Univerzita Karlova v Praze
Matematicko fyzikální fakulta
Studijní oddělení – doktorské studium

Review of doctoral thesis

Optical Recognition of Handwritten Music Notation

submitted by **Jan Hajič, Jr.**

The thesis deals with Optical Music Recognition (OMR) that is an important part of attempts to digitize historical heritage and convert documents of various genres to digital form. It is on the intersection of computer vision (CV), music theory, and general computer science (graph theory, etc). The thesis has 5 introductory chapters, and then it is based on publications, with 11 conference and journal papers each bearing a short introduction and quantification of candidate's contribution, 191 pages in total. This review first deals with the technical content of the thesis, then summarizes its technical quality, comments on the formal points and finally presents overall conclusion and recommendation to the PhD committee.

Technical content of the thesis and remarks to chapters

Chapter 1 introduces the reader into the topic of the thesis and motivates it. From the very start, it clearly distinguishes the musical content (semantics) useful for querying and replaying, from music notation, useful for reprintability. This division influences the rest of the terminology and many design choices and I appreciate that it is done clearly at the beginning.

Chapter 2 reviews necessary terms of music notation. Despite being a practicing musician, I found this part very useful, especially regarding the English notation terminology, and problems, that might (and will) arise when detecting different graphical primitives in (especially hand-written) scores. This chapter also intuitively defines the relations between individual elements.

Chapter 3 provides the basics of OMR, defining the 4 main blocks of traditional OMR systems and reviewing the available literature. Given the limited size of the OMR community and its late start (compared to other sub-domains of computer vision and machine learning), it is probably not far from covering all the essential literature on the topic, which is quite a rare situation. It also critically reviews the resources available for OMR, namely the lack of standard data, that is crucial for any serious scientific work. The author can be credited for creating an important resource – the MUSCIMA++ manual annotations – that will definitely serve the community for many years. As a whole, this chapter documents the depth and width of candidate's knowledge and his passion for the topic of his research. On the other hand, at the end of each "overview" section, I would appreciate a short summary of really important things that steered the author in his work.

Chapter 4 presents the main contributions – of these, the Music Notation Graph (MuNG) is probably the most interesting from the "computer science" point of view, however, its description is not very detailed, and the following publications do not fill this gap – it would be nice to inform the reader (in a tabular form) on the possible relations between the elements in the tree structure. The chapter also mentions the MUSCIMA++ data-set that is very valuable itself and that actually allowed for most of the R&D in the thesis. I appreciate the "DYI" approach the author adopted when creating this set, including writing own tools such as MUSCIMarker! Concerning the object detection and notation assembly stages, the author does not bring new CV paradigms, but rather creatively uses what is around, however, it is clear that the choices of what to choose were driven by his knowledge of the CV and machine learning domains and familiarity with the goals. Lots of work was also devoted to evaluation methodologies and actually performing the evaluations – I can only agree with this choice: especially in a field that is lacking standard resources and common tasks, this is an activity that deserves appreciation. Finally, I greatly value author's contribution in building the OMR community and educating colleagues and students on the reality of OMR.

Chapter 5 contains the main conclusions while all papers presented in the 2nd part of the thesis contain their partial discussions and conclusions.

Next come the comments on attached publications: Section 6.1 covers an introduction, mainly for the musicians' community, on the basics of OMR. It excellently defines the taxonomy of tasks, building blocks of an OMR system, SotA and open tasks. From a practitioner's point of view, I was missing at least some discussion on the metrics, and especially a short section on "what can one expect from nowadays OMR systems".

The paper in section 6.2 presents the MUSCIMA++ data-set and details the underlying hard work on it. While the score material is well presented, I would appreciate more details on the relations in the resulting graphs (see the same comment above) and better description of experiments – section V of the paper is very "hermetical" and hard to understand, and the author could have thought twice whether to include it or not – the paper is very valuable even without it.

The short paper in section 6.3 complements the information given in 6.2, but does not answer many questions, rather raises new ones about the annotation methodology (do the annotators first tag the objects and then their relations?) and again, the relations in the graphs (how are the "one to many" or "global" relations such as clef or key signature vs. notes handled ?).

Section 6.4 advocates "evaluating the evaluations" in OMR by defining a scheme allowing to correlate individual metrics to human judgements. I appreciate the simplicity the "A vs. B" tests not requiring any grading scales, etc. On the other hand, I was lacking a more precise description on how the "trials" were generated (what was the distribution of errors ? One type or more types of errors inserted ? How many of them ?) and I also missed a discussion on how to interpret the results of the assessment – do the coefficients obtained for TEDn mean that TEDn compares well with human judgement or not ? Could you compare for example with similar works judging human evaluation vs. BLEU score in machine translation ?

The following paper (6.5) complements the previous one but does not bring much more "flesh", as a position paper for a discussion or panel session it is definitely good, but I am not convinced about its use in the thesis.

