Duplicate detection in facsimile scans of early printed music

Christophe Rhodes, Tim Crawford, and Mark d'Inverno

Department of Computing, Goldsmiths, University of London
{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.

In performing OMR on the British Library's *Early Music Online* collection (Pugin and Crawford, 2013) of 16th century volumes we must deal with the problem of images which are rescans of the same pages. 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 SIFT features (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 from our distance measures: substantial reuse of graphical material; musical quotation; and title page detection.

References

PUGIN, L. and CRAWFORD, T. (2013): Evaluating OMR on the Early Music Online Collection. In: *Proc. International Society for Music Information Retrieval Conference*. Curitiba, Brazil, 439–444

LOWE, D.G. (1999): Object recognition from local scale-invariant features. In: *Proc. Int. Conf. on Computer Vision*. Corfu, Greece, 1150–1157

Keywords

MUSIC, OPTICAL MUSIC RECOGNITION, CLUSTERING, SIMILARITY MEASURES