Robust Transcript Alignment on Medieval Chant Manuscripts

Timothy de Reuse, Ichiro Fujinaga
Center for Interdisciplinary Research in Music Media and Technology
McGill University
Montreal, Canada
{timothy.dereuse, ichiro.fujinaga}@mcgill.ca

***Abstract*—We present a generalizable method of performing transcript alignment on the lyrics of medieval chant manuscripts. We use optical character recognition to generate a preliminary transcript of each page and then use a global sequence alignment method to match it up to the known correct transcript, combining the two incomplete sources of information into a high-quality alignment. We demonstrate this approach on manuscript pages using two different script styles from four different sources. This method requires little training data and works even when the transcript and the page itself have differing textual content, achieving per-syllable accuracies of 80–90% across the four sources.**

***Index Terms*—Optical Music Recognition, Transcript Alignment, Sequence Alignment, Historical Document Analysis**

I. Introduction

Western music notations has its origins in neume notation, which began as marks placed above syllables of text to denote their melodic contour. Accordingly, most notated Western European music up until the 15th century was for voice, and included lyrics [1]; however, little research in the field of Optical Music Recognition (OMR) has addressed the textual component of music notation. In order to encode early notated vocal music into a symbolic format, it is necessary to perform Optical Character Recognition (OCR) on the handwritten lyrics, which is not trivial and requires a large amount of training data to achieve a practical level of accuracy [2].

In some cases, we have access to a high-quality transcript of the lyrics of a manuscript. This tells us *what* text is on each page, but not *where* the text lies. In order to graphically display the location of text or associate musical content with each syllable, we need to ascertain where each syllable of the ground truth appears on the page; this task is called *transcript alignment*. Assigning musical notation to the aligned text is itself nontrivial, but for lack of space this issue is not addressed here. We focus on the lyrics of medieval chant manuscripts, which have many idiosyncrasies that make them difficult to analyze. To our knowledge, there is no previous research on transcript alignment on this type of manuscript.

The method we demonstrate here uses an OCR system to produce a preliminary transcription of each manuscript image, which misidentifies many characters but has estimated positions for all of them. We then use a dynamic programming-based sequence alignment algorithm to align this OCR transcript to an existing, correct transcript, combining the two sources of information into a high-quality alignment. Since chant manuscripts exist in a wide variety of notations and script styles, we focus on developing a method that can be adapted to other manuscripts; existing OCR models could be used with minimal effort spent on training and data preparation, since there is no requirement that the OCR transcript be highly accurate.

This research has been supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) and the Fonds de Recherche du Québec – Société et Culture (FRQSC).

II. Related Work

Only a handful of published works in OMR address text. Most research focuses on separating lyrical and musical content, using traditional document analysis techniques [3]–[5] or deep learning [6]. George [7] discusses using OCR on extracted lyrics and aligning identified words with musical symbols on the page. Hankinson et al. [8] incorporate an existing, pre-trained OCR system as a step into an OMR system intended for printed square-note neume notation, but note that the resulting text still has many errors.

On historical handwritten documents, transcript alignment is generally performed using Hidden Markov Models (HMMs) [9]–[11] or with dynamic programming methods such as dynamic time warping [12], [13]. A common paradigm with these methods is the exploitation of *anchor words*, which are words that appear only once in a given transcript and so can be located in the image with a higher degree of certainty, so that alignment can be performed on strings of text that lie between anchor words. It is also possible to render a transcript as an image file and directly map between regions of the manuscript image and regions of the synthesized image, though this requires the availability of a font that bears resemblance to the handwriting used in the manuscript [14].

Aligning the inaccurate output of an OCR system to a known transcript is not a novel technique, but it has most often been done as a step in training or evaluating an OCR system. Feng and Manmatha [15] use HMM-based sequence alignment to assign a score to the similarity between an OCR system's output and a ground-truth transcript. Romero-Gomez et al. [9] use HMM-based sequence alignment to automate the generation of OCR training data, though their method operates on individual text lines rather than whole pages.

