Towards musicologist-driven mining of handwritten scores

Masahiro Niitsuma, Yo Tomita, Weiqi Yan, *Senior Member* and David Bell

Abstract

Historical musicologists have been seeking for objective and powerful techniques to collect, analyse and verify their findings for many decades. The aim of this study was to show the importance of such domain-specific problems to achieve actionable knowledge discovery in the real the world. Our focus is on finding evidence for the chronological ordering of J.S. Bach's manuscripts, by proposing a musicologist-driven mining method for extracting quantitative information from early music manuscripts. Bach's C-clefs were extracted from a wide range of manuscripts under the direction of domain experts, and with these the classification of C-clefs was conducted. The proposed methods were evaluated on a dataset containing over 1000 clefs extracted from J.S. Bach's manuscripts. The results show more than 70% accuracy for dating J.S. Bach's manuscripts. Dating of Bach's lost manuscripts was quantitatively hypothesized, providing a rough barometer to be combined with other evidence to evaluate musicologists' hypotheses, and the practicability of this domain-driven approach is demonstrated.

Index Terms

domain-driven data mining, optical music recognition, historical musicology, music informatics, music information retrieval

M. Niitusma is with the Dept. of Media Technology,
College of Infotmation Science and Engineering,
Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu Shiga 525-8577 JAPAN
E-mail: mniitsuma@media.ritsumei.ac.jp
Tel: +81-77-561-3946

Towards musicologist-driven mining of handwritten scores

I. Introduction

In the development of musical scholarship, handwritten scores played a significant role—even after the invention of printing—as it is in this form that composers conceived their works and left them to posterity. They also captured otherwise hidden aspects of the composers' ideas and habits with all their subtleties. As they are often the only surviving evidence about writers and their work, they should be analysed with the utmost care and attention. In these extremely information-rich documents, however, such knowledge and patterns are not easily visible to non-experts.

Musicologists working in the area of manuscript studies have developed various methodologies to discover specific knowledge such as the date of origin, the identity of writers and how their handwriting changed over the years, from relatively small samples. They have managed to extract this knowledge from the available music manuscripts. But due to the lack of explanation as to how the samples are chosen, let alone musicologists standard practice of never disclosing all the factual data they have used to arrive at their conclusions in their publications, it has always been problematic when trying to verify their claims. One of the broader aims of this paper is to address such shortcomings in their traditional methodologies by making the analytical process more objective and transparent while at the same time allowing the expansion of data sets when newly discovered manuscripts become available.

II. Related work

While most of the work has been performed on modern handwriting recognition [1], far too little attention has been paid to historical documents, let alone music manuscripts. Although optical music recognition (OMR) has been investigated actively [2], [3], there has been little research investigating "deeper" aspects of music manuscripts beyond OMR, such as writer and chronology identification. A few publications have been made regarding manuscript dating beyond OCR (optical character recognition) [4]; however there have been no publications on chronology identification in music manuscripts using computer science.

The analysis of early music manuscripts still relies heavily on the work of musicologists, who can sometimes detect even a subtle change of handwriting, revealing some aspects of the situation under which the writer was working. This suggests that the current data-centered mining approach may have to be reconstructed with the help of domain experts.

In recent years, the strong connection between data mining and actual application domains has led to the third generation of data mining called *domain-driven data mining* [5], [6]. The principle aim of associated studies is to achieve actionable knowledge discovery by filling the gap between a number of algorithms found in academic publications and those really workable and meaningful in the application environment. Most recently this trend has reached domains in the humanities, such as the arts, design, and culture in contemporary society—leading to a number of projects addressing more expert-oriented analyses that are oriented to the domain experts' needs. In our application domain, the gap between existent research concerning music scores and musicologists' domain knowledge is still huge. To fill the gap, the actual problems in the domain have to be identified and considered.

