Duplicate detection in facsimile scans of early printed music

Christophe Rhodes, Tim Crawford, and Mark d'Inverno

Goldsmiths, University of London, New Cross, SE14 6NW, United Kingdom
{c.rhodes,t.crawford,dinverno}@gold.ac.uk

Abstract. There is a growing number of collections of readily-available scanned musical documents, whether generated and managed by libraries, research projects or volunteer efforts. They are typically digital images; for computational musicology we also need the musical data in machine-readable form. Optical Music Recognition (OMR) can be used on printed music, but is prone to error, depending on document condition and the quality of intermediate stages in the digitization process such as archival photographs. This work addresses the detection of one such error – duplication of images – and the discovery of other relationships between images in the process.

1 Introduction

1.1 Digitization and Early Music Online

Librarians have kept irreplaceable artifacts in trust for centuries. Now, with modern digital storage and networking technology, the opportunity has arisen to greatly widen access to heritage, and libraries and archives are taking this opportunity as and when resources permit. Normal digitization efforts involve taking pictures of sources; this is adequate for the most part, although in some cases (*e.g.* Henry Billingsley's 1570 translation of Euclid's Geometry, the first geometrical "pop-up" book printed in sixteenth-century England; see Swetz and Katz, 2011) essential information is lost.

In *Early Music Online* (Rose, 2011), a "Rapid Digitization" project funded by the Joint Information Systems Committee (JISC), over 320 printed volumes (35,000 pages) of music from 16th-century sources held in the British Library were digitised from microfilm, and made available to the community at large in the form of images, licensed for non-commercial use.

A photographic digitization process, as was carried out for *Early Music Online*, does not cause an immediate loss of information. The fact that digitization of the sources in *Early Music Online* was not from the originals but from microfilm has consequences for the published set of images – but the

digitization also offers an extra opportunity: just as images of text could be further processed to make the text on those pages available, so we might want to make available not just the images of the musical source but also a representation of the musical content contained within it, in order to facilitate further analysis (by the human scholar, by automated processes, or most likely by a hybrid of the two).

However, we need to deal with the problem of images which, for one reason or another, are rescans of the same pages, as they must not be treated as distinct entities. These images are not precise digital duplicates of each other, and so must be detected through some approximate means. As well as duplicate scans, there are other forms of similarity present in the collection, such as musical relatedness and movable type reuse.

We present our work on developing and combining image-based near-duplicate detection, based on Scale-Invariant Feature Transform (SIFT) descriptors (Lowe, 1999), with OMR-based musical content near-duplicate detection. We evaluate an order-statistic based method for finding duplicate scans of pages, and additionally identify a number of distinct kinds of approximate similarity emergent from our distance measures: substantial reuse of graphical material; musical quotation; and title page detection.

1.2 Optical Music Recognition

Although Optical Music Recognition (OMR, by analogy with Optical Character Recognition for text) has been a subject of research since the 1960s (Pruslin, 1966 and Prerau, 1970; see Kassler, 1972), it remains in general a difficult, unsolved problem (Rebelo *et al.*, 2012). Partly this is because, unlike text, common musical notation is made up of a number of intersecting graphical elements; partly because, again unlike most text, the two-dimensional layout of the page is highly significant to the interpretation of the glyphs.

In our particular context, there is the additional difficulty that we are dealing with historical artifacts, from before the standardization of musical layouts – indeed, the *Early Music Online* collection is at the very start of printed music, when each printer would have had their own collection of movable type. Nevertheless, accuracy rates of around 90% are achievable (Pugin and Crawford, 2013) on the majority of the collection, with some sources allowing OMR to be performed with far greater precision and recall than others.

In the long term, we aim to overcome these difficulties, to allow full-music search and other algorithmic processing, just as OCR has allowed scholars to perform full-text search over the contents of documents, not just their metadata. This paper deals with one piece of the puzzle: namely, identifying portions of the source on which the results of OMR should *not* be included in any such automatic transcription, but rather flagged for a human expert to investigate. In the next section, we describe measures of similarity between images of musical notation; we then use these measures to characterise particular relationships between pages from three of the sources (475 pages) from

Early Music Online.