III. Alignment Method

We operate on four manuscripts available in the Cantus manuscript database [16]: the Salzinnes Antiphonal (CDN-Hsmu M2149.14),[1] Einsiedeln (Stiftsbibliothek, Codex 611(89)),[2] St. Gallen 388 (CH-SGs 388), [3] and St. Gallen 390 (CH-SGs 390).[4] The Salzinnes and Einsiedeln manuscripts are written in Gothic script (Figure 1), while the two from St. Gallen are written in Carolingian minuscules (Figure 2). The Cantus database[5] contains plain text transcripts for each chant in these manuscripts. Any text that is not part of a chant is not transcribed. Chants often begin on one page and end on the succeeding page, but the Cantus database lists each chant as occurring on the page where it begins. So, given a folio number, we can retrieve a string of transcribed chants that mostly correspond to the text on the page of interest, but may contain text *not* on the page (the beginning and endings of chants on neighboring pages), and may exclude some text that *is* on the page (non-chant text). The goal of this method is to correctly align all textual content that lies in the intersection between the Cantus transcript and the manuscript page, while ignoring content that only appears in one or the other.

A. Pre-Processing

For the Salzinnes manuscript, we use a set of pages where text layers have already been extracted using the pixelwise classification method developed by Calvo-Zaragoza et al. [6]; on the other three manuscripts we use a set of pages where the text layer has been isolated from the background manually. After this step, the text layer is deskewed to straighten the text lines. Lines are identified by finding prominent peaks on the horizontal projection profile, and splitting into strips at local minima.

We assemble training data by manually transcribing manuscript pages, and train two OCR models using the OCRopus open-source OCR system [17]. In principle, we could use any OCR system capable of outputting per-character horizontal positions, but of the open-source options available we found OCRopus to be both accurate in its character segmentation and convenient in its utilities for generating training data. The first model is trained on forty pages of Gothic script from the Salzinnes manuscript, which comprises a total of 2302 words. The second is trained on Carolingian minuscule script from St. Gallen 390 (five pages) and St. Gallen 388 (two pages) comprising a total of 1140 words. Both of these models are trained until their output no longer seems to improve, which took about eight hours on an ordinary desktop PC for each of them. The character error rate of the OCR results against the training data is 0.127 for the Gothic script model and 0.125 for the Carolingian script model.

The scribes who wrote these manuscripts often abbreviated commonly occurring lyrics to save space on expensive parchment. A word may not always appear as an abbreviation, but there are rarely two different abbreviations that represent the same word. `dn̄s` → `dominus`, `dn̄e` → `domine`, `&` → `et`, and `alla` → `alleluia`. We also add an `m` after vowels that have bars over them, which is the most common meaning of that symbol (e.g., `ū` → `um`). This treats abbreviations as if written out in full, but with the letters of each syllable overlapping, allowing us to handle them normally when grouping characters into syllables. The assumption made here is that abbreviations never collapse more than syllable into a single letter, which is true for those that we handle.

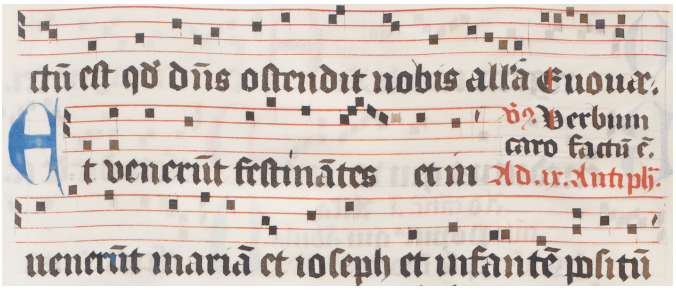

Fig. 1. A section from folio 042r in the Salzinnes Manuscript. Note abbreviations, non-chant text, and the large ornamental letter.

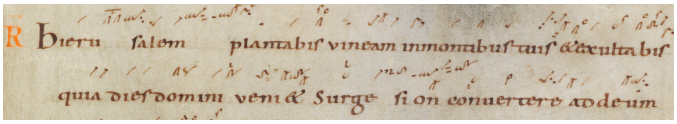

Fig. 2. A section from folio 23 in the St. Gallen 390 Manuscript.