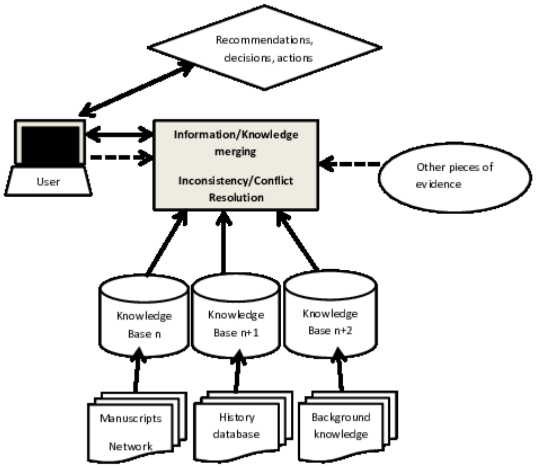


Fig. 1. An overview of the proposed database architecture

In our previous paper [7], musicologist-driven writer identification has been investigated, and the proposed algorithm is designed to be stand-alone. This paper concerns a method of extracting numerical evidence with uncertainty to be fused

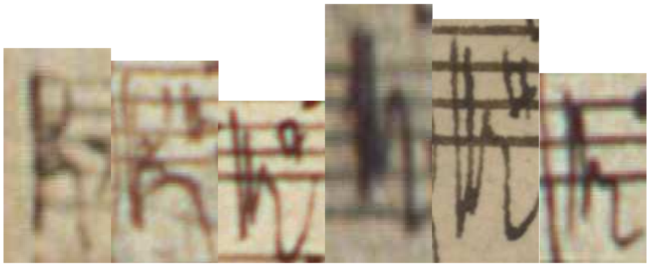


Fig. 2. The C-clefs of Bach's handwriting in the order of chronology suggested by musicologists.

with other evidence. It's use is demonstrated in chronology identification, which requires subtle analysis of handwriting.

Therefore, the proposed algorithm is designed to be incorporated into *musicology manuscript mining (MMM) work bench* (Fig. 1), to cope with uncertainty resulting from three problems in particular: 1) the lack of quantitative evidence; 2) the breadth of domain knowledge required to interpret sources; 3) the vast amount of evidence to be combined to reach final conclusions. The struggle to integrate the needs of historical musicologists with computer-based research is expected to reveal opportunities for developing data mining, shifting data-driven hidden pattern mining to domain-driven actionable knowledge discovery.

III. Chronology identification and C-clefs

Even if all the (significant) image processing problems could be resolved in this domain, the question of how to extract meaningful information from the data remains. Potentially valuable analyses suggested by Kobayashi, who is one of the most authoritative musicologists working on Bach's source studies, are facilitated to some extent by the MMM work bench. This paper is concerned primary with the paleographical value of music manuscripts focusing on C-clefs.

C-clef has been identified by Bach scholars as one of the most crucial symbols for dating manuscripts. Von Dadelsen [8] claims, for example, that Bach's C-clef can be categorized into three or four groups, each coinciding with a specific period, and von Dadelsen used this information to establish the chronology of Bach's manuscripts. Fig. 2 shows the C-clefs found in Bach's autograph manuscripts which are arranged in one chronological order suggested by musicologists. They demonstrate how the shape of Bach's handwriting changed over time.

IV. Selecting THE DATASET

As there is a controversy among Bach scholars regarding both the authorship and chronology of C-clef forms, our sample dataset has been created by selecting the most reliable manuscripts from an undisputed portion of Bach's fair copies that date between 1708 and 1748 under the limitation of their availability. This has been carefully done after discussion with historical musicologists. The details of this dataset are shown in Table 1. Two classification tasks were addressed using this dataset: one is eight-class classification using the dates proposed by Kobayashi as labels; the other is two-class classification which only distinguishes between the sets of pieces {A B C} and {D E F G H}. This corresponds to determining if a certain clef was written before he arrived at Leipzig (i.e. May 1723) to assume his role as Thomas cantor and director of music for the town, or after that date.

V. Extraction of C-clefs

Although texture-based features can be extracted from music manuscripts without segmentation [7], subtler analysis such as the specification of chronological order needs feature analysis from smaller symbols such as clefs as explained in the previous section. The extraction of C-clefs from manuscripts requires accurate segmentation. However, the segmentation of old music manuscripts has proved to be a difficult task, even in printed music. The difficulty is intensified in the case of handwritten scores. These difficulties seem to be caused by degradation such as show-through and bleed-through effects. In addition, microfiche, the primary medium for Bach's manuscripts in the study, gives the images in low-resolution, which creates further problems for image processing. For this reason, directions by domain experts play an important role in almost all the tasks involved. While we focus on the automated parts of the processing of manuscripts in the next section, we believe that it is clear that domain experts' input is required to initially drive the mining. Input is done through a graphical interface allowing the users to remove irrelevant pixels, to joint fragmented objects, to disconnect touching symbols, and to discard unusable objects.

