Simplified Optical Music Recognition
TNM034 - Advanced Image Processing

Johan Henriksson
johhe960@student.liu.se
Linköping University

Petra Öhlin
petoh078@student.liu.se
Linköping University

January 9, 2017

Abstract

Optical Music Recognition (OMR) is an application of optical character recognition. An OMR system is used for converting music sheets into a editable or playable digital representation. This report describes an implementation of an OMR system, the results from it provides and possible alternative methods. The results shows that the methods used gives a correct conversion in some conditions, but that alternative methods may provide better results in difficult conditions.

1 Introduction

The process of optical music recognition, converting a music sheet into notes and pitches in a digital representation, resembles the process of optical character recognition. It involves a lot of steps from the broad spectrum of image processing. Optical music recognition includes the steps from pre-processing the input image, to classifying extracted musical features into the output that is a string representation of the notes.

The quality of the software covered in this report is evaluated by the result of test runs, compared to manually classified sheets. The area of optical music classification is well documented and this specific software can be improved in many ways. Improvements and alternative methods are therefore also presented in the report.

2 Background

2.1 Project description

This aim of this project is to implement a optical music recognition software that is able to take images of music sheet and turn them into a text representation. This task can be divided into multiple subproblems, such as pre-processing, segmentation, matching, feature extraction and decision theory.

A set of 16 input images were given, both scanned and photographed, that covered multiple normal conditions such as different rotation, skewing, lighting settings, brightness, resolutions, etc. These images of a music sheet also ranged is what type of music tones and the number of staff line systems it had. The converted string representation of the music sheet contain the notes recorded from left to right following the systems, and a "line break" is simply marked at the end of each staff line system with the character 'n'.

The implementation software is Matlab and all functions described in the report is therefore Matlab functions. Matlab and it's Image Processing Toolbox provides a lot of built in functions for image processing and are therefore a good language to implement an OMR system in.

The main limitation of the program is that not all musical objects are classified, in fact only quarter notes and eighth notes are registered. Another musical limitation is that the pitch detection is assumes that there are only G-clef systems.

2.2 Image processing

Some image processing operations were a fundamental part of the solution as they are used several times in the project for different purposes.

2.2.1 Morphological operations

The basic idea in binary morphology is to probe an image with a simple, pre-defined shape, drawing conclusions on how this shape fits or misses the shapes in the image. This is done to increase or decrease the characteristic of the of the structural element that itself is a binary image. There are two fundamental morphological operations that can be used in different combinations to achieve different effects, these are erosion and dilation.

2.2.2 Horizontal and vertical projection

A projection is an decrease in dimensions, for a binary image this means going from two dimensions to one [3]. This is done by representing one of the dimensions as one value that is the sum. Summing up the pixels horizontally or vertically is resulting in a one dimensional vector with information about where objects are located in the image.

3 Method

3.1 Preprocessing

Some input images are digital photos of a music sheet on a table. They require more preprocessing than the scanned images, e.g. the table needs to be cropped off. The cropping function implemented in this software is taking a naive approach assuming that the table is of highest contrast to the music sheet in the blue color channel. After converting the input image into binary representation and removal of noise is done it simply checks the sum of each side. If a side contains mainly black pixels it is assumed to be a part of the table and is cropped off. The loop stops when no sides needs to be cropped. The digital photos are also of different sizes and since some thresholds later in the program are assuming a certain size of the note heads all images are re-sized to fit that assumption.

The Hough transform is a feature extraction technique and can be used to detect lines in the image [5]. From these lines, the angle which they vary from the main axis can be calculated and used to rotate the image so that all lines in the input image are uniform and parallel. In the case of music sheets, the staff lines are long straight lines and give a significant contribution to use for detection. The Image Processing Toolbox in Matlab provides the function `hough` to compute the Hough transform and houghpeaks to derive the coordinates of the peaks. When used together, the angle of rotation is derived.

When the image is rotated, the image contains edge pixels without information, that needs to be cropped. Since music sheets generally has some white space on the sides from the beginning, this can be performed without any information loss. To do this, the function `imrotate` used to rotate the image also has an optional parameter that crops the image based on the rotation angle.

