IDENTIFICATION AND CROSS-DOCUMENT ALIGNMENT OF MEASURES IN MUSIC SCORE IMAGES

Simon Waloschek, Aristotelis Hadjakos
Center of Music and Film Informatics
Detmold University of Music, Germany
{s.waloschek, a.hadjakos}@cemfi.de

Alexander Pacha
Institute of Information Systems Engineering
TU Wien, Austria
alexander.pacha@tuwien.ac.at

ABSTRACT

In the course of editing musical works, musicologists regularly compare multiple sources of the same musical piece, such as composers' autographs, handwritten copies, and various prints. For efficient comparison, cross-source navigation is essential, enabling to quickly jump back and forth between multiple sources without losing the current musical position. In practice, measures are first annotated by hand in the individual source images and then related to each other. Our approach automates this time-consuming and error-prone process with the help of deep learning. For this purpose, we train a neural network that automatically finds bounding boxes of all measures in images. A second network is trained to compute the similarity between two measures to determine if they have the same musical content and should, therefore, be linked for navigation. Sequences of outputs from the second network are matched using Dynamic Time Warping to provide the final proposal of measure relationships, so-called *concordances*. In addition to cross-source navigation, the results can be used to spot structural differences across the sources which are essential for editorial work, so that musicologists can focus more on analytical tasks.

1. INTRODUCTION

Modern musical editions are the result of a long musicological process. From the composer's manuscript to the printed music book, a musical work usually undergoes a large number of iterations and minor corrections, occasionally even substantial changes, such as striking or reworking complete parts [1]. Many of these changes are either unintentional—e.g., errors in handwritten copies, typographical errors by publishers—or generally not documented in a transparent manner. Musicologists, therefore, work on this genesis when editing a work and try to record the chronological order and causalities in their edition creation process.

The first step in this process is, therefore, the screening of the source material to identify differences between the various sources of a work. To facilitate this process, links are created between the sources so that editors can quickly switch back and forth between them. Adequate granularity of these links are usually musical measures, a feasible compromise between annotation effort and accuracy [29]. Currently, the measures of all sources are manually annotated with bounding boxes and related to each other in a very time-consuming and error-prone way.

We have automated this multi-stage process by first recognizing and sorting measures in score images (both handwritten and typeset) and then linking them according to their musical content. For this purpose, deep learning was used to develop a distance metric in an end-to-end fashion without an intermediate representation. The results can be further processed using classic alignment algorithms from the MIR community such as *Dynamic Time Warping* (DTW). While DTW-based approaches have achieved sufficient quality for practical use, audio-to-score alignment is still an active field of research [31]. Promising approaches for the synchronization of scans and sound recordings [5,6] are currently limited to monophonic and piano music and have not yet achieved sufficient accuracy for most real-world scenarios. With the contribution of this paper, we decrease a potential gap in the "audio – symbolic score – image" triangle and offer a new way for measure-accurate alignment across modal boundaries.

2. RELATED WORK

Detecting measures can be seen as a preprocessing step in Optical Music Recognition (OMR). Therefore, it was rarely singled out as a dedicated task. While Pedersoli and Tzanetakis perform document segmentation, they only distinguish between music scores and text blocks [22]. The only research we know of, that specifically addresses the automatic extraction of measures is by Vigliensoni et al. [30]. In their work, they attempt to extract measures with a traditional computer vision approach by heuristically finding all bar lines and then joining them into measures. Their approach requires human intervention for each page and straight bar lines to work well.

