Exploring the Alignment of Transcriptions to Images of Encrypted Manuscripts

Goio Garcia, Pau Torras, Alicia Fornés
Computer Vision Center, Computer Science Dept.
Universitat Autònoma de Barcelona, Spain
goio.garcia@autonoma.cat
{ptorras,afornes}@cvc.uab.cat

Beáta Megyesi
Department of Linguistics
Stockholm University
Sweden
beata.megyesi@ling.su.se

Abstract

The automatic transcription of encrypted manuscripts is a challenge due to the different handwriting styles and the often invented symbol alphabets. Many transcription methods require annotated sources, including symbol locations. However, most existing transcriptions are provided at line or page level, making it necessary to find the bounding boxes of the transcribed symbols in the image, a process referred to as alignment. So, in this work, we develop several alignment methods, and discuss their performance on encrypted documents with various symbol sets.

1 Introduction

Historical encrypted manuscripts often employ invented alphabets for encryption. These alphabets are usually composed of digits, Latin or Greek letters, Zodiac or alchemical signs, invented symbols and diacritics. Not surprisingly, the automatic generation of transcription of such documents is a challenge due to difficulties in dealing with many different symbols and handwriting styles presented in a few lines or pages. Despite the advances in the image processing research field, current Handwritten Text Recognition (HTR) techniques are based on deep learning architectures, which require a considerable amount of labelled data to train. Unluckily, in the case of encrypted documents, the amount of available transcriptions is insufficient and the few existing transcriptions are provided at line or page level, without any labelling of the area where the symbol is located. Normally, in HTR, the symbols in a document are typically labelled with bounding boxes with respect to the coordinates of its top left and bottom right points. The consequence of the lack of information about the symbols' location is that only end-to-end segmentation-free transcription methods can be employed, restricting the use of other high-performing types of recognition methods, such as symbol spotting and symbol classification models.For this reason, it is desirable to devote efforts in aligning the existing transcriptions represented as text sequences to the actual cipher symbols in the image, that is, leverage the information granted by the transcription, such as the presence and order of symbols, and obtain accurate bounding boxes, so that all type of recognition methods can be applied.

In this work, we explore different methods for aligning the transcriptions with symbol images, a task referred as text-to-image alignment.

2 Previous Work

The alignment of graphical sources to their transcriptions has been explored through various means in the literature. The most common is the idea of employing a text recognition model and then finding the mapping between the ground truth and the recognised text. The earliest instances of this idea can be found in (Tomai et al., 2002; Kornfield et al., 2004), in which the authors segment the page into words and then align the ground truth using dynamic programming algorithms. Further work combine Hidden Markov Models (HMMs) with the Viterbi algorithm for the word image-to-text alignment phase (Rothfeder et al., 2006; Feng and Manmatha, 2006; Toselli et al., 2007; Fischer et al., 2011), with the latter also dealing with writing-transcript inconsistencies. Good word segmentation is usually difficult to obtain, so this family of techniques is usually combined with some form of correction algorithm that prevents catastrophic failure cases (De Gregorio et al., 2022; De Gregorio et al., 2023).

In (Torras et al., 2021) an alignment method is proposed where an attention-based Sequence-to-Sequence model is trained to recognise symbol lines from the Copiale (Knight et al., 2011) and

Borg (Aldarrab et al., 2018) ciphers. The highest attention activations are used to locate the correspondence between each symbol and its display in the image. To account for possible transcription mistakes, the Levenshtein algorithm is used to find the minimum edit path between the ground truth and the prediction of the model. The same authors propose the use of Recurrent and Convolutional Neural Network-based recognition architectures trained using the Connectionist Temporal Classification (CTC) loss (Torras et al., 2023). The alignment is found by forcing the known ground truth sequence when decoding the output of the model – for every slice of the input image, the corresponding transcription class is generated. Their advantage over Sequence to Sequence models is that they require much less data and compute time to train and can recover from mistakes much more reliably.

3 Connectionist Temporal Classification

In this work, we build from the foundation of CTC-based alignment methods and attempt to find ways of preventing its common failure cases. In this section, we briefly introduce the CTC loss to describe the improvements we propose.

CTC models (Graves et al., 2006) are a popular paradigm for sequence-learning. They have the capability to process arbitrary length input and output sequences, as long as the input sequence is longer than the target. In addition, they do not need alignment information during training. Given an input sequence $X = (x_i)_{i=1}^{T}$ and a target sequence $L = (l_i)_{i=1}^{U}$, where each element of the target sequence belongs to an alphabet A of length a, the CTC model predicts an output sequence $Y = (y^{(i)})_{i=1}^{T}$, where $y^{(i)} \in \mathbb{R}^{a+1}$ is a vector containing the probability for every symbol in alphabet $A' = A \cup \emptyset$ at time-step i. Symbol $\emptyset$ is known as the "blank" character. By selecting a symbol for each time-step, one obtains a path π. Since input sequences may differ (e.g. in speech or handwriting there are varying speeds or character widths), the CTC output should represent the same target sequence despite different input sequences.