3.2 Staff line identification

Staff lines consists of parallel straight lines, have consistent space between them and are always grouped in systems of five; Features that can be used when detecting them.

3.2.1 Horizontal projection

In the process of detecting the staff lines, horizontal projection was used to get the sum of white pixels in the image, in the horizontal direction. All data points that is larger than its two neighbouring points are classified as a local maxima, a local peak. All peaks that are larger than one third of the magnitude of the largest peak are identified as staff lines. This assumption was made to filter out all peaks that does not correspond to a staff line such as the song title. Figure 1 is an example of how the projection looks like, this example is a music sheet with four systems of staff lines. When projecting the magnitude of the peaks onto the original image, it is clear that the result is accurate, see figure 2.

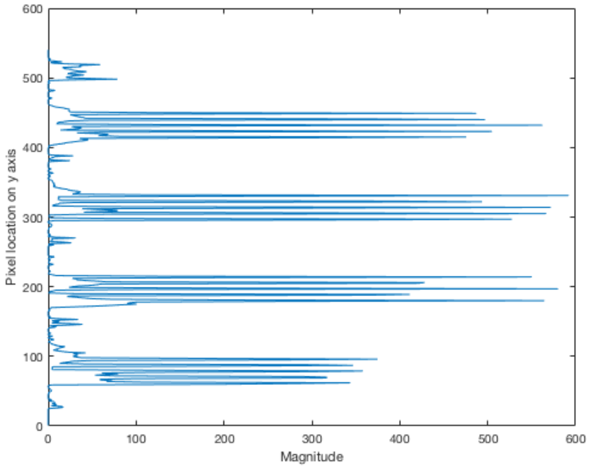

Figure 1: Horizontal projection of a music sheet.

Prior to calculating the projection, an erosion is performed on the image with a horizontal line as the structuring element. This is done to enhance the horizontal lines in the image. This results in a more stable calculation since the peaks in the projection are more characteristic.

Lastly, a semantic quality check is performed to ensure that it only systems with five identified staff lines are marked as correct. This is done because because an uneven result would mean that the rest of the calculations based on the staff lines are inaccurate.

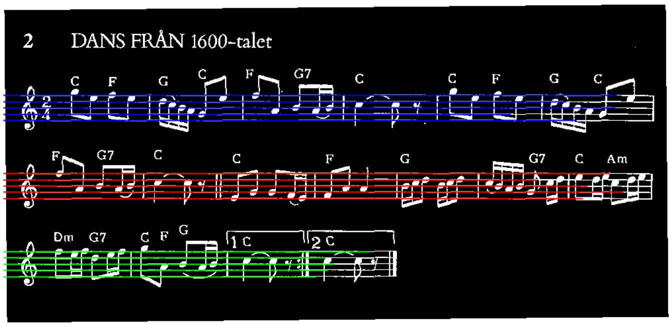

Figure 2: Detected staff line positions and the relative peak magnitude of the horizontal projection.

3.3 Identify notes

The goal of the identification phase is to find all locations for all interesting notes. The notes of interest in this task have one common feature that can be utilized in the identification phase, that all notes have filled note heads. Another fact that is used, is that the notes are stored in a specific order from left to right in staff line systems.

3.3.1 Identify notes

When the different systems of staff lines and their order on the music sheet are identified, the image is divided into sub images, containing only one staff line system each. The subimages are divided based on half of the distance from the last staff line on the upper group to the first staff line on the lower group. The top subimage is also cropped from the title on the music sheet, if there is one.

The located staff lines are removed to ease the notes classification later in the process [2]. This is done by setting the row of the staff line's location and its closest neighboring pixels to zero. This removal creates holes in notes that are located on a staff line, this distortion is fixed by a dilation with a vertical line as the structural element. This process is shown in figure 3 where the staff lines are removed, and in figure 4 where the dilation have been performed.