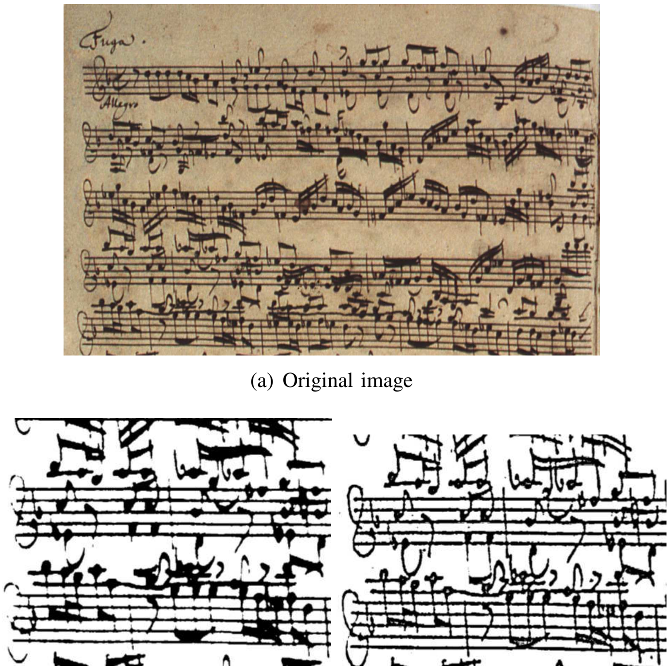
The input image is first pre-processed; this includes binarization and noise removal to prepare for feature extraction. Conventional global thresholding such as Otsu's binarization often breaks the shape of musical symbols as shown in Fig. 3(b). This problem can be resolved using adaptive binarization techniques such as Niblack's method. As shown in Fig. 3(c), the shape of symbols (clefs for example) is then sufficiently retained for further analysis.

In the present study, extraction of C-clefs from these manuscripts was mainly conducted with manually specified bounding boxes due to the problems caused by poor quality of the input images, as shown in Fig. 4. This extraction is followed by both morphological operations and staff-line removal to procure a clear image, in order to prepare for feature extraction. This is shown in Fig. 5(b) and (c). In feature selection, 15 features implemented in Gamera were used, and the effect of each feature is shown below.

- *area*
  The area of the bounding box.
- *aspect ratio*
  The aspect ratio of the bounding box.
- *black area*
  The number of black pixels.
- *compactness*
  The volume to surface ratio.
- *moments*
  The centre of gravity on x and y axis normalized by width and height.

TABLE I
DATA SET USED FOR THE EXPERIMENT.

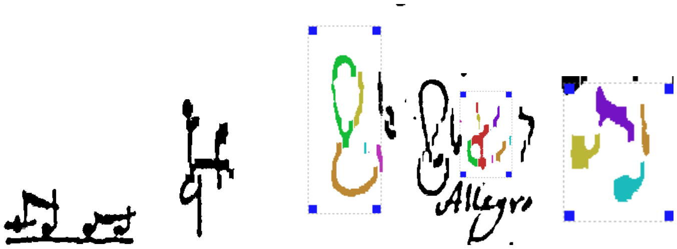
ID	Title	BWV	Source	Date	Sample size
A	Cantata “Gott ist mein König”	BWV71	D-B, Mus. Ms. Bach P 45	1708	89
B	Alles mit Gott und nichts ohn’ Ihn	BWV1127	D-W, Ra B 24	1713	11
C	Inventions and Sinfonias	BWV772–801	D-B, Mus.ms. Bach P 610	1723	188
D	Magnificat	BWV243	D-B, Mus.ms. Bach P 39	1733	221
E	Mass in B-minor, Kyrie-Gloria	BWV232	D-B Mus.ms. Bach P 180	1733	248
F	St Matthew Passion	BWV244	D-B, Mus.ms. Bach P 25	1736	633
G	Well-Tempered Clavier II, No. 10, 19, and 24	BWV879, 888, and 893	GB-Lbl, Add.MS. 35021	1739	69
H	Canonic Variations on Vom Himmel hoch	BWV769	D-B, Mus.ms. Bach P 271	1748	22


(a) Original image

(b) Otsu’s binarization

(c) Niblack’s adaptive binarization

Fig. 3. Two different thresholds.