For retrieval of sixteenth-century musical texts, Crawford et al. [4] have recently proposed a two-step procedure. They run an OMR algorithm to obtain an intermediate format, followed by a second step that uses n-grams and minimal absent words (MAWs) to find duplicates, related texts, or parts that have the same musical material. Neural networks make such intermediate formats partly obsolete and allow for learning bimodal embeddings end-to-end as shown by Dorfer et al. [5, 6], who correlate the scanned music score with a sound recording. For this purpose, synchronization was considered either a reinforcement learning problem [6] or a metric learning problem [5]. In the metric learning approach, Dorfer et al. use the *pairwise ranking loss*—also known as *triplet loss* [26]—that draws triplets from a dataset consisting of an anchor, a positive example (picture fits the audio) and a negative example (picture does not fit the audio). This loss function creates an embedding, where images and audio with the same content are appear close together, while non-matching images and audio are placed relatively far apart. Their approach has successfully been used before in other application domains, such as facial recognition [26]. We resort to a similar cost function for metric learning (see section 4.2).

As the basis for our detection, we use a convolutional neural network (CNN). While CNNs are currently an active field of research for OMR, the most influential approaches come from the research area of computer vision. They are used for many tasks, including image recognition, semantic segmentation, object detection, and instance segmentation. R-CNN [9] performs object detection by analyzing a large number of heuristically generated region proposals that are classified into background or one of the classes of interest. Additionally, the bounding box is refined with regression. R-CNN uses a CNN that extracts features for object detection. These features are used in a downstream SVM for classification and regression. Faster R-CNN [23] improves the process by incorporating both the region proposal step as well as the classification and regression into the architecture of the neural network.

CNN-based computer vision approaches are largely transferable to OMR and actively used for Music Information Retrieval: Gallego and Calvo-Zaragoza are using auto-encoders to remove staff lines [8]. Pacha et al. compare various CNN-based approaches for detecting music symbols in scores [21]. CNNs can also be used for semantic segmentation for staff-line removal, music and text separation as well as for layout analysis as shown by Calvo-Zaragoza et al. [3]. Using U-Nets [25], Hajic et al. do semantical segmentation of handwritten music [10]. Pacha and Calvo-Zaragoza recognize musical objects in mensural notation using region-based CNNs [20]. By learning energy levels that are used as inputs to a watershed algorithm, Tuggener et al. recognize music symbols [28]. In addition to the energy levels, the network also predicts class labels and bounding boxes. And finally, Calvo-Zaragoza and Rizo use convolutional recurrent neural networks trained with a Connectionist Temporal Classification (CTC) loss to recognize musical symbols in monophonic music scores [2]. To simulate non-ideal image conditions, they artificially distort the images.

3. DATA & ANNOTATIONS

The success of Deep Learning approaches largely depends on the amount and diversity of data used during training. Since no dataset of sufficient size was available for measure recognition or the concordance task, we created a large dataset ourselves in cooperation with musicologists and professional musicians.

Our dataset contains measure annotations that were created manually by musicologists for digital music editions. In most cases, the image sources are high-resolution scans of facsimiles, occasionally supplemented by early music prints and PDFs exported directly from music engraving software. Due to an imbalance between handwritten and typeset scores, we additionally obtained scores from the *IMSLP/Petrucci Music Library* while paying attention to varying image quality, the used engraving mechanism as well as diverse musical content. We complemented our collection with 140 pages from the MUSCIMA++ dataset[1] [7, 11].

Our data collection has a total of 8 251 pages with 81 124 annotated measures. The distribution according to engraving type and the number of systems per page is given in Table 1. One category is particularly overrepresented: handwritten music scores with just one system per page because of a large quantity of full orchestral scores from operas by Carl Maria von Weber. Book covers, text pages, and empty pages have zero systems.

Systems per page	**Pages per engraving type**	
	Handwritten	Typeset
0	413	113
1	5627	932
2	175	553
3	122	175
4 or more	102	39
Total pages	6439	1812

Table 1. Overall distribution of the dataset used.

The accuracy of the measure annotations varies. Since the exact boundaries are not relevant for musicologists, they were recorded only roughly. That is why many bounding boxes contain small overlaps with adjacent measures as shown in Figure 1.

To annotate the measures in the individual pictures, the Android app *Vertaktoid*[2] [18] was used. It allows to conveniently draw bounding boxes for all measures with a pen directly on the tablet screen. The results can then be exported to the MEI format [24] and used as ground truth training data.

