. Ingold, and M. Liwicki, "Open Evaluation Tool for Layout Analysis of Document Images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 1st International Workshop on Open Services and Tools for Document Analysis (OST)*, Kyoto, Japan, nov 2017, pp. 43–47.