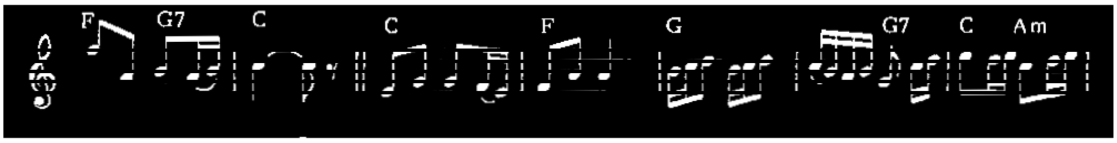

Figure 3: System with removed staff lines.

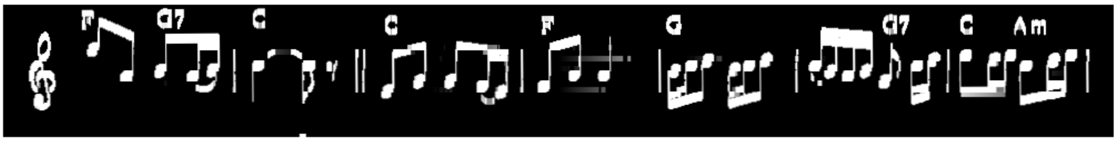

Figure 4: System with restored objects after removal of staff lines.

3.3.2 Fill gaps and filter out irrelevant objects

To remove all musical objects that are not of interest in the classification, such as pauses and chord specifiers, the image is eroded with a disk element with the size of a filled note head. The result of the erosion can be seen in figure 5 where five objects is left. One of the objects correspond to the G-clef and the rest is corresponding to notes that are interesting. The unwanted objects that are left after the erosion, such as the G-clef, are then filtered out. Since most of the remaining objects are notes, a good reference for the filtering is the to filter out the noise by removing all objects smaller then a percentage of the area of the biggest object.

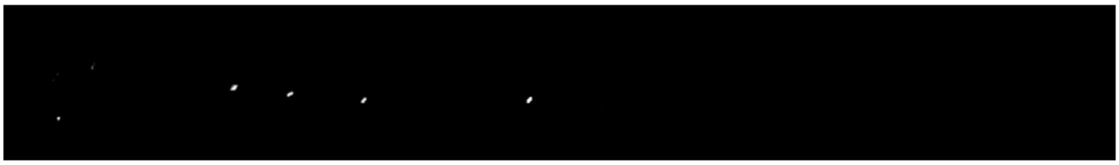

Figure 5: Filled note heads.

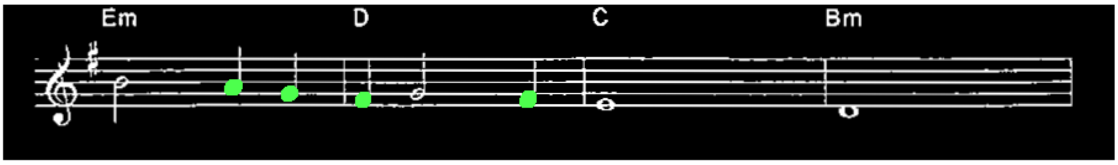

Figure 6: Filled note heads overlay on original image.

A dilation is performed on the image to merge objects that are close to each other, hence corresponding to the same note head. The dilated objects are plotted as an overlay on the image in figure 6. To get the exact location of the note heads the centroid, the geometric center of the shape, is used. The centroid is a property of Image Processing Toolbox function `regionprops`.

When the objects have been identified and located, the image needs to be cleaned up the image before the classification can be performed. This is because, even though the staff lines are removed, some objects might be located so close to each other that they would interfere in the classification and therefore aggravate the result of the classification.

From the image with removed staff lines, the bounding box of each object is obtained by `regionprops`, see figure 7. A bounding box is the smallest possible box surrounding a cohesive object. All bounding boxes are looped through and the only the interesting bounding boxes are left in figure 8. The content of all non relevant bounding boxes are set to zero, when all the irrelevant objects are filtered out, only the relevant objects are left, see figure 9.