Data coming from digital music editions are partly provided with concordance annotations between the measures.

[1] The measure annotations are published as separate dataset at https://apacha.github.io/OMR-Datasets/#muscima

[2] https://github.com/cemfi/vertaktoid

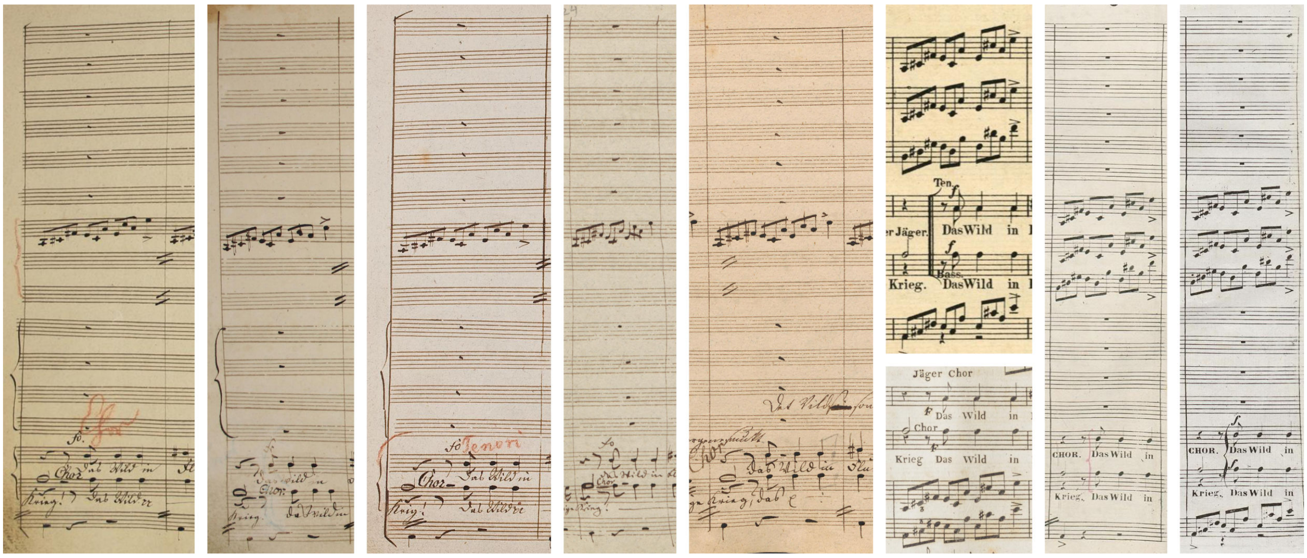

Figure 1. Examples of cropped measures originating from different sources of the same work. All measures represent the same musical position, i.e. the same measure, within the work, but are in part extremely diverse in terms of instrumentation, graphic representation and also image resolution.

4. ALIGNING MEASURE SEQUENCES

Our proposed solution for the given task can be split into three individual parts. First, we have to find the bounding boxes of all measure in the score images. Then we need a metric in order to compute the similarity between two given measure in terms of musical content. And finally, we have to compute actual concordances for multiple sources of the same music.

4.1 Optical Measure Recognition

For automatically detecting measures in complete music scores, we propose a machine-learning approach with deep convolutional neural networks and a Faster R-CNN detector [23]. Faster R-CNN has been shown to work well in a range of situations, including detecting music objects [21]. In this case, there is just one class of objects that needs to be detected, and the objects typically cover large portions of the entire image with little overlap. Our implementation is based on the TensorFlow Object Detection API framework [14] and freely available online [3] .

We split the dataset randomly into 80% for training, 10% for validation, and 10% for testing. To avoid a bias toward scores with just one system, we sample the images equally from the ten categories depicted in table 1. The only exception are images without systems which are sampled only half as often as the other categories.