2 Similarity measurements

2.1 Image similarity

As a basic measure of image similarity, we follow Lowe (1999) in computing SIFT descriptors for each image, which are invariant to (uniform) scaling and rotation, and robust against affine distortion and lighting changes. In order to compare the image similarity between a source image and a target, we compute for each descriptor in the source the *two* nearest (as measured by the Euclidean distance) descriptors in the target, and count a 'hit' if the distance to the nearest is less than two-thirds of the distance to the second nearest. The overall similarity score for the pair of images is the sum of the 'hits' from image to source, without reference to relative position or orientation. Note that this similarity is not necessarily symmetric, as the source and target images are treated differently.

2.2 Musical similarity

We use the Aruspix software (Pugin, 2004; Pugin and Crawford, 2013) in untrained mode to extract musical data from images. Note that Aruspix will attempt to extract musical information no matter what the source image: for images containing no musical notation at all, this of course means that the output will be musically nonsensical, resulting from chance agglomerations of glyphs and graphical material which look 'enough' like music to Aruspix's recognizer. We convert the output of Aruspix, a representation of the musical data identified to strings representing either the diatonic melody or the diatonic intervals present on each staff, for example:

kind	string
melodic	`SSQRSRPRQPNPONOPQRSTSSRTSRP`
interval	`-bAAabBaabBaaAAAAAAa-aBaab`

The melodic string encodes the diatonic pitch (similar to chromatic pitch, but with 7 notes per octave rather than 12, thus disregarding accidentals) as the ASCII character with code point 48 + the diatonic pitch. The interval string encodes the diatonic interval between successive diatonic pitches, with `-` indicating no change, capital letters representing movement upwards (`A` representing up one step, `B` up two) and lower-case letters movements downwards (`a` representing down one step, `b` down two, and so on)

We thus obtain one of these strings for each of the cases (melodic and interval) per line of music. We compute the similarity score of a source image to a target image by: first, taking the string for each line in the source image; second, finding and scoring the closest match in the target image using the

Wu-Manber algorithm (Wu and Manber, 1994, as implemented in `agrep`); and finally, summing those scores over all lines in the source page.

2.3 Outlier analysis

We identify various possible scenarios for a scan X or a pair of scans (X,Y), which we encode as predicates:

music(X)
: the scan X is primarily of musical notation

duplicate(X,Y)
: the scans X and Y are near-duplicates of each other

musicsim(X,Y)
: the scans X and Y contain substantially similar musical material

graphicsim(X,Y)
: the scans X and Y contain substantially similar graphical material

Some of these predicates imply other relations;

- duplicate(X,Y) → graphicsim(X,Y)
- duplicate(X,Y) → (music(X) → musicsim(X,Y))
- duplicate(X,Y) → duplicate(Y,X)

the asymmetry arising from the fact that all scans contain graphical material, but not all scans contain musical material.

An ordered pair of scans (X,Y) will have two similarity scores associated with it: a similarity score based on image similarity, and a second score based on the musical similarity imputed from comparing the output of the OMR process. These similarity scores tell us nothing *a priori*; in order to extract meaning from them, we must compare them against thresholds. However, there is also no way of *a priori* deriving thresholds of similarity for "interestingness", so we use the distribution of similarity scores between X and all other scans as a way of establishing a threshold.

Specifically, we fit a lognormal distribution to the central 80% of similarity scores, for each of the measures (image and music) separately; we then treat as a threshold the 0.5% level of improbability, accepting the default thresholds from the implementation in the extremevalues R package (van der Loo, 2010). This then gives us three possible diagnostics for each similarity measure:

- (X,Y) are unusually *similar* to each other;
- (X,Y) are unusually *dissimilar* to each other;
- (X,Y) have a similarity score which is not particularly distinctive.

These diagnostics, when the two similarity scores are combined, give a total of nine possible outcomes for each pair of scans.

2.4 Hypothesis

Our hypothesis is that we can use the combination of our music and image similarity measures to identify near-exact duplicates resulting from multiple images of the same pages on the microfilm.