B. Sequence Alignment

After running the trained OCR models on each text line in a page, we retrieve a transcript that is inaccurate, but contains correct positions for most of the characters. We use the Needleman-Wunsch algorithm [18], a global sequence alignment method, to match these sequences together. This algorithm lines up both transcripts alongside each other, and tries to make the sequences match in as many positions as possible by adding *gaps* to both sequences, which let the alignment skip over characters that are only in one of the sequences. More formally, a gap represents a position where one would have to insert or delete a character in order to transform one sequence into the other. Where there is non-chant text in the OCR transcript, the algorithm tends to insert gaps into the Cantus transcript; where there is text in the Cantus transcript that is not on the page being processed, the algorithm tends to insert gaps into the OCR transcript. Our implementation uses affine gap penalties, which encourages the algorithm to use fewer long gaps rather than many short ones, encouraging matched regions to be unbroken, contiguous strings.

Table I shows the result of this sequence alignment on an excerpt from the top of a page of the Salzinnes manuscript (Figure 1), where the text starts in the middle of the chant:

[1] `cantus.simssa.ca/manuscript/133/`
[2] `www.e-codices.unifr.ch/en/list/one/sbe/0611`
[3] `www.e-codices.unifr.ch/en/csg/0388/`
[4] `www.e-codices.unifr.ch/en/csg/0390`
[5] `cantus.uwaterloo.ca/`

TABLE I
AN EXCERPT FROM THE CANTUS TRANSCRIPT OF THE PAGE SHOWN IN FIGURE 1, AN EXCERPT FROM THE OCR TRANSCRIPT OF THE SAME PAGE, AND THE RESULTS OF ALIGNING THESE TWO STRINGS WITH THE NEEDLEMAN-WUNSCH ALGORITHM.

Cantus Transcript	OCR Transcript
bethleem et videamus hoc verbum quod fa	ctū est qd ds ostendit notbis alla Euouae.
ctum est quod dominus ostendit nobis alleluya euouae	vus perbum
et venerunt festinantes et	t venerūt festinātes et m Ad.x. Antipl.
invenerunt mariam et joseph et infantem positum	uenerūt mariā et ioseph et infantē positū

↓

Global Alignment

```
---------------------------------------ctum est q--d d-----s ostendit notbis alleluia Euouae.
bethleem et videamus hoc verbum quod factum est quod dominus ostendit no-bis alleluya euouae--

vus perbumt venerumt festinamtes et m Ad.x. Antipl.uen erumt mariam et ioseph et infantem positums
--- -e----t venerunt festinantes et -----------i--nven erunt mariam et joseph et infantem positum-
```

"-ctum est quod dominus..." Because of this, the transcript here includes a portion of the text from the previous page, which is necessary in practice because we have no information as to where a chant crosses a page break. Wherever two characters are matched up in this alignment, we can assume that both characters refer to the same textual material.

C. Grouping into Syllables

Chant is more naturally segmented into syllables than into words, since each syllable is sung to one or more neumes. The last stage of the alignment involves splitting the Cantus transcript into syllables and analyzing which syllables have been matched to which characters in the OCR transcript. If the syllable is aligned with any characters, whether or not the characters themselves match, the syllable is assigned the union of those characters' bounding boxes on the page. This also applies if a syllable is assigned to some characters and some gaps. In Table I, the second instance of quod in the transcript, assigned to q--d in the alignment, will take the bounding box of qd in the original OCR. If a syllable is aligned only with gaps, then it is assumed to refer to text not on the page, and is ignored. Any OCR characters that are not aligned with any syllable of the transcript are assumed to be non-chant text and are ignored. Finally, we reposition each bounding box to compensate for the deskewing of the whole page in the preprocessing step, so that they correctly match up with their content (syllables) in the original manuscript.

IV. RESULTS

We evaluated this method against manually assembled ground-truth annotations of alignment on five pages from the manuscripts under examination. The annotations consisted of a list of text syllables, each associated with a single bounding box that fits the characters on that page. We calculated the intersection-over-union (IoU) of each syllable's bounding box in the ground truth and its corresponding bounding box from our alignment. However, the size of our bounding boxes is influenced by the precision of our text line segmentation method. To ensure that we were not also implicitly evaluating how well we segment text lines, we binarized the text layer first, and ignored white pixels in the IoU calculation. This means that two bounding boxes were given an IoU score of 1 only if all of the black pixels in one are also in the other, regardless of their total size. We place each syllable into one of three categories based on its IoU under this scoring method:

- Matches: IoU > 0.5
- Partial matches: IoU < 0.5
- Misses: IoU = 0, or the corresponding bounding box is missing from our alignment altogether.