(a)

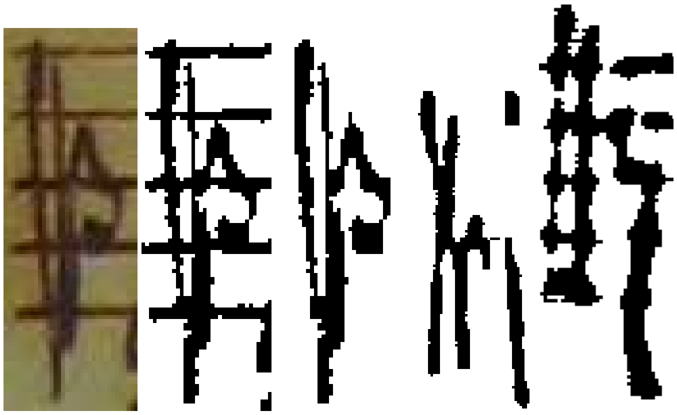
(b)

(c)

(d)

(e)

Fig. 4. Two crucial problems of OMR: (a)(b) collided symbols; and (c)(d)(e) fragmentation. In (c)(d)(e) the correct bounding box is specified by user, thus connecting fragmented objects into one music symbol.


(a)

(b)

(c)

(d)

(e)

Fig. 5. C-clefs cropped with the manually specified bounding box and prepared for feature extraction: (a) original clef; (b) binarization using Niblack’s method; (c) staff-line removal using Dalitz’s method and noise reduction with morphological operation; (d)(e) other examples including irrelevant pixels.

- *ncols feature*
  The number of columns.
- *nholes*
  The averaged number of white runs not touching the border. This is computed both for each row and each column.
- *nholes extended*
  Divides the image into four strips and then does a nholes analysis on each of those strips. This is first done vertically and then horizontally, resulting in a total of eight feature values.
- *rows feature*
  The number of rows.
- *skelton features*
  Generates a number of features based on the skeleton of an image.
- *top bottom*
  The first feature is the first row containing a black pixel, and the second feature is the last row containing a black pixel.
- *volume*
  The percentage of black pixels within the rectangular bounding box of the image.
- *volume16regions*
  Divides the image into a 4 x 4 grid of 16 regions and calculates the volume within each.
- *volume64regions*
  Divides the image into a 8 x 8 grid of 64 regions and calculates the volume within each.
- *zenrike moments*
  Computes the absolute values of the normalized zernike moments up to order six.

VI. TRAINING CLASSIFIERS

The performance of random forest (RF), which worked the best in the preliminary experiment, was investigated using 10-fold cross-validation and compared with other methods: support vector machine (SVM), bagging, and boosting. RBF (radial basis function) kernel was used as the kernel function of SVM, and this was automatically estimated from the result of a preliminary experiment. CART (classification and regression tree) algorithm was used for weak learning in all the ensemble classifiers and the other parameters were set as default.

Table II shows the result of 10-fold cross-validation (averaged over 100 repetitions). The best accuracy for two-class classification was 100% obtained by RF. This accuracy seems significant considering that the classification of Bach's handwriting has been attempted by only a few experts. The eight-class classification is far more complicated and thus extremely difficult even for human experts. The best accuracy of 72.53% was achieved by RF for eight-class classification. The value of the area under the ROC (receiver operating characteristic) curve (AUC) shows no significant differences, and RF is the best (AUC can be more accurate in the case of unbalanced data where one of the class is rare). It is interesting that Bayes outperformed SVM despite its simplicity. While there is much room for improvement, these classifications may serve as a rough barometer for musicologists.

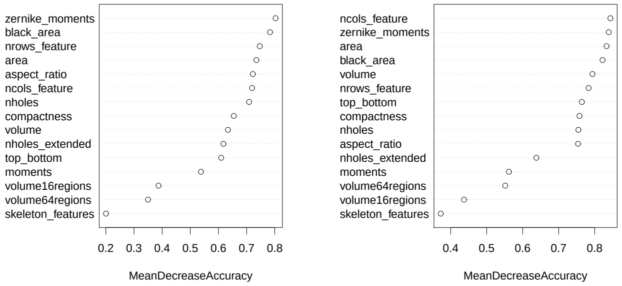
VII. Feature analysis

In addition to simple classification, RF is often used as permutation based variable importance measures under predictor correlation. This not only improves the result of each classification by proper feature selection, but also enables deep analysis of the changes in handwriting.