We tested the three different backbones, ResNet50, ResNet101 [13], and Inception-ResNet-V2 [27] and restricted ourselves to these to enable transfer-learning by initializing the networks with weights trained on ImageNet which generally improves the learning process, especially at the beginning. Input images are resized to be no longer than 1024 pixel on the longest edge. The Intersection over

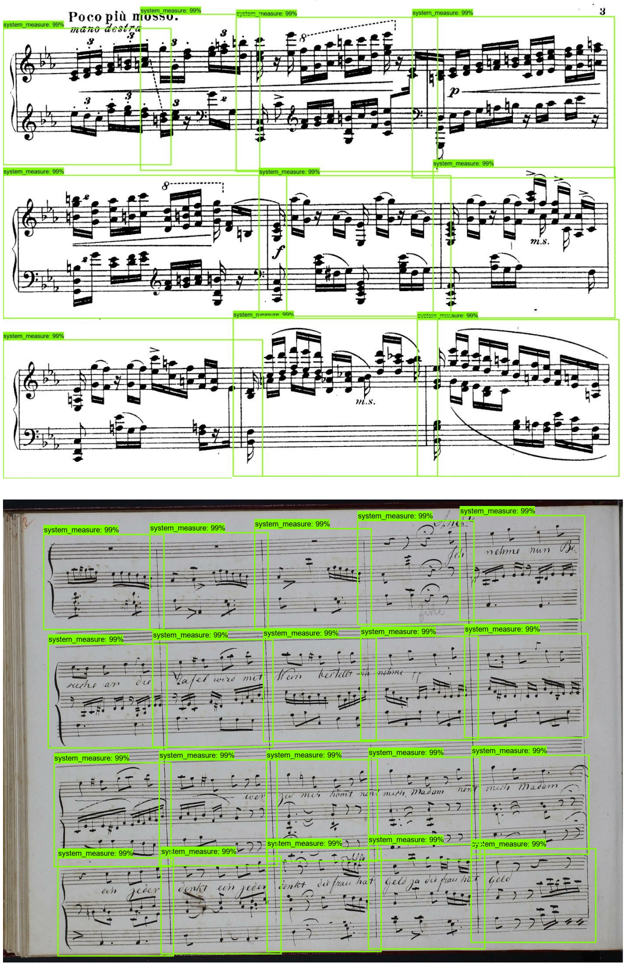

Figure 2. Two samples of the detection results. Measures are detected robustly in typeset and handwritten scores without the need for preprocessing the images.

[3] https://github.com/OMR-Research/MeasureDetector

Union (IoU) measures how well two bounding boxes overlap. If two predictions are very close, non-maximum suppression filters the box with the lower score. The IoU threshold is set to 0.6 and a maximum of 600 objects are detected per image. These parameters are derived from statistical analysis of the entire data set and cover $> 99.99\%$ of the dataset.

We evaluated the optical measure detection with the commonly used average precision (AP) metric, as defined for the COCO detection challenge [15]. It produces a single number that measures how well objects were detected. A detection is considered a match with the underlying ground truth if the IoU is above a certain threshold. The trained models achieve very good results with 78.7% AP (IoU=0.5:0.95) on the test set for the top-performing model with Inception-ResNet-V2 [27] backbone. A few samples of the detection output are depicted in Figure 2.

Given that the measure recognition step does not necessarily return the measures of a page in the musically correct order, we sort them according to the measure numbering rules outlined by Mexin et al. in [18].

4.2 Metric Learning

Now that the scans of all scores are divided into individual measures, they have to be compared with each other to identify equivalent measures. Again, a deep learning approach is used to learn such a musical similarity metric between two measures directly from the images. The neural network is trained to compute an embedding for measure images so that similar measures are placed in the proximity of one another in the embedding space. This allows for convenient comparison of two measures by computing their distance, e.g., using the L^2 norm.