The results of this evaluation are in Table II, where the Accuracy column denotes $\frac{\text{Matches}}{\text{Total}}$. Figure 3 shows the result of the alignment from the same excerpt shown as an example in Table I. Larger images illustrating the results of the alignment on pages from each manuscript are available in the appendix (Section VI). In each of these images, syllables of text are highlighted with yellow if they correspond to a partial match, and red where they correspond to a miss. It is difficult to directly compare these results to others in the literature, since transcript alignment on chant manuscripts has (to our knowledge) not been previously addressed; however, Fischer et al. [11] achieved word-label accuracies of 92% with their HMM-based alignment algorithm on a Latin manuscript with a similar style as the two St. Gallen manuscripts we used here.

The method performed best on long, unbroken strings of chant text, with few abbreviations. Most errors were a result of non-chant text, occurring frequently at the beginning or end of larger segments of non-chant text that are bookended by musical text; the sequence alignment algorithm can "figure out" that there exists text in the OCR transcript to skip, but it begins or ends the gap too early or too late. Rubrics[6] or other isolated markings can also cause similar off-by-one errors, especially when an adjacent character is transcribed incorrectly by the OCR model; in that case, the sequence alignment has no cost incentive to align a gap to one over the other. Uncommon abbreviations also cause errors, since the sequence alignment would have to insert a pattern of several gaps to correctly align only the characters in the abbreviation.

[6]Short instructions or descriptions related to the liturgical process, commonly inserted between chants.

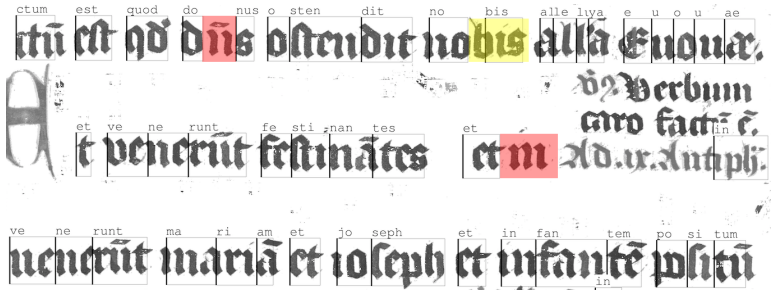

Fig. 3. The results of the transcript alignment performed on the image in Figure 1 using the sequence alignment results shown in Table I. The yellow highlight marks a partial match, while the two red highlights mark misses.

TABLE II
THE PER-SYLLABLE ACCURACY OF OUR ALIGNMENT METHOD, TESTED AGAINST MANUALLY CONSTRUCTED GROUND TRUTH.

Folio	Matches	Partial	Misses	Total	Accuracy
Salzinnes, 013r	157	10	5	172	91.2%
Salzinnes, 020v	151	9	3	163	92.9%
St. Gallen 390, 23	99	12	15	126	78.6%
St. Gallen 388, 28	310	21	20	351	88.3%
Einsiedeln, 004v	211	23	17	251	84.1%

More often, abbreviations resulted in the alignment attempting to cram extra characters into neighbouring words.

V. CONCLUSION

We have demonstrated a generalizable method of performing transcript alignment on medieval chant manuscripts. Generalizing this approach to other manuscripts requires only an OCR model that can achieve adequate per-character accuracy on the script under consideration. Compared to HMM-based models that require manual creation of ground-truth alignment data for each manuscript under consideration, this approach has the potential to save a significant amount of time in the end-to-end OMR process, especially when alignments on multiple manuscripts are desired. Future work on this method may focus on testing how well the OCR models need to perform to achieve optimal alignment results, and searching for optimal sets of parameters to fine-tune the behavior of the global sequence alignment.

REFERENCES

[1] H. Strayer, "From Neumes to Notes: The Evolution of Music Notation," *Musical Offerings*, vol. 4, no. 1, pp. 1–14, 2013.