Specifically, we invert the relationships in Section 2.3, and attempt to infer higher-level information from the low-level outlier information. If (X,Y) are unusually similar according to the music similarity measurement, we assert the musicsim(X,Y) relation; similarly with image similarity and graphicsim(X,Y); and we further infer duplicate(X,Y) from graphicsim(X,Y) $\wedge$ musicsim(X,Y).

Other outlier cases (pairs where one similarity score is high but not the other, and pairs where at least one similarity score is low) are also potentially of interest, and we can attempt to characterize the relationships between pages that give rise to those scores more qualitatively in the results below.

3 Results

Our test collection is 475 images resulting from scans of three sets of part-books of parody masses (mass settings based on a pre-existing piece of music) published in 1545-6 by Tielman Susato in Antwerp. This is an interesting test set from the point of view of our similarity measures. Firstly, the nature of parody masses is that there will be significant reuse of musical content, within a single work (in the same voice and different mass section, and in the multiple voices) and between distinct works (for example, if there are multiple masses on the same original material, though this does not in fact occur in this set of images). Secondly, since the books were printed by the same printer there is the likelihood that graphical material might be reused without any musical similarity between the material on the pages.

Given this test collection, there are 225,625 pairwise comparisons between images, given that our definition of these comparisons is not symmetric, and including the comparison of a scan with itself. We would expect the identity comparison to show up as an outlier in both measures – indeed, this is useful as a consistency check – and at least 180,160 (80% of the rest) to be considered as having uninteresting distances (since we are fitting the distribution to the central 80%).

similarity	low (graphic)	medium (graphic)	high (graphic)
low (music)	1083	3215	7
medium (music)	6091	213,122	1592
high (music)	0	18	497

Table 1. Counts of similarity judgments between all pairs of pages in our dataset, for both similarity measures. Outliers according to the lognormal fit are labelled "low" and "high", while "medium" indicates a non-outlier.

From Table 1, we can observe firstly that the lognormal fit is presumably working reasonably well: the number of non-outlier pairs is comfortably above the 180,160 which would be the minimum. This view of the aggregate data does not of course preclude there being individual cases for which the lognormal fit was inappropriate; however, on the dataset as a whole it appears to be justifiable.

Secondly, the number of high-melodic/high-image similarity pairs is 497, 22 above the 475 identity matches. From just this table it is not possible to say, but one way that this can arise is if there are 11 duplicate image pairs, all of which are detected in both directions. In fact, because of artifacts arising from the musical similarity measure applied to pages with no musical content, it turns out that two of the identity matches are misclassified, and there are in fact 12 duplicate image pairs detected by this measure, which we publish on the semantic web (retrievable using `curl -H 'Accept: text/n3' http://duplicate-pages.emo.data.t-mus.org/`). Figures 1 and 2 illustrate some of the duplicate image pairs found using this method.

103a 104a

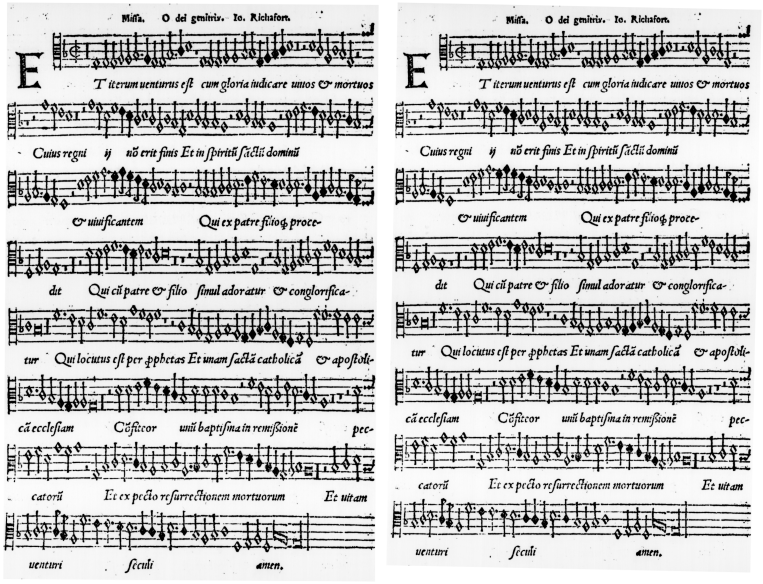

Fig. 1. Two pages with high image and musical similarity, from Susato (1545): these are most likely successive photographic shots of the same physical page.