The idea is based on *triplet loss* [26]: A pair of equivalent measure images from two different sources is drawn from the list of concordances. We will call them the *anchor* image and the *positive* image. Additionally, a *negative* measure image is drawn from the same source as the positive image, serving as a counterexample, i.e. having no musical relation to the anchor or the positive measure image. Each of these three images is fed separately into the same neural network, resulting in three k-dimensional vectors. The loss function is defined as

$$\mathcal{L} = max(d(f^a, f^p) - d(f^a, f^n) + \alpha, 0) \tag{1}$$

with f^a, f^p, and f^n being the resulting vectors from the network f for the three images and a distance measure d. Training with this loss function minimizes the distance from the anchor to the positive image while maximizing the distance between the anchor and the negative image. The additional margin α defines how far away the least dissimilarity should be. Finally, the surrounding $max(...)$ function ensures that the loss never gets negative.

We chose ResNet50 as the base network and replaced the usual final average pooling and classification layers by a fully connected layer with k-dimensional output. (Other CNN-based networks used for computer vision would most likely work comparably well.) All measure images are resized to 512×512 pixels but the original width and height information is also passed to the network as additional input.

The success of the used loss function depends heavily on the sampling strategy for the image triplets as discussed by Wojke and Bewley in [32]. In our context, there are three specific problems in the dataset:

1. A randomly sampled negative image might accidentally have the same musical content as the two other images. Those cases are not covered in the concordance dataset since not all measures with equal content have to be linked together.
2. Intuitively, it seems beneficial to take the previous or subsequent measure of the positive sample as the negative measure with the goal of enhancing the contrast between them in terms of increased distance in the embedding space. This would make adjacent measures more distinguishable. But again, the chance of these measures having the same content is higher compared to random sampling.
3. Especially handwritten sources sometimes exhibit excessive use of measure repeats and other abbreviations as can be seen in the left part of Figure 1. Such symbols are meaningless if their immediate context is not given.

The first two problems could be solved by manually adding all measures with the same content to the list of concordances. Given the amount of images, we decided against doing so and rely on rare collisions thanks to the large number of data. We also discarded the (perfectly valid) idea of looking at adjacent measures to form the triplets.

The third problem—presence of measure repeats and abbreviations—has a direct impact on the appropriate choice of the distance metric d in our loss function; When using triplet loss, it is common practice to normalize the embedding vectors. This constraint puts all embeddings on a k-dimensional hypersphere, leading to some advantages for further processing (see [26]). Furthermore, *cosine distance* is often used to calculate the distances. Both decisions make it impossible to get an embedding vector that is equally distant to all other possible vectors. This very property, however, characterizes the meaning of measure repeats if no context is given. We, therefore, opted for no vector normalization and chose the L^2 norm as our distance metric, resulting in

$$\mathcal{L} = \sum_{i=1}^{N} \left[\|f_i^a - f_i^p\|_2 - \|f_i^a - f_i^n\|_2 + \alpha \right]_+ \tag{2}$$

for a training batch with size N. To speed up training and ensure fast convergence we select triplets that violate the following constraint:

$$\|f_i^a - f_i^p\|_2 + \alpha < \|f_i^a - f_i^n\|_2 \,. \tag{3}$$

This filter step is performed for each batch during training and makes sure that only those triplets are used that significantly contribute to the learning process. It also prevents the network from overfitting.

4.3 Concordance Computation & Manual Adjustments

Given the embedding vectors for all measures of each source of a musical work, we can compare two sources by computing the distances between all measures from one source to the other. The resulting similarity matrices can then be used for *dynamic time warping* (DTW) as described by Müller in [19] to get an alignment path between the sources as shown in Figure 3.

We implemented the canonical DTW algorithm without any noteworthy modifications to the core. Allowed step sizes inside the similarity matrix during path computation are $(0,1)$, $(1,0)$, and $(1,1)$. It rarely happens that a measure gets divided into two parts at system or page breaks, so we penalized steps along a single axis by a factor of 2 to slightly enforce one-to-one mappings of the measures.

The quality of the alignment was evaluated using a dataset with two sources and given ground truth concordances as outlined in Table 2. We have decided in favor of this particular dataset because it offers several challenges that occur only rarely in other works:

Split measures: Some measures are split into two parts at page breaks. Therefore, one measure of source A maps to two other measures of source B.