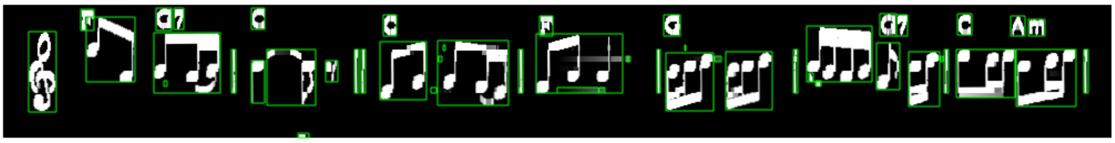

Figure 7: All bounding boxes.

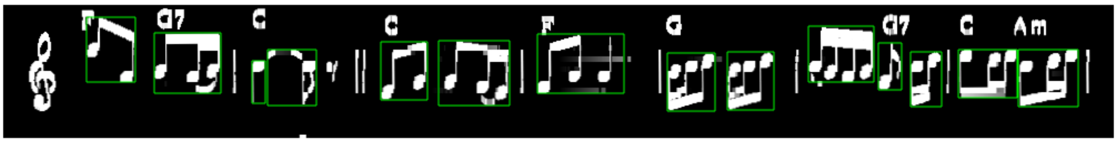

Figure 8: All interesting bounding boxes.

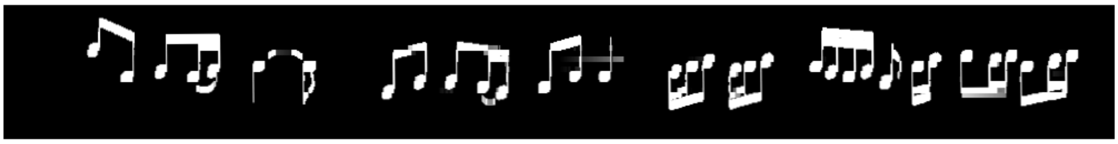

Figure 9: All interesting objects.

3.4 Classify musical features

When the objects have been identified and located, the musical feature classification can be determined. In this implementation, only quarter, eighth notes and their pitch are identified.

3.4.1 Classify note type

A decision tree like structure is used to both discard objects that does not need to be classified and to determine the notes that are going to be classified.

On the first level there are all notes with filled note heads, this is a result of the identification phase where all other musical objects have already been discarded. In the first decision the single notes are separated from the group notes, this is because the flag that corresponds to the single notes and the beam that corresponds to the group notes can be treated differently and used to do the final decision. In figure 10, only notes with stems above the note head is represented but notes with the stem under the note head are treated with the same decisions.

In both decisions the notion of bounding boxes are used. The bounding box is a property of all objects in a binary image that is wrapping the object in the smallest possible box, examples can be found in figure 7, marked with a green border. In matlab, this property is stored in `regionprops`.

In the first decision, it is simply checked if the note head is sharing the same bounding box with one or more other note heads, in that case it is part of a group note and otherwise it is a single note.

In the second decision, the bounding box is cropped so that each note head has each own bounding box just surrounding the stem, see the yellow border in figure 11 and 12. To know if the stem is above or under the note head, the vertical location of the note head is compared to the location of the bounding box. In the group case, the mean value of all note heads in the group's vertical location are used in the comparison to cover the case when a large difference in tone. The cropping of the single note takes advantage of the fact that the flag is always situated to the right of the stem. The decision is then based on the percentage of white pixels along the right vertical side of the cropped bounding box, marked with a red line in figure 11 and 12. As an example,

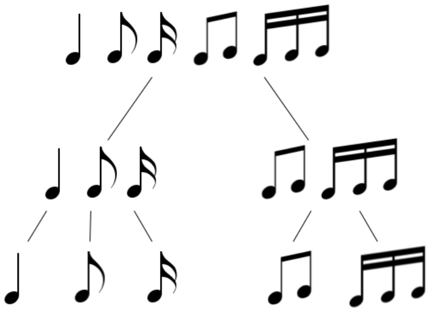

Figure 10: Decision tree like structure of the note type classification process.

the quarter note will not have any white pixels along the right side of the not, but the eight note will. For the group notes, the vertical side on the dropped bounding box that have the maximum amount of pixels is used in the comparison, this is used to be able to use the same method for all notes in the group. This algorithm was inspired by the technique used by Andy Zeng [1] described in the paragraph about alternative methods.