[2] F. Cloppet, V. Églin, V. C. Kieu, D. Stutzmann, and N. Vincent, "ICFHR2016 Competition on the Classification of Medieval Handwritings in Latin Script," in *Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition*, Oct. 2016, pp. 590–595.

[3] J. A. Burgoyne, Y. Ouyang, and T. Himmelman, "Lyric Extraction and Recognition on Digital Images of Early Music Sources," in *Proceedings of the 10th International Society for Music Information Retrieval Conference*, 2009, p. 5.

[4] C. Dalitz, G. K. Michalakis, and C. Pranzas, "Optical Recognition of Psaltic Byzantine Chant Notation," *International Journal of Document Analysis and Recognition*, vol. 11, no. 3, pp. 143–158, 2008.

[5] C. M. Dinh, H. J. Yang, G. S. Lee, and S. H. Kim, "Fast Lyric Area Extraction from Images of Printed Korean Music Scores," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 6, pp. 1576–1584, 2016.

[6] J. Calvo-Zaragoza, F. Castellanos, G. Vigliensoni, and I. Fujinaga, "Deep Neural Networks for Document Processing of Music Score Images," *Applied Sciences*, vol. 8, no. 654, 2018.

[7] S. E. George, "Lyric Recognition and Christian Music," in *Visual Perception of Music Notation: On-Line and Off-Line Recognition*, S. E. George, Ed. IRM Press, 2004, pp. 198–226.

[8] A. Hankinson, J. A. Burgoyne, G. Vigliensoni, A. Porter, J. Thompson, W. Liu, R. Chiu, and I. Fujinaga, "Digital Document Image Retrieval Using Optical Music Recognition," in *Proceedings of the 13th International Society for Music Information Retrieval Conference*, 2012.

[9] V. Romero-Gómez, A. H. Toselli, V. Bosch, J. A. Sánchez, and E. Vidal, "Automatic Alignment of Handwritten Images and Transcripts for Training Handwritten Text Recognition Systems," in *Proceedings of the 13th IAPR International Workshop on Document Analysis Systems*, 2018, pp. 328–333.

[10] J. Rothfeder, R. Manmatha, and T. M. Rath, "Aligning Transcripts to Automatically Segmented Handwritten Manuscripts," in *Document Analysis Systems VII*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2006, pp. 84–95.

[11] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription Alignment of Latin Manuscripts Using Hidden Markov Models," in *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, 2011, pp. 29–36.

[12] R. Cohen, I. Rabaev, J. El-Sana, K. Kedem, and I. Dinstein, "Aligning Transcripts of Historical Documents Using Energy Minimization," in *Proceedings of the 13th International Conference on Document Analysis and Recognition*, 2015, pp. 266–270.

[13] A. Kumar and C. V. Jawahar, "Content-Level Annotation of Large Collection of Printed Document Images," in *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, 2007, pp. 799–803.

[14] G. Sadeh, L. Wolf, T. Hassner, N. Dershowitz, and D. S. Ben-Ezra, "Viral Transcript Alignment," in *13th International Conference on Document Analysis and Recognition*, 2015, pp. 711–715.

[15] S. Feng and R. Manmatha, "A Hierarchical, HMM-based Automatic Evaluation of OCR Accuracy for a Digital Library of Books," in *Proceedings of the Joint Conference on Digital Libraries*, Chapel Hill, NC, 2006.

[16] *Cantus: A Database for Latin Ecclesiastical Chant – Inventories of Chant Sources*. Directed by Debra Lacoste (2011-), Terence Bailey (1997-2010), and Ruth Steiner (1987-1996). Web developer, Jan Koláček (2011-). [Online]. Available: http://cantus.uwaterloo.ca/

[17] T. M. Breuel, "The OCRopus Open Source OCR System," in *Proceedings of the Society of Photo-Optical Instrumentation Engineers*, 2008.