The CTC defines a mapping from paths to target sequences $\mathcal{B}$. For some path π, $\mathcal{B}$ removes consecutive repeated symbols and then deletes every blank symbol. We denote as $\mathcal{B}^{-1}(L)$ the set of feasible paths, all paths that map to target sequence L. To obtain the probability of a target sequence L, we must marginalize over all feasible paths

$$p(L|X) = \sum_{\pi \in \mathcal{B}^{-1}(L)} p(\pi|X) \tag{1}$$

where the probability of path π is simply the product of probabilities for the selected symbol at every time-step. The loss function is taken as the negative log-probability of the target path

$$L_{ctc} = -\log p(L|X). \tag{2}$$

4 Methods

For text-to-image alignment of symbols, we test three different methods, based on the CTC model.

Method 1: Maximum entropy regularization

The CTC loss function has been modified with the goal of discouraging peaky distributions, as proposed by Liu et al. (2018) and using the implementation given in the same paper. A new regularization term is introduced, which encourages exploration when decoding by maximizing the entropy of the set of feasible paths and as a result distributes the probabilities over multiple alignments, avoiding the convergence into a single feasible path. This way the CTC decoding process is improved, as it requires for the probabilities of multiple feasible paths to be summed. The parameter β controls the importance of this term.

Given a target path L and an input sequence X, the new loss function reads

$$L_{entctc} = L_{ctc} - \beta H(p(\pi|L, X)) \tag{3}$$

where $H(p(\pi|L, X))$ is the entropy of the set of feasible paths. The entropy of feasible paths is given by the Shannon entropy

$$H(p(\pi|L, X)) = -\sum_{\pi \in \mathcal{B}^{-1}(L)} p(\pi|L, X) \log p(\pi|L, X) \tag{4}$$

Method 2: Define anchors based on single model confidence

By leveraging the reduced overconfidence in incorrect outputs of Method 1, it is possible to identify high confidence alignments directly through the CTC output probabilities. These high confidence alignments are referred to as anchors, and are used to further subdivide the image and repeat predictions on the low confidence segments while keeping the anchored alignment intact. An example of the process is illustrated in Figure 1.

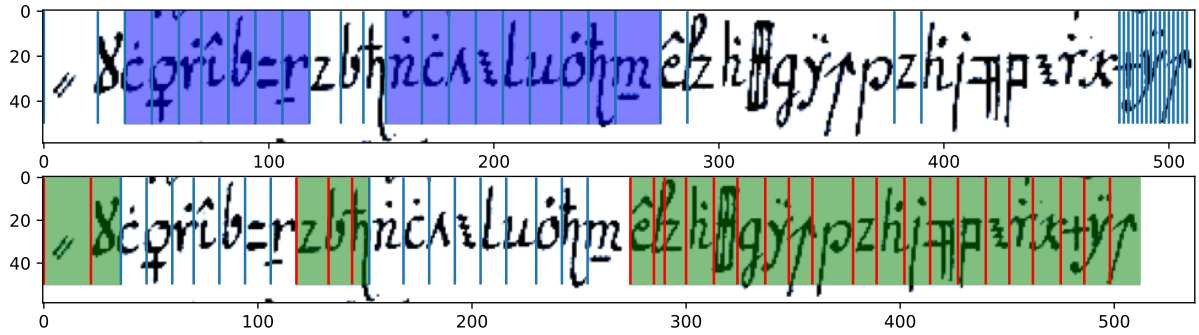

Figure 1: Top image: initial alignment of a Copiale string (blue vertical bars) outputted by the $\beta = 0.1$ regularized model, and selected anchors based on single model-confidence (blue regions). Bottom image: final alignment after reprediction of non-anchored alignments (green regions). The changed alignments are marked in red.

The determination of the threshold for what constitutes a high confidence alignment is performed automatically through Otsu's method (Otsu, 1979), which identifies the optimal cutoff point between two classes by minimizing the intraclass variance. Then, for the high confidence class, the anchor threshold is taken by substracting 1.5 standard deviations from the median.

A final filter is performed on the alignments that have been classified as high confidence, only leaving as anchors those for which their direct neighbours are also high confidence. The reason for this is that the confidence outputted by the CTC is still high for the first incorrect alignment, but falls off for the following ones.