Kobayashi asserted that Bach's handwriting changes over time. On the other hand, scholars of forensic handwriting such as Headric [9] insist that such changes in handwriting are normally slight and gradual compared to changes between writers, unless they are the result of some acute situation. However, there has been no further analysis concerning these changes. In fact, the same handwriting analysis varies by the metrics we focus upon, as follows:—

- difference in handwriting by different scribes
- change which occurs over time, in handwriting by a specific scribe
- deviation in handwriting by a specific scribe during a specific period
- deviation in handwriting by the same scribe affected by circumstances (e.g. how busy he was; how certain he felt when writing / copying that the information he was writing was the final version; the surface on which he was writing: e.g. on writing desk / on music stand on his harpsichord.)

Fig. 6 shows the variable ranking for each classification estimated on the basis of mean decreased accuracy using out of bag data. Even though both these classifications concern changes which occur overtime, in handwriting by Bach himself, there are significant changes in the ranking. For instance, the difference in the importance of zenrike moments implies the importance of feature concerning angles. We can combine this analysis with the implementation of higher level features such as the length of double bars to clarify the difference between the above three changes. This leads to meaningful knowledge discovery for handwriting experts.



(a) Variable ranking for two-class. (b) Variable ranking for eight-class.

Fig. 6. Variable ranking on the basis of mean decreased accuracy estimated using out of bag data.

VIII. Dating lost autograph

This section explores how the proposed method can be applied to problems in a musicological context. As one such example, we demonstrate dating of lost manuscripts, only copies of which have survived.

Bach's autograph manuscripts were often duplicated by his copyists including his wife Anna Madalena (AMB). After the act of copying, some information in the original contents, particularly that often considered unessential such as calligraphic features of the beaming and stemming, were lost. However, AMB preserves these features more closely than many other Bach copyists; as a result; her copies look so similar to JSB's original that even specialists often have difficulty in distinguishing between them. Therefore, it may be possible to date a lost autograph by carefully studying AMB's copies.

For this experiment, 10 C-clefs were extracted from the d-minor prelude and fugue from Ms. 35021, which is believed to be in AMB's hand. These C-clefs were classified by RF which was trained in the first experiment using the dataset shown in Table I.

The result shows that 80% of the C-clefs were classified as F (1739) and 20% were classified as C (1723) (Fig. 7 shows clefs extracted from each manuscript). Assuming that all the clefs were copied from the lost autograph manuscript in the same period, it is highly likely that the manuscript was copied from the autograph manuscript which was written in the F period (1739). These results can be used as quantitative evidence to check musicologists' hypothesis. Contemplating the quantitative evidence which does not support the hypothesis, could reveal defects in the hypothesis, thereby preventing oversights. In this case, the 20% classified as C should be investigated further, to elaborate the hypothesis. For example, we can attribute 20% to the incomplete accuracy of the classifier (it misclassifies 5% of F as C) or the changes caused by the process of copying by AMB. It should be noted that the proposed algorithm can yield numerical evidence with uncertainty to be fused with other pieces of evidence, rather than hard factual conclusions. In the case of historical musicology, this is practical and compatible decision making, as rough approximation of imperfect evidence may lead to

TABLE II
COMPARISON OF ACCURACY AND AUC FOR SEVERAL CLASSIFIERS EVALUATED BY 10-FOLD CROSS-VALIDATION (CI= 99.95%).

Dataset	RF	Boosting	Bagging	SVM	K-NN	Bayes
two(accuracy)	100.00±0.00	99.90±0.26	99.93±0.22	92.61±2.13 ●	97.94±1.09 ●	94.49±1.76 ●
eight(accuracy)	72.53±3.31	68.97±3.45 ●	70.11±3.34 ●	65.67±3.14 ●	63.30±3.57 ●	66.02±3.51 ●
Average	86.27	84.44	85.02	79.14	80.62	80.26
two(AUC)	1.00±0.00	1.00±0.00	1.00±0.00	0.95±0.02 ●	1.00±0.01	0.99±0.01 ●
eight(AUC)	0.92±0.03	0.89±0.05 ●	0.91±0.04	0.88±0.05 ●	0.86±0.06 ●	0.89±0.06
Average	0.96	0.95	0.95	0.92	0.93	0.94

○, ● statistically significant improvement or degradation evaluated by the corrected paired t-test.