[18] S. B. Needleman and C. D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

VI. Appendix

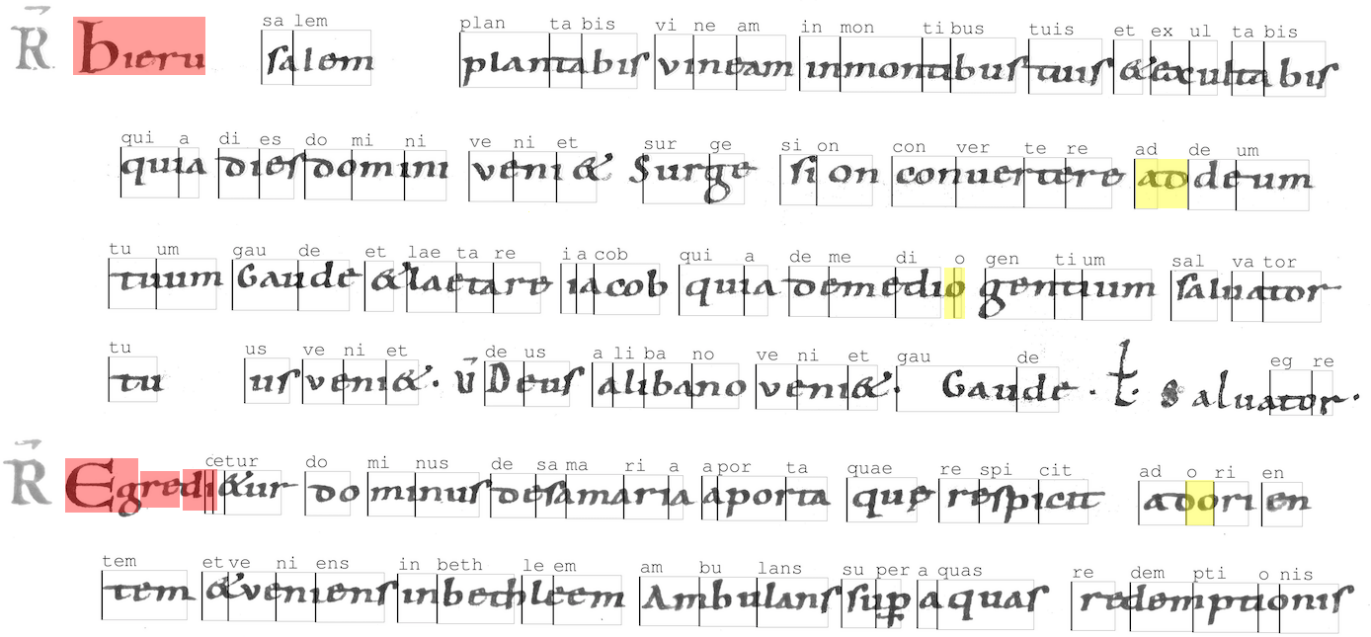

Fig. 4. An excerpt from the results of our alignment method on St. Gallen 390, folio 23. Yellow highlights on syllables mark partial matches, and red highlights on syllables mark misses.