Method 3: Define anchors based on multiple model agreement

By training two separate models and comparing their decoded outputs, it is possible to select anchors based on similarity of both alignments. Alignments were selected as anchors when their Intersection over Union (IoU) was over a 0.95 threshold. The separately trained models included the original unregularized model, the regularized model presented in Method 1 and a version of the regularized model trained on horizontally inverted images (H.Flipped model).

5 Experiments

Given the three methods described above, we carry out experiments with various symbols sets.

5.1 Datasets and Metrics

The models have been evaluated in two different datasets. Annotations include the 1-dimensional bounding boxes of every symbol present in the image as ground-truth. a) Copiale: Training dataset of 253747 images of line fragments from the Copiale cipher, extracted from a dataset of 648 lines of text. Validation dataset of 126 images of full lines, and test dataset containing 139 images of full lines. b) Borg: Training dataset of 2868 images of line fragments from the Borg cipher, extracted from 178 lines of text. Validation dataset composed of 22 images of full lines, and test dataset containing 49 images of full lines. Line fragments used for training are a form of data augmentation and may include overlap between several images.

To evaluate the models' performance, we use the Average Intersection over Union (AIoU) metric and the Hits@X metric. AIoU quantifies the average overlap of the predicted bounding boxes with the ground truth, with 1 being ideal and 0 being worst. The Hits@X metric indicates the percentage of bounding boxes that have achieved an IoU greater than X%.

5.2 Results of Method 1

The first goal was to determine the optimal value for the regularization parameter β. Values 0.05, 0.1 and 0.2 were tested, and the results for the best value is shown in Table 1. The best performing value depends on the dataset: 0.1 for Copiale, and 0.05 for Borg.

While the overall performance of the model has been reduced with the exception of the Borg dataset when $\beta = 0.05$, by looking at the distribution of output confidences we have observed that a reduction in high confidence alignments has been achieved, as desired.

Table 1: Results for all tested methods. Regularized columns always show results for the best β parameter on each dataset, 0.1 for Copiale and 0.05 for Borg.

			Method 1	Method 2		Method 3	
		Unreg.	Regularized	Unreg.	Reg.	Unreg. + Reg.	Reg. + H.Flip.
Copiale	AioU	0.807	0.763	**0.826**	0.794	0.813	0.780
	Hits@25	0.972	0.935	0.994	0.976	0.976	0.953
	Hits@50	0.952	0.911	0.973	0.951	0.954	0.925
	Hits@75	0.763	0.695	0.785	0.722	0.780	0.720
Borg	AioU	0.545	0.559	0.645	**0.654**	0.552	0.559
	Hits@25	0.786	0.748	0.960	0.893	0.780	0.755
	Hits@50	0.685	0.681	0.788	0.797	0.683	0.678
	Hits@75	0.284	0.394	0.340	0.429	0.305	0.389

5.3 Results of Method 2

The described method has been applied to both datasets, yielding results shown in Table 1. There is a significant performance uplift when compared to Method 1 for the Borg dataset, while in the Copiale dataset it achieves similar performance. The usage of the regularized model over the unregularized one does not have a clear impact. For instance, in the case of the Borg dataset it increases the rate of accurate alignments at the same time it reduces the number of rough alignments, causing the average IoU to stay within a 1% range of each other. Usually, it is desirable to maximize the number of perfect alignments even at the cost of reducing the quality of other alignments. This is because in case the model predictions are manually corrected, perfect alignments do not need intervention and so it results in a workload reduction. In the Copiale dataset, however, the overall performance is better with the unregularized model. This seems to indicate that the regularized model is better for smaller datasets with few labelled data.

5.4 Results of Method 3

Method 3 outperforms the baseline and Method 1 in the Copiale dataset when using the agreement between the unregularized and regularized models, as shown in Table 1. However, Method 2 is still the best-performing one in both datasets. Using the model trained on horizontally flipped images does not seem to provide a significant advantage.

One of the main challenges that have been observed when using the agreement between the regularized and unregularized models, is the fact that both models tend to produce the same mistakes, marking the incorrect predictions as anchors. As they have been trained using the same data and have a very similar architecture, this situation is not uncommon. In the case of the agreement between the regularized and the H.Flipped model, it is hard to obtain any anchor at all, since the regularized model tends to produce mistakes at the end of the image, while the H.Flipped model produces them at the start, leaving no region where both models agree. The performance of this method might be improved if a different architecture can be used in the aggregation process.