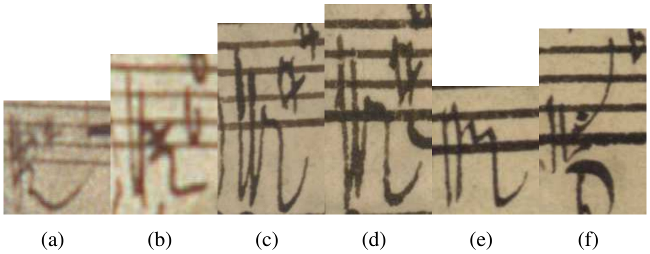


Fig. 7. Comparison of the C-clefs by two different writers; (a)(b) By JSB (1723); (c)(d) By JSB (1739); (e)(f) By AMB.

meaningless conclusions in the end — especially conclusions derived from only limited pieces of evidence.

IX. CONCLUSION AND FUTURE WORK

This study indicates that prior work in writer identification can lead to more meaningful knowledge through the interaction of domain experts with automated tools. However, some limitations are observed. Although some processes are automated, the extraction of C-clefs still requires many corrections by domain experts when manuscripts qualities is inadequate. Therefore, applying this method to large data can be prohibitively expensive. Further investigation is needed to improve our method for music symbol extraction and classification, to allow use of low-quality microfiche such as used in this study. The accuracy of C-clef classification can be improved by investigating the incorporation of digitised musical knowledge, and this should be explored in collaboration with musicologists. Another limitation is the assumption made in this study, that all the clefs from the same page were written in the same period. There is deviation in shape even in the clefs on the same page and sometimes they look as if they were added subsequently or even possibly by a different hand. This level of analysis requires more sophisticated image processing and feature extraction techniques that are capable of handling more subtle changes in each music symbol.

Establishing the chronology of handwriting is not a straightforward task. In contrast to OMR systems, musicologists would take a complex approach, taking account of chronological, compositional, and notational information, placing these against the historical background of the source. The latter includes the situation under which the initial copying and revisions took place, the diplomatic polices that might reveal the purpose for which the score was made, and so on, to verify the initial hypothesis. This situation justifies the significance of the proposed method to produce numerical and uncertain evidence from music manuscripts under the novel database scheme (Fig. 1), which can handle uncertain reasoning with multiple pieces of imperfect evidence. C-clef study is a starting point, and we suggest investigating many other symbols in the same manner, first extending to symbols such as other clefs (G-clefs and F-clefs), quaver rests, quavers, minims (esp. where the stem is connected to note-head), C (common-time signature, and Cut-time with a vertical stroke), flagged notes (quavers, semiquavers), and accidentals (sharp, flat, natural), then to extend those that he did not explore but valuable, e.g. numbers 1, 2, 3, 4, 6, 8 that appears in time-signature. Once the system is able to process and make sense of all these individual symbols, the system may then be enhanced to cover combinations of symbols as the manifestation of a scribe's engagement with notation, which may be detected in the way the placement of various symbols are negotiated. Future work should include a data fusion method to combine the numerical evidence obtained with domain knowledge. It is hoped that quantification and statistical analyses such as demonstrated in this paper will be significantly enhanced in future research, and that they can be adopted by future musicologists to discover many more exciting facts hidden deep in the beautiful manuscripts of Johann Sebastian Bach. Moreover, we hope that the result of this study will encourage a paradigm shift in the current data-centered data mining to domain driven data mining in a wide range of domains, such as this.

ACKNOWLEDGMENT

This work is financially supported by the Nakajima foundation.

REFERENCES

[1] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[2] D. Bainbridge and T. Bell, "The challenge of optical music recognition," *Computers and the Humanities*, vol. 35, no. 2, pp. 95–121, 2001.

[3] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. S. Marcal, C. Guedes, and J. S. Cardoso, "Optical music recognition: state-of-the-art and open issues," *International Journal of Multimedia Information Retrieval*, Mar. 2012.

[4] S. He and L. Schomaker, "Beyond ocr: Multi-faceted understanding of handwritten document characteristics," *Pattern Recognition*, vol. 63, pp. 321 – 333, 2017.

