Title

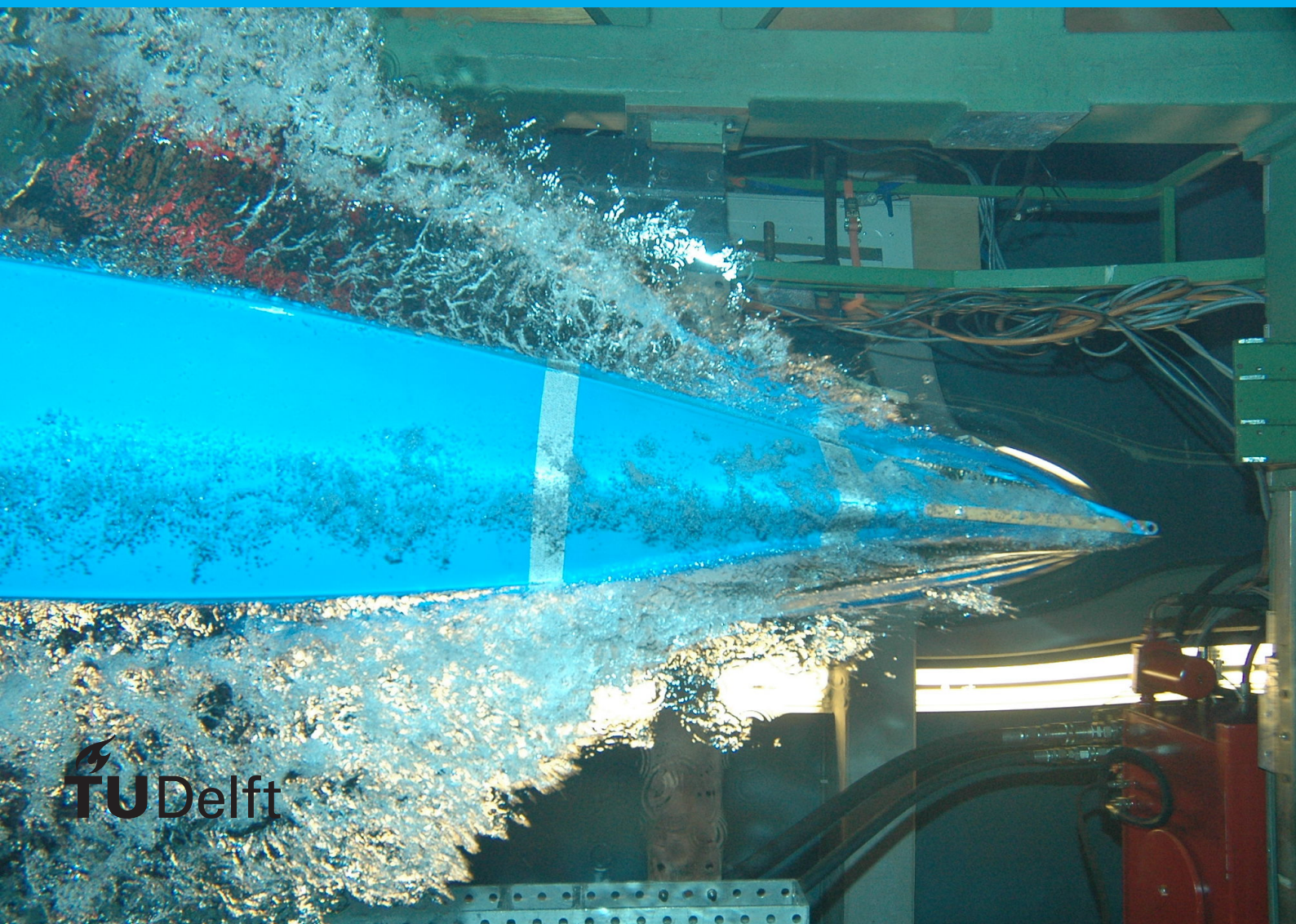
;

Optional subtitle

M. Meuleman

Cover Text
possibly
spanning multiple lines

ISBN 000-00-0000-000-0



Title

Optional subtitle

by

M. Meuleman

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday January 1, 2013 at 10:00 AM.