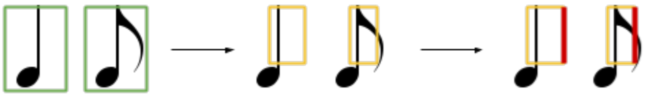

Figure 11: Bounding boxes and sub bounding boxes of single.

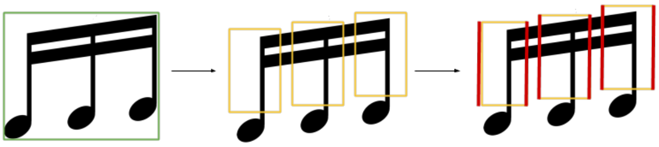

Figure 12: Bounding boxes and sub bounding boxes of group notes.

3.4.2 Classify pitch

The method used to classify the pitch of the notes, found in [1], is using the vertical difference from the note head to the lowest line in the system. The fact that the difference between the staff lines is fixed is used to know which pitch the note head corresponds to. The distance is than compared to a list of all possible pitches with respect to the reference pitch. A limitation in this approach is the list where the pitches are stored. If there is a note with a larger distance than the highest or lowest note in the list this method will not be able to classify it.

4 Alternative methods

4.1 Staff line identification

As described in the paragraph about staff line identification, the Hough transform is a method for locating lines in an image. In fact, this method can be used for both identifying the staff lines and the rotation in the image. The reason why this method is only used for finding the rotation angle in this project is simply because the horizontal projection method was implemented earlier. The horizontal projection is also more intuitive to understand and works well for the images in the test set.

4.2 Staff line removal

The current method for removing the staff lines sets the whole row containing the detected staff line and one pixel over and one under it, to zero, effectively making it black. This removes the staff lines, but leaves a gap in the notes. To fill them, a dilation is used with a line element, but this makes everything else thicker. A solution for this would be to not remove the staff line if two pixels over the staff line is white as well as two pixels under. That leaves the note more intact and the dilation unnecessary, which makes conditions better to do a good correct classification.

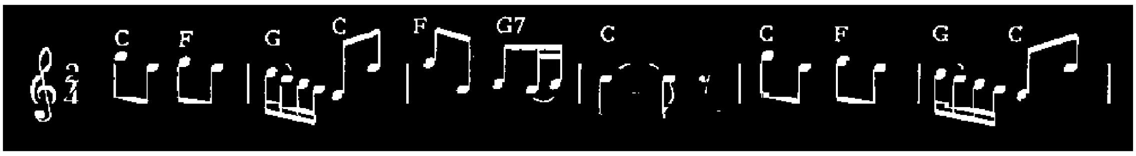

Figure 13: Alternative method for removing staff lines.

4.3 Musical type classification

There are many methods to use in the last decision of the classification process, that is distinguishing between quarter, eighth and sixteenth notes.

The first method used were horizontal projection of the bounding box. Different thresholds based on number of peaks in the projection and the strength of the peaks were evaluated before the new method were tested and concluded to result in a more stable classification.

An additional method presented by Andy Zeng in the paper [1] were discussed that could have been used if the reconstruction of the notes had been better after removal of the staff lines. This technique counts the number of holes along the vertical side of the bounding box. In this implementation, the case of two beams are sometimes merged as one in the dilation, and therefore wont work.

4.3.1 Template matching

Template matching is another technique that can be used for music classification. The main idea of the technique is to compare a clean bounding box surrounding the object that is going to be classified with a set of possible note types. The object in the bounding box is then classified as the image in the set that are the most similar. This method was rejected because the lack of reference set of note types and since the number of possible note types are relatively few in this specific program. This method would probably perform better than the chosen solution in a software that is classifying more types of musical types.