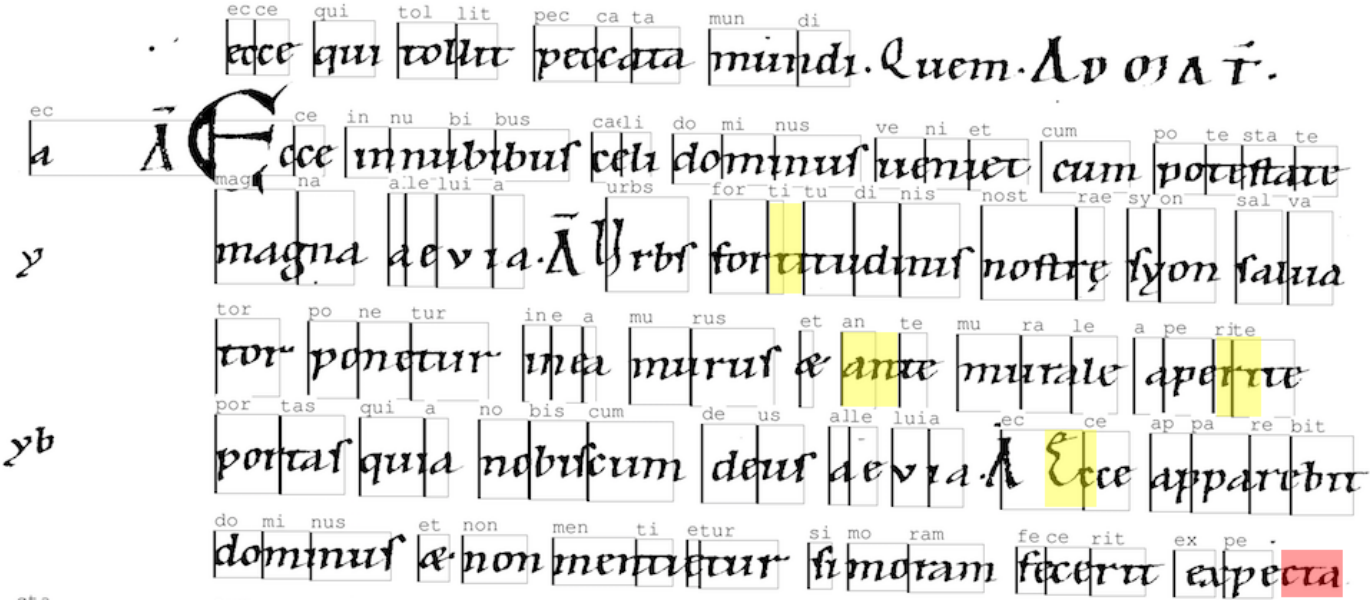

Fig. 5. An excerpt from the results of our alignment method on St. Gallen 388, folio 28. Yellow highlights on syllables mark partial matches, and red highlights on syllables mark misses.