Student number:	4375629	
Project duration:	March 1, 2012 – January 1, 2013	
Thesis committee:	Prof. dr. ir. J. Doe,	TU Delft, supervisor
	Dr. E. L. Brown,	TU Delft
	Ir. A. Aaronson,	Acme Corporation

This thesis is confidential and cannot be made public until December 31, 2013.

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Abstract

Preface

Preface...

M. Meuleman
Delft, January 2013

Contents

1 Introduction

1.1. Optical Music Recognition

Music has always been communicated in two ways: aural transmission, where music is played or performed and people can listen to it, and written transmission, where music is formalized in a document. These written formats come in many forms, one of which is western musical notation. These documents, mostly referred to as (music) scores, used to be handwritten, but were later printed on paper and can nowadays be found in digital format on computers, mostly as images or scans. Quite some effort has been taken to collect all of these digital scores and store them in central places or make them available to the public where possible. An example is the IMSLP library, where vast amounts of scores for classical music in the public domain are available. Collections such as these give rise to many opportunities, such as affordable scores and sheet music for musicians, or easy access to musical data for researchers. The formats in which these scores are stored do have their limitations however. When text is contained in PDF documents (or many other document forms for that matter), the text is recognized by the computer as text, meaning it can be searched, edited and reformatted. For music this is not the case. Scanned music scores are simply stored as images in PDF documents, meaning that none of the applications for text can be applied to these scores. This means that a lot of applications that the digital age provides cannot yet be applied to music scores and as such there is a huge gap for potential left open.

The field of Optical Music Recognition (OMR) works toward solving this problem by finding methods that can translate the scanned formats to a format that can provide semantic meaning of the music scores to computers, so that these aforementioned applications can become available to the written music domain. Translating text from an image to computer-readable text has long been solved in the field of Optical Character Recognition (OCR)[citation needed:]. Its musical counterpart however remains for a large part unsolved. A lot of this is due to the more complex structures we see in music scores. Where reading text is a single dimensional operation, classifying each character in a line of text, reading music is a two dimensional operation, where the dimensions are time on the horizontal axis and pitch on the vertical axis. Additionally, in music different symbols can be stacked on top of one another to signify multi-tones or chords, or can be connected horizontally to one another to improve readability for musicians. Two connected notes can also be translated relative to each other, meaning the connection can become stretched or tilted. Also there are symbol that are often small in comparison to the notes, but have significant meaning, such as dots, dashes and accents. Draw some examples All of these illustrated cases can be combined and many of these combinations occur frequently in music scores, making it impossible to apply simple heuristics when translating music scores to a computer-understandable format. Many approaches have been taken to apply machine learning and end-to-end learning [citation needed:]to this problem. There are promising results in there, but for a lot of use cases the translated music has to almost perfectly correspond to the original.

This is especially true for scores of classical music. Many classical scores are complex, and the scanned format of many of these scores are often hard to read because they are old editions or scanned in low quality. This, combined with the fact that there is a lot of classical music available in the public domain makes digitization in this computer-understandable format especially attractive for classical music.

1.2. TROMPA

TROMPA (Towards Richer Online Music Public-domain Archives) is an international organization of scientists and scholars that push towards this goal of making more applications available to the music domain and increase engagement with classical music and music scores[citation needed:]Some more info about TROMPA. A group of researchers at Delft University of Technology is part of this organization. This group focusses on translating the scans of music scores into the MEI (Music Encoding Initiative) format, an XML-like format for music scores that allows computers to give semantic meaning to these scores[citation needed:]. This process is far from perfect, and therefore research is currently ongoing in crowd computing solutions.