4.3.2 Neural network

One of the newest approaches that have been proven to be very powerful is the machine learning technique, neural network. Neural networks were originally inspired from the central nervous system and is now implemented in many areas far from the biological roots [2]. An implementation of neural networks generally results in superior classification capability [4]. This idea seems really exciting but it was discarded because it is not classic image processing that was the focus of the project.

4.4 Musical pitch classification

The first method used to classify the pitch used the position of all staff lines and gaps in the system. The vertical location of the note head were compared to this list and classified as the pitch which had the closest value in the list. This method was discarded because the notes above the staff line system where all classified as the pitch of the highest staff line. A combination of classifying the pitch to the closest line or gap and the distance to the lines could have been implemented but the accuracy of only calculating the distance to one of the staff lines where good enough for our solution.

5 Result

The result is test runs of the program compared to manually classified sheets. All files named with an s are scanned images, generally easier to classify, and all files named with an c are photographs, with different difficulties. The following table shows the result. The column named I/M specifies if the output correspond to the implementation (I) or the manually tagged (M) result.

Table 1: Result; Comparing output strings.

Image name	I / M	Output
im1s.jpg	I	c2g3e3f3e3g2e3f3a2b2C3C3g3e3f3e3d3g2e3n c2a2b2C3c3e2g2g2f2e2f2a2a2d3b2d3g2e3d3e3c3e3n c2f3d3g3a2b2C3c3C3c3
im1s.jpg	M	g3e3f3e3g2e3f3a2b2C3c3g3e3f3e3g2e3n f3a2b2C3c3e2g2g2f2a2A2d3d3g2e3c3n f3d3g3a2b2C3c3C3c3
im1c.jpg	I	a2f3d3e3d3d3b2a2g2e3f3a2b2C3c3g3e3f3e3g2n f3a2b2C3c3e2g2g2f2e2f2a2A2d3b2d3g2e3c3n d2e3a3b2a2b2c3c3c3c3
im1c.jpg	M	g3e3f3e3g2e3f3a2b2C3c3g3e3f3e3g2e3n f3a2b2C3c3e2g2g2f2a2A2d3d3g2e3c3n f3d3g3a2b2C3c3C3c3
im3s.jpg	I	G3g3a3G3E3d3f3E3D3d3e3d3b2c3d3e3f3G3n c2G3g3a3G3D3g3a3B3C4E3F3G3n c3c3c3d3e3d3c3d3E3C3d3d3d3e3f3e3d3e3F3D3n b1e3f3F3f3d3f3g3A3g3f3g3a3B3a3g3C4C4
im3s.jpg	M	G3g3a3G3E3e3f3E3D3d3e3d3b2c3d3e3f3G3n G3g3a3G3D3g3a3B3C4E3F3G3n c3c3c3d3e3d3c3d3E3C3d3d3d3e3f3e3d3e3F3D3n e3f3G3f3e3f3g3A3g3f3g3a3B3a3g3C4C4
im3c.jpg	I	b1f3a2a3G3E3e3f3E3D3d3e3d3b2c3d2c3f3n c2F3f3a2g3F3C3f3g3A3B3E3E3F3n c2c3b2b2c3d3c3b2c3D3B2c3c3c3d3e3d3c3d3E3C3n c2d3f2e3f3e3d3e3f3f3f3e3f3g3A3g3
im3c.jpg	M	G3g3a3G3E3e3f3E3D3d3e3d3b2c3d3e3f3G3n G3g3a3G3D3g3a3B3C4E3F3G3n c3c3c3d3e3d3c3d3E3C3d3d3d3e3f3e3d3e3F3D3n e3f3G3f3e3f3g3A3g3f3g3a3B3a3g3C4C4

5.1 Improvements and discussion

It can be concluded that the result of the implemented software are not fully accurate. Some notes are still classified wrong and sometimes more notes are found than there are on the sheet, especially in the test runs with photographed note sheets. A couple of flaws are identified in the implementation and these are presented in the following discussion.

In some test images, the G-clef is identified as a note head. This was not an issue until late in the development process when the filtering was changed to somewhat fit the more difficult test runs. To take care of this flaw the G-clef can be identified separately in the beginning and filtered out in the bounding box filtering step. One way to identify the G-clef could be to do a vertical projection of the image and search for the highest peak.