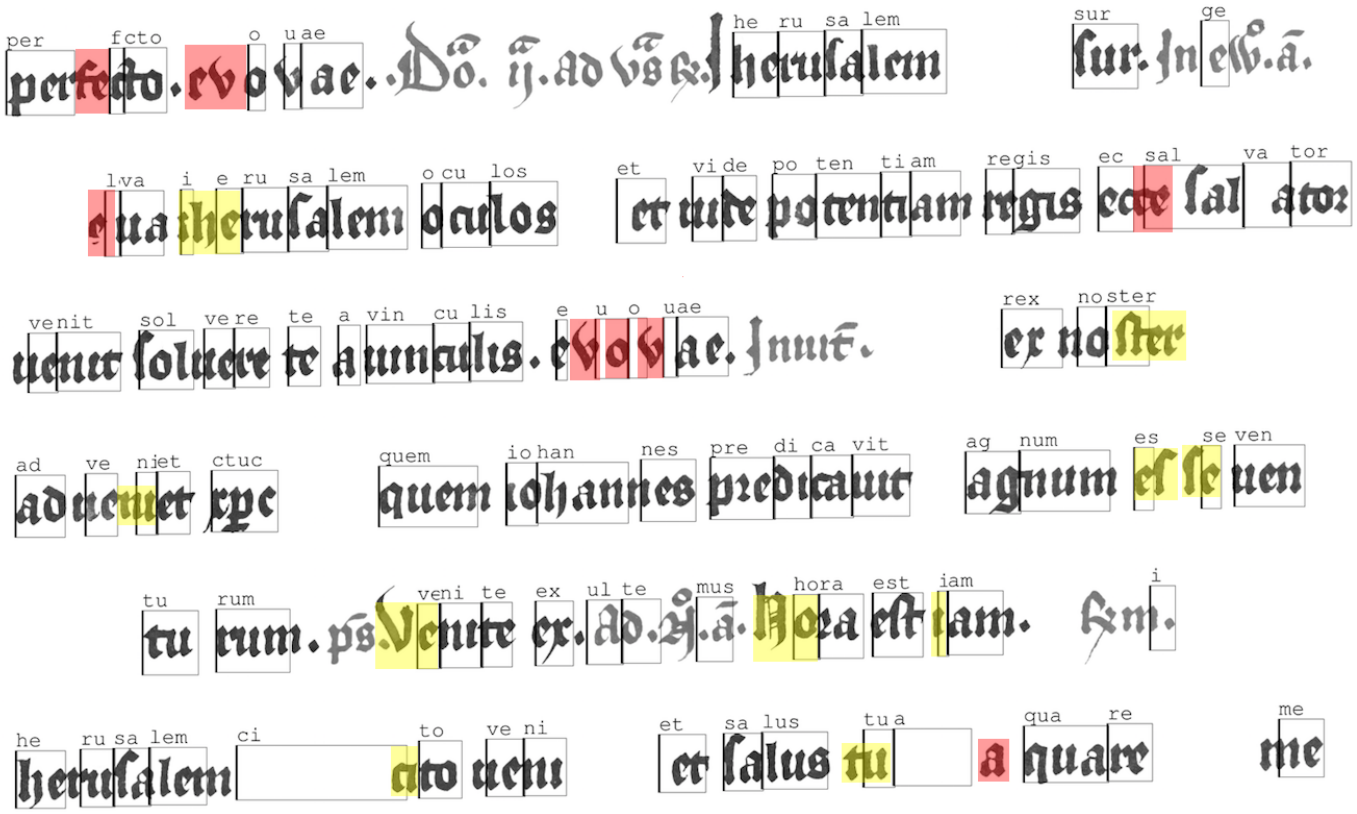

Fig. 6. An excerpt from the results of our alignment method on Einsiedeln, folio 004v. Yellow highlights on syllables mark partial matches, and red highlights on syllables mark misses.

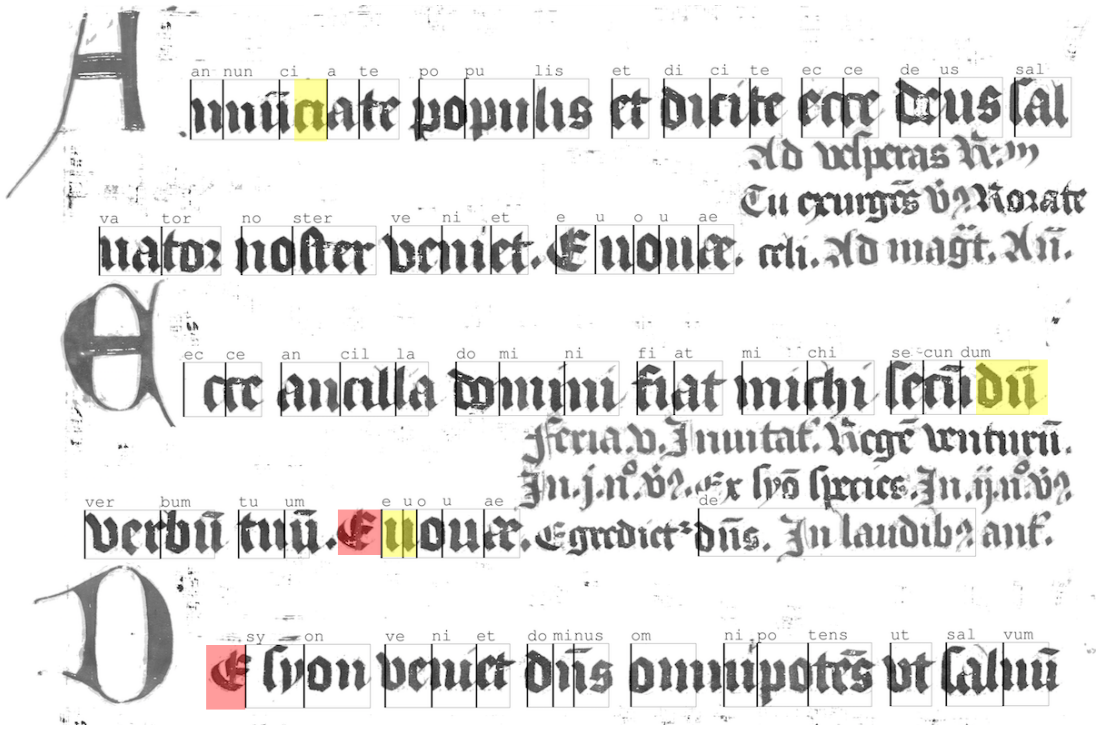

Fig. 7. An excerpt from the results of our alignment method on Salzinnes, folio 020r. Yellow highlights on syllables mark partial matches, and red highlights on syllables mark misses.