1.3. Crowd Sourced solutions

Maybe the crowd can help out a bit

1.3.1. Struggles for crowd sourcing

Repetition legitimises Repetition legitimises Repetition legitimises

But maybe see if we can do without transcribing the same measure 100 times

Also, if we could measure the complexity of a measure somewhat easily, this could help with prioritizing tasks

1.4. Research questions

Main research question

> How can we apply standard image processing methods to improve the process of transcribing orchestral music scores?

1.5. Main contributions

1.6. Outline

2

Related work

2.1. OMR

- OMR is an extension of OCR, but with many more complex problems. The 2 dimensional nature of OMR, in contrast to the 1 dimensional nature of OCR, can cause troubles in accurately transcribing music [citation needed:]. Besides this, there is a vast collection of symbols possible in sheet music, with a very uneven occurrence distribution, resulting in a large number of false positives for the symbols that occur infrequently (Chen, Stolterman, 2016). There is also a large portion of written music that adheres strongly to a fixed ruleset, advocating for a rule-based recognition system, which is offset against the small but long tail of exceptions on these rules, making any rule-based recognition system inaccurate in too many cases.

- Throughout the last two decades a generally accepted OMR pipeline has been devised, first started by Bellini et al. (2004??) and later refined by Rebelo et al. (2012). Various stages of this pipeline remain unsolved.

- To find solution for these open issues, two trends can be seen over the last decade or so: deep learning and human-aided recognition. (include lots of sources for both pls)

- In deep learning, promising results are shown for certain stagese of the pipeline (Gallego, Calvo-Zaragoza, 2017), (Pacha, Calvo-Zaragoza, Hajic, 2019) as well as for a full-pipeline end-to-end learning approach (Calvo-Zaragoza, Valer-Mas, 2017), (Wel, Ullrich, 2017). The main difficulty with the deep learning approach is the cost of collecting data. There is only a small collection of datasets available, getting more of these is expensive.

- Meanwhile human-aided solutions have also started to get some traction. Examples are the Allegro system (Burghardt, Spanner, (2017), human-in-the-loop systems on measure level (Chen, Stolterman, 2016) and on symbol level (Chen, Raphael, 2016)

2.2. Crowd computing

3

Data description

As described before, for this work we want to focus on music scores for larger ensembles. In this section we will describe the requirements for this data and the way the data is collected.

3.1. Requirements

The requirements of the data are somewhat foggy. [1]

3.2. Aquisition

Aquisition was done through IMSLP

4

Methodology

In this chapter we will lay out the work that was done for the measure detector. First we will cover the general structure of music scores and the assumptions made that follow from that structure. After that we will cover the implementation of the measure detector, followed by the evaluation of results.

4.1. Music scores

To understand the steps that need to be taken when implementing a measure detector, we first describe the general structure of a music score in this section. A music score consists of

4.2. Measure detector

4.2.1. Previous work

The segmenter currently embedded in the OMR pipeline is the one taken from the work of Waloschek, Hadjakos and Pacha [1]. Their approach consisted of manually annotating measures on pages of orchestral scores, defining a distance metric between annotated measures and training a CNN to detect measures in new input data. There are a few shortcomings to this approach. First of all, this model is far from perfect and makes too many mistakes to be used in a reasonable manner in this OMR pipeline. Even on pages that contain high quality scans and straight barlines, the detection is not perfect and therefore requires post-processing, see Figure 1 for two examples. Second, the measures that are detected are restricted to separation over time only. This means that when applying this to music scores with multiple instruments or voices -which is common in orchestral music, choir music, or piano scores- the segmenter will only segment horizontally, but leaves the different voicings grouped together in the same block. Besides this, the model is fixed and cannot be easily improved upon. Retraining the model to overcome the mistakes it currently makes would require manual annotation of these pages which is a very costly process. Finally, this model is relatively slow compared to other approaches. Figure 1: Two examples of errors of the CNN method. In the first example we see that a large part of the entire page is classified as a single measure, in the second examples we see that smaller subsections of measures are detected as measures.