Completely different sections: An entire part of the piece was replaced in source B. Finding the "correct" concordance is impossible.

Additional parts: Source B contains a 16-measure Aria that is not present in the other source.

Missing measure annotations: We also intentionally removed measures from source A to simulate annotation errors.

	Pages	**Measures**
Source A (typeset)	250	3098
Source B (handwritten)	532	3176
Total	782	6274

Table 2. Structure of the evaluation dataset.

In the MIR community, DTW is often used to synchronize audio and/or symbolic score sources with each other [12]. The time resolution of the features in such scenarios is usually in the range of several dozen milliseconds. Deviations in the alignment path are therefore undesirable, but can often be neglected as long as they do not exceed certain limits. In our context, however, any deviation from the ground truth marks a significant error. We took this into account and defined a very simple score for the overall performance:

$$score = 1 - \frac{\text{Number of } (x,y) \text{ pairs from} \\ \text{alignment not in ground truth}}{\text{Total number of concordances} \\ \text{in ground truth}} \quad (4)$$

Our evaluation showed 14 errors in relation to 3079 concordance pairs, resulting in a score of **99.545%**.

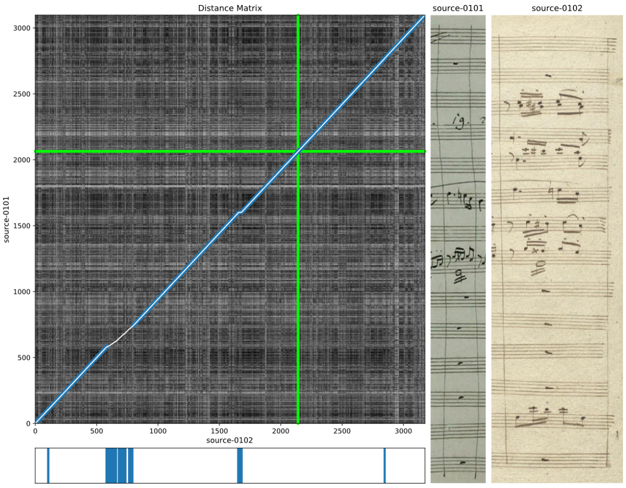

Figure 3. Interface for inspecting the computed measure concordances. The alignment (white) and ground truth (blue, only available in evaluation dataset) are plotted over the currently visible part of the similarity matrix. Measures of both sources (right) can be compared by moving a cursor within the matrix (green crosshair). A plot at the bottom indicates potentially interesting positions.

As pointed out, the remaining 0.455% error rate still present a non-negligible problem. Therefore, we developed an interface for manual adjustments to the alignment. Apart from being able to quickly compare the measures from two sources as shown in Figure 3, users can define points in the similarity matrix that have to be part of the alignment path. Each of these points splits the matrix into two parts and computes the warping path for each part individually, ensuring that either the beginning or end of the path matches the desired point. An event plot at the bottom of the matrix helps to identify regions with potential errors by showing where the alignment path is not diagonal, i.e. taking a step in $(0,1)$ or $(1,0)$ direction.

The mentioned obstacles for correct alignment have been handled successfully by either resulting in a correct alignment or—in case of substantial structural differences—indicating a problem that cannot be solved without human intervention by marking these parts in the plot below the similarity matrix.

This alignment and adjustment step has to be repeated for each source in regard to a *master source* of choice. The corrected alignment data can then finally be imported into the tools used by musicologists for their editorial work.

5. CONCLUSIONS AND FUTURE WORK

In this paper we proposed an approach to automate the tedious task of annotating and linking measures in heterogeneous score images, thereby allowing for cross-source navigation between measures without losing the current musical position. We used deep learning to find bounding boxes of measures in score images, learned a distance metric for measures, and used that to align measures from various sources, effectively linking equivalent musical positions across sources. The evaluation showed that our approach is feasible and solves a real-world problem while still retaining complete flexibility in case editors need to make manual adjustments, thanks to an interactive correction tool.