A skewed image or an image scan with a slight rotation would be a significant source of error when detecting the staff lines. In order for the horizontal projection to work properly, the image has to have been corrected of this in the preprocessing phase. Figure 14 is an example of when the detection is not working. When the detected lines are not dividable by five the shortest line is removed until the detected lines are dividable by five. This is in some cases this is not a correct solution to the problem and better preprocessing or more advanced technique depending on the distance between the lines is needed to achieve an accurate result.

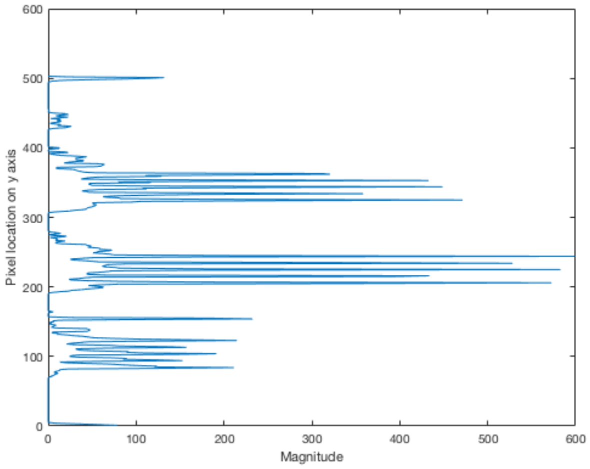

Figure 14: Horizontal projection when identifying staff lines is not working.

The decision technique have not been tested on all possible valid situations and in some cases, the distortion from the removal of the staff lines makes the decision theory classify notes incorrectly. Two examples of cases that have not been tested are chords and single sixteenth note (a note with two flags). These decisions therefore probably needs to be tuned more carefully to be classified right.

The current note detection algorithm is not perfect in the sense that noise can survive the operations that aim to filter them out, and be incorrectly identified as notes. The disk element that the filtering uses, resembles the note head shape and filters everything else out. Still, some objects in the music sheet that have a similar shape and size are incorrectly classified as notes. Additional checks or alternative methods like template that ensures that the detected notes actually are notes would improve the end result.

To improve the program so that it works for all types of input images, more pre-processing needs to be added to fix distortions and account for conditions that are common in digital images.

In general, more musical knowledge can be utilized to make more accurate decisions and simplify algorithms. An example is when identifying the staff lines, we are only checking if each staff line system is containing five lines, in addition to that it is known that the distance between each staff line is constant and that there is an relationship between that distance and the note head. This facts could be used to decide the size of the disk segment when identifying note heads, since the staff line identification is done first.

References

[1] A. Zeng, *Optical Music Recognition CS 194-26 Final Project Report*, 19 December 2014. Available: https://www.cs.princeton.edu/ andyz/omr.pdf

[2] A. Rebelo, G. Capela and J. S. Cardoso, *Optical recognition of music symbols - A comparative study*, 17 November 2009. Available: https://pdfs.semanticscholar.org/72d4/292ecb9cb3cc9158203d829898fe539c2187.pdf

[3] D. Bainbridge and T. Bell, *The Challenge of Optical Music Recognition*, 2001. Available: https://ai2-s2-pdfs.s3.amazonaws.com/8e7b/3859813a746f9b16a584ecb024103fce48cb.pdf

[4] C. Wen, A. Rebelo and J. S. Cardoso, *A new Optical Music Recognition system based on Combined Neural Network*, 2 February 2015. Available: https://www.researchgate.net/profile/Cuihong-Wen/publication/272753740-A-new-Optical-Music-Recognition-system-based-on-Combined-Neural-Network/links/55c497bb08aeca747d6135d7.pdf

[5] D. Ringwalt, R. B. Dannenberg and A. Russel, *Optical Music Recognition for Interactive Score Display*, 31 June, 2015. Available: http://www.nime.org/proceedings/2015/nime2015-198.pdf