4.2.2. Image processing based segmentation

To overcome these issues, another approach is currently under development. The main problem with the CNN approach is the supervised nature of it, requiring annotations which are costly to come by. Our algorithm takes an image processing approach, removing the requirement of annotated examples. The structure of a page is analysed top-down, first separating systems, which are subsequently segmented into vertical blocks, which are segmented into measures, as can be seen in Figure 2. This process is explained here in a little more detail.

Before any segmentation is done, some preprocessing is performed on the page. The contrast of the page is maximized, after which the page is binarized. Next, any rotations in the page that might have occurred due to scanning of the original score is rectified. These are all standard preprocessing techniques in the image processing field.

The first segmentation is on the system level. There is a high level of separation between systems on a page (there are no connected components between them), which allows for binary propagation. This binary propagation will fill up each of the systems, allowing for easy detection of one or a few large blocks on the page, each of which is a system. Figure 2: Structure of a music score

Next, each of these systems are segmented into vertical blocks, making use of the horizontal intensity profile of the system. Due to the structure of music scores, each system will have barlines that span almost all of the systems height. These vertical lines are detected as peaks in the intensity profile, finding these peaks will provide the location where the system should be segmented into separate blocks. After this step, the output will be similar to the previous segmenter, which was only trained to segmented blocks.

After the blocks have been segmented, each of these blocks are segmented into measures. There are currently two methods for this correction: a smallest intersection method, and a largest region method.

Smallest intersection method

The smallest intersection method works in two parts: first a set of baselines per system are established, and then these baselines are corrected for each individual block in that system. The baselines are established from the vertical intensity profile of the system. Each of the bars consist of 5 small peaks, corresponding to the 5 lines in a bar. These 5 peaks grouped together can be detected as one broader peak. The baselines are set as the middle points between each of these detected peaks. Next the baselines are corrected on a per block basis. This second step is necessary, since in the scores it can occur that notes and related annotations can cross an established baseline into the “territory” of measures above or below it. This crossing over can change per block, and therefore a correction per block is necessary. This second step finds within a predefined distance from the baseline the points with the lowest value in the intensity profile. These points indicate that when segmenting at these points, the least amount of information, indicated by white pixels, will be segmented, and therefore these points should be considered as good segmenting points. When finding these candidate points a small margin from the minimum intensity value is taken, and the point closest to the baseline is chosen as the segmenting point.

Largest region method

The largest region method divides the region in between two peaks into regions, where regions are separated from each other by intensity values above a certain threshold. The largest of these regions is taken, as this indicated the largest part between two measures where there is little to no information. The middle of this region is chosen as the segmenting point. In Figure 3 two examples are given of segmented pages using the image processing approach with the largest region method.

Figure 3: Two examples of score pages segmented with the largest region method. Evaluation The drawback of this image processing method is that performance of the segmenter is hard to evaluate. The CNN based method had in this respect the advantage of having annotated examples. Unfortunately, these examples are not applicable to evaluation of the image processing approach, since that segments a level deeper; where the CNN approach segments at the block level, the image processing approach segments at the measure level. Currently evaluation has to be done by hand because of the lack of a corpus of segmented music scores. A tool is under development that can hopefully make this process more efficient and over time can hopefully help to create a corpus of segmented measures.

[1] Waloschek, S., Hadjakos, A., & Pacha, A. (2019). Identification and Cross-Document Alignment of Measures in Music Score Images. In ISMIR (pp. 137-143).

Bibliography

[1] David Bainbridge and Tim Bell. The challenge of optical music recognition. Computers and the Humanities, 35:95–121, 2001. ISSN 00104817. doi: 10.1023/A:1002485918032. URL https://link.springer.com/article/10.1023/A:1002485918032.