The presented solution still does not cover all possible situations that might occur in the editorial process. If the measure sequences to be compared have a different order, the alignment fails for these parts if not completely. We will address this specific problem in the future by identifying such passages and proposing reasonable re-ordering.

Having a musically meaningful distance metric for measures also allows closing the gap between score images and symbolic scores. The latter can be rendered with suitable engraving software and divided into individual measures, followed by the steps of our alignment pipeline. Since audio can also be rendered from symbolic scores, alignments between all three modalities are possible.

Another interesting application of our distance metric is the ability to visualize datasets in image fields as shown in Figure 4. Using dimensionality reduction algorithms such as T-SNE [16] or UMAP [17], the measures are positioned such that musically similar measures appear proximate to one another, giving new insight into a musical piece but also into the inner workings of the distance metric. For example, the visualization shows that measure repeats are placed almost in the center, indicating that their learned embedding retains the musical property of being close to basically every other measure in the embedding space.

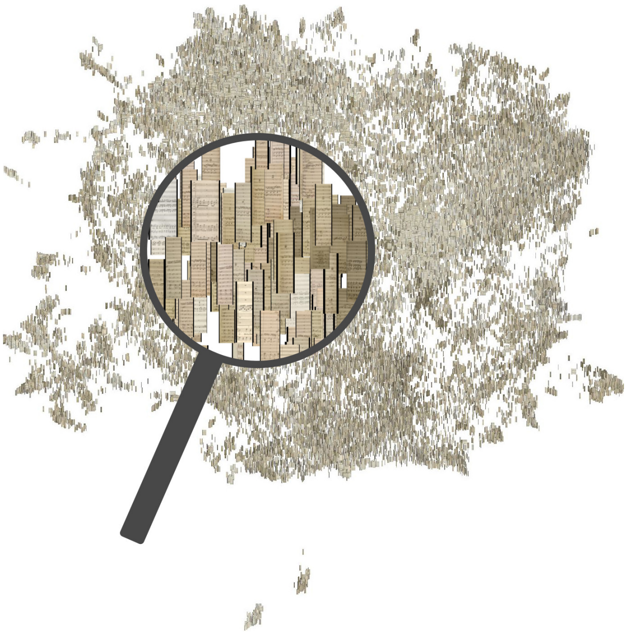

Figure 4. 46 344 measure images from 15 different sources of the same piece are projected into a two-dimensional manifold with the UMAP algorithm. The map is interactively zoomable.

6. REFERENCES

[1] Benjamin W. Bohl, Axel Berndt, Simon Waloschek, and Aristotelis Hadjakos. Dem Igel Sitte lehren... Musikedition: von der digitalen Verfügbarkeit zur aktiven Nutzung. In Kristina Richts and Peter Stadler, editors, *„Ei, dem alten Herrn zoll' ich Achtung gern'“ – Festschrift für Joachim Veit zum 60. Geburtstag*, chapter 12, pages 141–163. Allitera Verlag, Munich, Germany, 2016.

[2] Jorge Calvo-Zaragoza and David Rizo. Camera-primus: Neural end-to-end optical music recognition on realistic monophonic scores. In *19th International Society for Music Information Retrieval Conference*, pages 248–255, Paris, France, 2018.

[3] Jorge Calvo-Zaragoza, Gabriel Vigliensoni, and Ichiro Fujinaga. A machine learning framework for the categorization of elements in images of musical documents. In *3rd International Conference on Technologies for Music Notation and Representation*, A Coruña, Spain, 2017. University of A Coruña.

[4] Tim Crawford, Golnaz Badkobeh, and David Lewis. Searching page-images of early music scanned with OMR: A scalable solution using minimal absent words. In *19th International Society for Music Information Retrieval Conference*, pages 233–239, Paris, France, 2018.