The following sections aim at the methods developed for OMR: all of them advantageously use the MUSCIMA++ data-set. Section 7.1 starts with detection of note-heads with region-proposal convolution NNs, working in three steps. The paper presents a nice work, but (similarly as above), I was lacking a discussion whether precision=0.81 and recall=0.97 are rather good or rather bad, and how much can be gained by the follow-up note assembly stage, probably able to correct many errors. Also, I would appreciate more intuitive explanation of the properties of note-head proposal filter (section II C).

Section 7.2 presents a short but important paper on another CV paradigm for detecting all necessary symbols in one step, the U-Net. This is further elaborated in 7.3 where the full pipeline is presented, including notation assembly stage. Several things should be made more clear in the paper, for example, it is not clear, how the multi-channel training was done (were different output layers used ?) and it not clear whether the result is monophonic. As a whole however, the paper presents promising results despite its concentration on symbol detection and very simple notation assembly stage.

Section 7.4 presents an exhaustive paper with simple (and efficient) evaluation based just one bounding boxes, working on three datasets with three different CV techniques. It contains wealth of results and many well documented, and discussed examples. It is a pity that the paper concentrates on object detection only – while it is interesting for the CV community, it might sound rather depressing for the musicians (especially regarding the hand-written MUSCIMA++ set) – for them, it would be nice to present results with post-processing.

The publication part is concluded by two short papers – 7.5 advocates the use of OMR in digital libraries and describes a case study in musicology solvable with OMR (although rather rudimentary, it is a nice demonstration of how OMR could help even now) and 7.6 is a short paper accompanying the presentation of MUSCIMarker in a demo session.

Summary on the technical content of the thesis

I have some critical comments on the thesis (mainly missing technical details such as resolution of images, important for "engineering" comprehension of the text) but as a whole, the thesis clearly demonstrates the qualities of the candidate – capability to study non-trivial literature from several fields, suggest own novel solutions, implement them, carefully test and discuss the results. I highly appreciate the quantity and quality of experiments done on different data-sets and the fact that the data and software is publicly available – this contributes to the *trust* one can have in the experimental results.

In a field that is relatively narrow, it can be easy to obtain outstanding results also by lower standard scientific work, but this is definitely not the case for the candidate – his work meets the highest criteria of serious scientific work and publications. The author also gives an impression of a great team worker (judged from numerous collaborations and joint papers).

Comments on the formal aspects

The thesis is written in a nice, literary and almost error-free English and its structure is logical and easy to follow, with some isolated exceptions mentioned above. The mathematical writing is correct, sometimes, the candidate prefers verbal expression of what is done – this is fine for literature surveys, but in the technical sections describing own development, I would sometimes appreciate a bit more "flesh", especially in the sections on music notation graph. The figures and schemes are well executed (except for the occasional unreadability of legends). There is a limited number of typos and grammatical errors, the candidate will receive a commented version of the document to help her fix these problems, in case corrections are allowed for the final publication of the thesis.

The selected form of thesis based on publications is not usual, and has advantages and drawbacks: while it clearly places author's work on a time-line, makes it possible to clearly say what was done in cooperation with other researchers (and give them credit) and allows to say "we were the first to …", it also requires the reviewer to read several times the same introductory texts, limits space for important information and takes much longer time to read. The chosen form is at the discretion of the author, but at the institution/department level, I would rather lobby for the classical form of dissertations.

Summary and recommendation

I have carefully examined the doctoral thesis of Mr. Jan Hajič Jr. Despite the criticism raised above (many points are rather recommendations than critique), in my opinion, it is a solid work that contributes to progress in OMR and in related research fields (computer vision, structured representation of documents, information retrieval and preservation of cultural heritage). I also examined his publication track including OMR related works and previous NLP publications, and I find it exceeding the standards for a PhD candidate at a respected University.

To conclude, I do recommend accepting the Thesis as a partial requirement for granting Mr. Jan Hajič Jr. the Doctoral degree at the Charles University in Prague.

For the defense, I have the following **questions**:

1. I did not find any information on the use of "language models" that could cover historical periods, styles of music, composers, down to individual compositions. Did you use them in your work and/or do you know about their results in the works of others ?
2. State clearly, what was the document resolution in dpi and relate the dimensions of NN input layers to the usual sizes of individual graphical elements.
3. Provide more details on introduction of notation errors in assessing the metrics – see comments to section 6.4.
4. Detail, how the multi-channel training was done (section 7.3).
5. Provide a summary of practical usability of OMR, for example in a table with rows corresponding to difficulty of musical notation and columns standing for type-set and hand-written music.

In Brno, May 28th 2019

Dr. Jan "Honza" Cernocky, Associate Professor
Head of Department of Computer Graphics and Multimedia
Responsible of BUT Speech@FIT group
Faculty of Information Technology, Brno University of Technology
Bozetechova 2, 612 66 Brno, Czech Republic
Tel: +420 5 41141284 Fax: +420 5 41141290,
mailto:cernocky@fit.vutbr.cz, http://www.fit.vutbr.cz/~cernocky
http://www.fit.vutbr.cz/ https://cs-cz.facebook.com/FIT.VUT
http://speech.fit.vutbr.cz https://www.facebook.com/BUT-Speech/