Thirdly, there are some interesting cases to investigate: in particular, the large number of high-image/medium-melodic cases; the seven cases of high-image similarity and low-melody similarity; and the low-image/low-melodic similarity cases. Since these are not in fact exact duplicates, it is apparent

142b 199b

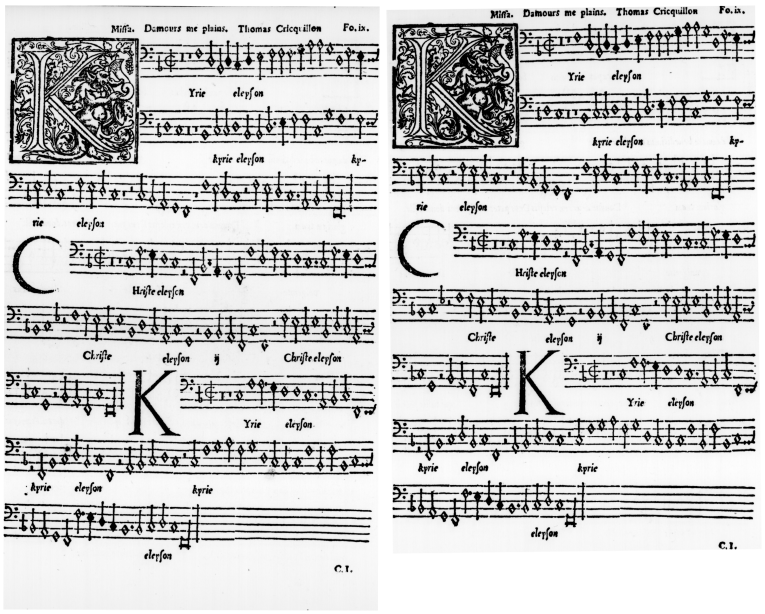

Fig. 2. Two pages with high image and musical similarity, from Susato (1546b): not shots of the same physical page, but most likely a misbound gathering.

that combining the outlier judgments of the two similarity measurements was necessary for the basic task; the cases with one or other measure (but not both) showing high similarity reveal other relationships between the material on each page.

Figure 3 shows a pair of pages with high image similarity, but a melodic similarity between the pages that is no higher than expected according to the fit. Note the reuse of decorated initial capitals, a feature of the printing technology and resources of an individual printer in the sixteenth century – an individual printer (Tielman Susato, in this case) would not have a wide repertoire of type for decorated capitals, and so would reuse one of the appropriate size each time there was a call for one. Since we are here dealing with mass settings, there will be many examples of initial 'K's and 'C's for *Kyrie* and *Christe* movements.

Figure 4 highlights another feature of this set of works: many of the mass settings are 'parody masses': settings based on musical material of another work, which gets reused throughout the mass setting. In this case, we have the ending of the *Gloria* and the start of the *Credo* from Thomas Crequillon's *Missa Kein in der Welt so schön*, both using the material from the song for a substantial fraction of the page.

072a 180b

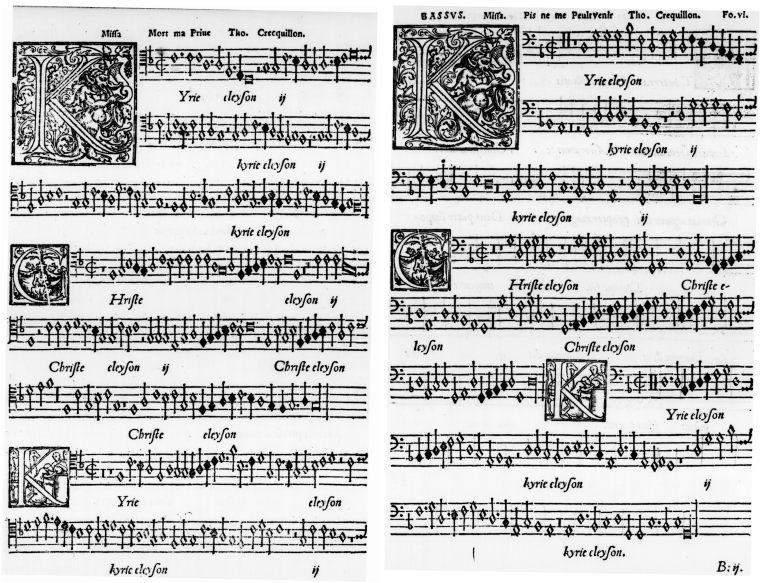

Fig. 3. A pair of pages with high image and medium melodic similarity, from Susato (1546a).