[5] Matthias Dorfer, Jan Hajič jr., Andreas Arzt, Harald Frostel, and Gerhard Widmer. Learning audio–sheet music correspondences for cross-modal retrieval and piece identification. *Transactions of the International Society for Music Information Retrieval*, 1(1):22–33, 2018.

[6] Matthias Dorfer, Florian Henkel, and Gerhard Widmer. Learning to listen, read and follow: Score following as a reinforcement learning game. In *19th International Society for Music Information Retrieval Conference*, pages 784–791, Paris, France, 2018.

[7] Alicia Fornés, Anjan Dutta, Albert Gordo, and Josep Lladós. CVC-MUSCIMA: A ground-truth of handwritten music score images for writer identification and staff removal. *International Journal on Document Analysis and Recognition*, 15(3):243–251, 2012.

[8] Antonio-Javier Gallego and Jorge Calvo-Zaragoza. Staff-line removal with selectional auto-encoders. *Expert Systems with Applications*, 89:138–148, 2017.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, 2016.

[10] Jan Hajič jr., Matthias Dorfer, Gerhard Widmer, and Pavel Pecina. Towards full-pipeline handwritten OMR

with musical symbol detection by u-nets. In *19th International Society for Music Information Retrieval Conference*, pages 225–232, Paris, France, 2018.

[11] Jan Hajič jr. and Pavel Pecina. The MUSCIMA++ dataset for handwritten optical music recognition. In *14th International Conference on Document Analysis and Recognition*, pages 39–46, Kyoto, Japan, 2017.

[12] Yun Hao. Real-time audio to score alignment (a.k.a score following). `https://www.music-ir.org/mirex/wiki/2019:Real-time_Audio_to_Score_Alignment_(a.k.a_Score_Following)`, 2019.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recogntiion (CVPR)*, pages 770–778, 2016.

[14] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014.

[16] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[17] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[18] Yevgen Mexin, Aristotelis Hadjakos, Axel Berndt, Simon Waloschek, Anastasia. Wawilow, and Gerd Szwillus. Tools for annotating musical measures in digital music editions. In *14th Sound and Music Computing Conf. (SMC-17)*, Espoo, Finland, 2017. Aalto University.

[19] Meinard Müller. *Information Retrieval for Music and Motion*. Springer-Verlag, Berlin, Heidelberg, 2007.

[20] Alexander Pacha and Jorge Calvo-Zaragoza. Optical music recognition in mensural notation with region-based convolutional neural networks. In *19th International Society for Music Information Retrieval Conference*, pages 240–247, Paris, France, 2018.

[21] Alexander Pacha, Jan Hajič jr., and Jorge Calvo-Zaragoza. A baseline for general music object detection with deep learning. *Applied Sciences*, 8(9):1488–1508, 2018.

[22] Fabrizio Pedersoli and George Tzanetakis. Document segmentation and classification into musical scores and text. *International Journal on Document Analysis and Recognition*, 19(4):289–304, 2016.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*, pages 91–99. 2015.

[24] Perry Roland. The music encoding initiative (MEI). In *1st International Conference on Musical Applications Using XML*, pages 55–59, 2002.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015.

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[27] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017.

[28] Lukas Tuggener, Ismail Elezi, Jürgen Schmidhuber, and Thilo Stadelmann. Deep watershed detector for music object recognition. In *19th International Society for Music Information Retrieval Conference*, pages 271–278, Paris, France, 2018.

[29] Joachim Veit and Kristina Richts. Current status and perspectives of MEI usage in musicology and in libraries. *Bibliothek Forschung und Praxis*, 42(2):292–301, 2018.

[30] Gabriel Vigliensoni, Gregory Burlet, and Ichiro Fujinaga. Optical measure recognition in common music notation. In *14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[31] S. Waloschek and A. Hadjakos. Driftin' down the scale: Dynamic time warping in the presence of pitch drift and transpositions. In *19th International Society for Music Information Retrieval Conference*, Paris, France, 2018.

[32] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 748–756. IEEE, 2018.