6 Discussion and Conclusions

In this paper, we have explored different methods to improve the accuracy of handwritten text-to-image alignment of symbols to known transcriptions using CTC models. We presented a regularized model based on maximum entropy of feasible paths, and two methods making use of the concept of anchors: high confidence alignments that remain fixed during subsequent predictions. The effectiveness of each method depends on the dataset, but Method 2 achieves the best results in general, with similar performance as the original model for the Copiale dataset, and a significantly improved alignment for the Borg dataset. Method 1 accomplishes its goal of reducing high confidence outputs, but does not improve performance by itself in the Copiale dataset. However, the regularized model can be a useful tool when used in conjunction with other methods. Although Method 3 is an improvement over single-model predictions, the similarity in the predictions of the aggregated models leads to a lack of capacity to detect anchors and correct errors. Further work will include experiments on encrypted sources with various alphabets.

Acknowledgments

This work has been partially supported by the Swedish Research Council (grant 2018-06074, DECRYPT), the Spanish projects PID2021-126808OB-I00 (GRAIL) and CNS2022-135947 (DOLORES) and the Spanish FPU Grant FPU22/00207. The authors acknowledge the support of the Generalitat de Catalunya CERCA Program to CVC's general activities.

References

Nada Aldarrab, Kevin Knight, and Beáta Megyesi. 2018. The Borg.lat.898 Cipher. https://www.su.se/english/research/research-projects/decipherment-of-historical-manuscripts/the-borg-cipher-1.688283.

Giuseppe De Gregorio, Ilaria Citro, and Angelo Marcelli. 2022. Transcript Alignment for Historical Handwritten Documents: The MiM Algorithm. In Cristina Carmona-Duarte, Moises Diaz, Miguel A. Ferrer, and Aythami Morales, editors, *Intertwining Graphonomics with Human Movements*, Lecture Notes in Computer Science, pages 45–60, Cham. Springer International Publishing.

Giuseppe De Gregorio, Giuliana Capriolo, and Angelo Marcelli. 2023. End-to-End Transcript Alignment of 17th Century Manuscripts: The Case of Moccia Code. *Journal of Imaging*, 9(1):17.

Shaolei Feng and R. Manmatha. 2006. A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '06, pages 109–118, New York, NY, USA. Association for Computing Machinery.

Andreas Fischer, Volkmar Frinken, Alicia Fornés, and Horst Bunke. 2011. Transcription alignment of Latin manuscripts using hidden Markov models. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, HIP '11, pages 29–36, New York, NY, USA. Association for Computing Machinery.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Kevin Knight, Beáta Megyesi, and Christiane Schaefer. 2011. The Copiale Cipher. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 2–9, Portland, Oregon. Association for Computational Linguistics.

E.M. Kornfield, R. Manmatha, and J. Allan. 2004. Text alignment with handwritten documents. In *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.*, pages 195–209.

Hu Liu, Sheng Jin, and Changshui Zhang. 2018. Connectionist temporal classification with maximum entropy regularization. In *Advances in Neural Information Processing Systems*, pages 837–847.

Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Jamie Rothfeder, R. Manmatha, and Toni M. Rath. 2006. Aligning Transcripts to Automatically Segmented Handwritten Manuscripts. In Horst Bunke and A. Lawrence Spitz, editors, *Document Analysis Systems VII*, Lecture Notes in Computer Science, pages 84–95, Berlin, Heidelberg. Springer.

C.I. Tomai, Bin Zhang, and V. Govindaraju. 2002. Transcript mapping for historic handwritten document images. In *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 413–418.

Pau Torras, Mohamed Ali Souibgui, Jialuo Chen, and Alicia Fornés. 2021. A Transcription Is All You Need: Learning to Align Through Attention. In Elisa H. Barney Smith and Umapada Pal, editors, *Document Analysis and Recognition – ICDAR 2021 Workshops*, volume 12916, pages 141–146. Springer International Publishing, Cham.

Pau Torras, Mohamed Ali Souibgui, Jialuo Chen, Sanket Biswas, and Alicia Fornés. 2023. Segmentation-Free Alignment of Arbitrary Symbol Transcripts to Images. In Mickael Coustaty and Alicia Fornés, editors, *Document Analysis and Recognition – ICDAR 2023 Workshops*, Lecture Notes in Computer Science, pages 83–93, Cham. Springer Nature Switzerland.

Alejandro H. Toselli, Verónica Romero, and Enrique Vidal. 2007. Viterbi Based Alignment between Text Images and their Transcripts. In Caroline Sporleder, Antal van den Bosch, and Claire Grover, editors, *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).*, pages 9–16, Prague, Czech Republic, June. Association for Computational Linguistics.