088a 088b

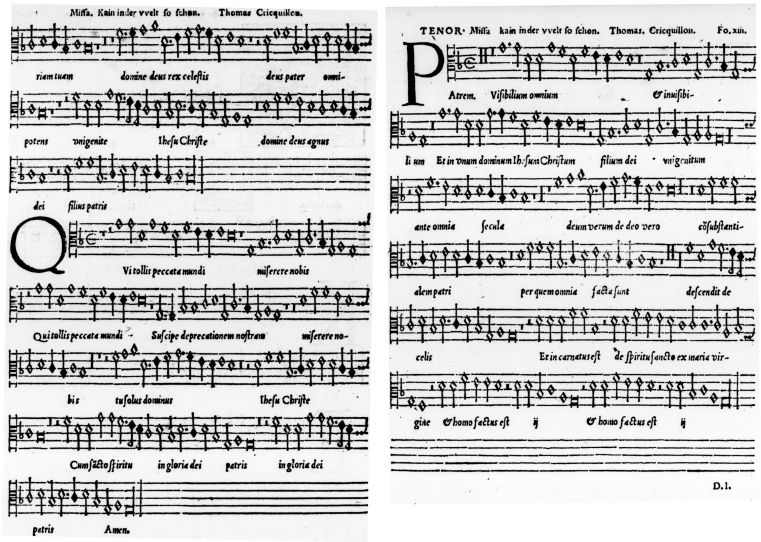

Fig. 4. A pair of pages with high melodic and medium image similarity, from Susato (1545).

076a 017b

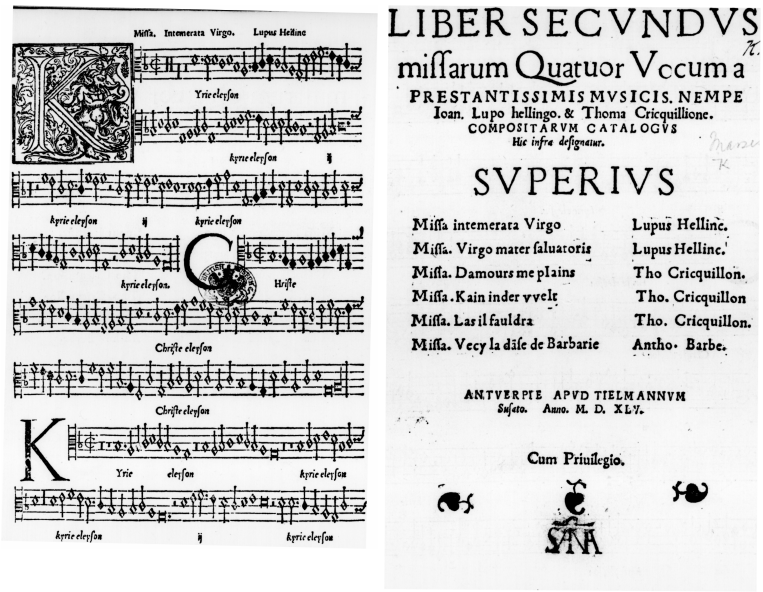

Fig. 5. A pair of pages with low melodic and low image similarity, from Susato (1545)

Finally, Figure 5 illustrates that this consideration of outliers also catches non-musical material: Aruspix will attempt to perform OMR on images that it is given, and there is no metadata accompanying the set of images to indicate which contain musical material and which do not. However, the essentially random output from OMR on title pages will be dissimilar to most of the detected content, as will the image features compared with image features from pages which do contain musical material; this also explains the seven cases with high image similarity and low melodic similarity, in which one or both of the pages contain substantial amounts of text.