[5] C. Zhang, P. S. Yu, and D. Bell, "Introduction to Domain-Driven Data Mining Special Section," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 6, pp. 753–754, 2010.

[6] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17–37, Jan 2017.

[7] M. Niitsuma, L. Schomaker, J.-P. V. Oosten, Y. Tomita, and D. Bell, “Musicologist-driven writer identification in early music manuscripts,” *Multimedia Tools and Applications*, vol. 74, no. 9, pp. 1–19, May 2015.

[8] G. v. Dadelsen, *Beiträge zur Chronologie der Werke Johann Sebastian Bachs*. Tübinger Bach-Studien, vol. 4–5. Trossingen: Hohner, 1958.

[9] R. Huber and A. Headrick, *Handwriting identification: facts and fundamentals*. CRC Press, 1999.

Masahiro Niitsuma is currently an assistant professor at Ritsumeikan University Japan, working on sound language processing and music informatics research. Before he came to Ritsumeikan, he received a BS in Information Technology and a MS in Computer Science from Keio University Japan, and Ph.D in Creative Arts from Queen’s University Belfast. He has published numerous articles on computational analysis on Baroque music, and is currently exploring personality in movement patterns and their effects on musical experiences. He is interested in more expert-oriented data mining from musical performance and music manuscripts, from both academic and musical perspectives to achieve ‘deeper’ knowledge discovery.

Yo Tomita is Professor of Musicology at School of Arts, English and Languages at Queen’s University Belfast, Northern Ireland, and Senior Fellow of the Bach-Archiv Leipzig. He received his doctorate from the University of Leeds in 1991 with a dissertation on the sources of Bach’s The Well-Tempered Clavier, part II. He has published widely in Bach studies, from those pursuing to identify Bach’s compositional choices and decisions as manifested in Bach’s own scores to the reception history of Bach’s music in the late 18th and early 19th centuries. His recent publication includes Bach: The Baroque Composers (Fahnhan: Ashgate, 2011) and Exploring Bach’s B-minor Mass (Cambridge University Press, 2013), and is currently working on a two-volume monograph The Genesis and Early History of Bach’s Well-tempered Clavier, Book II: a composer and his editions, c.1720-1850 (Routledge), and The Cambridge Bach Encyclopedia (Cambridge University Press).

Wei Qi Yan is an associate professor with the Auckland University of Technology (AUT); his expertise is in digital security, surveillance, privacy and forensics, and he is leading the Computer and Cyber Security (CCS) Research Group at AUT. Dr. Yan is the editor-in-chief (EiC) of the International Journal of Digital Crime and Forensics (IJDCF); he was an exchange computer scientist between the Royal Society of New Zealand (RSNZ) and the Chinese Academy of Sciences (CAS), China, he is the chair of ACM Multimedia Chapter of New Zealand, a member of the ACM, a senior member of the IEEE, TC members of the IEEE. Dr. Yan is a guest (adjunct) professor with PhD supervision of the State Key Laboratory of Information Security (SKLOIS), Chinese Academy of Sciences, China.

David Bell graduated in 1969 in Pure Mathematics, and has three research degrees in Computing topics. He has been a full professor since 1986 – at Queen’s University, Belfast since 2002, where he is now Professor Emeritus and Visiting Research Professor.

He has produced several hundred publications, and supervised about 40 PhDs to completion. He was prime investigator on a large number of national projects and on many EU-funded programmes (eg MAP, ESPRIT, DELTA, COST, AIM,...) in IT since 1981.

His activities have included: PC chairmanship (eg joint Programme Committee Chair of VLDB’93 and ICDE’97), other international conference PC memberships and journal editing/ editorial board memberships — eg Computer Journal and North-Holland’s Information Systems. He has been guest editor of well-known journals such as IEEE Trans KDE, on (eg) Knowledge Discovery, Data Driven Data Mining and Semantic Web. Prof Bell is also Author/Editor of several books, including a ‘popular science’ book published this year — ‘Superintelligence and World-views’. He served on the UK Technology Foresight Panel for several years, and was member of a number of national and international advisory and funding groups, including, for example, a sub-group of the Wellcome Trust in the UK. His research interests are centred on data and knowledge management — the linking of reasoning under uncertainty, machine learning, and other artificial intelligence techniques with database work.