4 Conclusions

We have shown that a combination of image and music similarity measures can be used to identify duplicates and near-duplicate photos in digitized archives, and also to identify pairs of pages of possible interest falling short of being considered duplicates. Even though the similarity measures themselves are simple, their combination is sufficient to identify all the duplicates with no false positives, in this particular dataset. Analysis of other outlier cases shows potential to identify reuse of musical material, reuse of type, and classification into music-containing and non-musical pages.

4.1 Further work

The distance measures between items we have used in this investigation are very simple; we have used SIFT image features without attempting to detect higher-level objects, and musical features with no attempt to consider perceptual similarity or even duration of individual notes. We could improve the image distance measure to take into account the coherence of groups of matches, as is done in pose estimation, though this would not address the most obvious false-positive of reuse of type for decorated initials. We could also attempt to deal with this by considering image features only on those regions which are detected as music by the Optical Music Recognition program.

We would also like to make our approach scale. At present, the method is workable on datasets of this size, 475 pages, corresponding to individual books or restricted sets of books, and in practice there are already interesting duplicates present in sets of that size. In principle, we would like to run our method on larger datasets as a whole to investigate whether there is any contamination or other connections between sources; however, the pairwise comparison leads to $O(N^2)$ time complexity, and so building a feature index is a necessary step to apply this to larger collections.

We have published our similarity judgments from this investigation as Linked Data at `http://duplicate-pages.emo.data.t-mus.org/`, and we will expand this resource as we generate more data. As well as publishing individual duplicate pairs, we aim to publish higher-level judgments, such as the presumed cause of the duplication – from the photographic process as in Susato (1545) or the binding in Susato (1546b). Finally, in the *Transforming Musicology* project, we aim to apply a similar method to similarity judgments of more general musical artifacts, such as musical recordings and editions of musical works.

4.2 Acknowledgments

This work was supported by the *Transforming Musicology* project, AHRC AH/L006820/1.

5 References

- KASSLER, M. (1972): Optical Character-Recognition of Printed Music: A Review of Two Dissertations. *Perspectives of New Music 11(1) 250–254*
- VAN DER LOO, M. P. J. (2010): extremevalues, an R package for outlier detection in univariate data, R package version 2.1
- LOWE, D. G. (1999): Object recognition from local scale-invariant features. *International Conference on Computer Vision, 1999, 1150–1157*
- PRERAU, D. S. (1970): Computer Pattern Recognition of Standard Engraved Music Notation. MIT Libraries
- PRUSLIN, D. H. (1966): Automatic recognition of sheet music. MIT Libraries

- PUGIN, L. (2006): Aruspix: an Automatic Source-Comparison System. In: W. B. Hewlett and E. Selfridge-Field, E. (Eds.), *Music Analysis East and West*, MIT Press, Cambridge, MA, 49–60
- PUGIN, L. and CRAWFORD, T. (2013): Evaluating OMR on the Early Music Online Collection. *Proceedings of ISMIR 2013, 439–444*
- REBELO, A., FUJINAGA, I., PASZKIEWICZ, F., MARCAL, A. R., GUEDES, C. and CARDOSO, J. S. (2012): Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval, 1(3), 173-–190*
- ROSE, S. (2011): Introducing Early Music Online. *Early Music Review, (143), 14–16*
- SUSATO, T., ed. (1545): Missarum quatuor vocum : Liber secundus / a prestantissimis musicis Nempe Ioan. Lupo hellingo. & Thomas Cricquillione. Compositarum catalogus hic infra designatur. Antwerp
- SUSATO, T., ed. (1546a): Missarum quinque vocum : Liber primus / a diversis musicis compositarum, quarum nomina catalogus indicabit. Antwerp
- SUSATO, T., ed. (1546b): Missarum quatuor vocum : Liber tertius / a diversis musicis compositarum. Antwerp
- SWETZ, F. J. and KATZ, V. J. (2011): Mathematical Treasures - Billingsley Euclid. *Loci, January 2011*
- WU, S. and MANBER, U. (1994): A Fast Algorithm For Multi-Pattern Searching. TR-94-17, Department of Computer Science, University of Arizona