

INTERPRETING RHYTHM IN OPTICAL MUSIC RECOGNITION

Rong Jin

School of Informatics and Computing
Indiana University, Bloomington
rongjin@iemail.iu.edu

Christopher Raphael

School of Informatics and Computing
Indiana University, Bloomington
craphael@indiana.edu

ABSTRACT

We present a method for understanding the rhythmic content of a collection of identified symbols in optical music recognition, designed for polyphonic music. Our object of study is a measure of music symbols. Our model explains the symbols as a collection of voices, while the number of voices is variable throughout a measure. We introduce a dynamic programming framework that identifies the best-scoring interpretation subject to the constraint that each voice accounts for the musical time indicated by the known time signature. Our approach applies as well to the situation in which there are multiple possible hypotheses for each symbol, and thus combines interpretation with recognition in a top-down manner. We present experiments demonstrating a nearly 4-fold decrease in the number of false positive symbols with monophonic music, identify missing tuplets, and show preliminary results with polyphonic music.

1. INTRODUCTION

Throughout the history of the ISMIR community symbolically-represented music has figured prominently in a wide variety of applications, analysis techniques, as well as search and retrieval schemes. In spite of this demonstrated need, symbolic music data are still in short supply, greatly limiting the scale and variety of scientific music research. For music in machine-generated common Western notation, optical music recognition (OMR) provides, in principle, a direct path to create rich and extensive symbolic databases, thus OMR is among the most important problems for the classically-oriented music scientist. Significant advances in core OMR technology would lead to large scale symbolic music libraries, digital music stands, content-based search, as well as many specific applications well-known in this community.

OMR is a deeply challenging problem, well-known to ISMIR stalwarts [1–3], though less well-represented over recent years in published research. Blostein and Baird [4] present a 1992 OMR overview that is not so different from

a more current description [5]. One does not work long in this domain without encountering longstanding themes and conflicts of recognition science.

Our strong bias is for *top-down* recognition: approaches that begin by clearly articulating the world of possible hypotheses or answers, then scoring these hypotheses according to their *a priori* plausibility and their agreement with the data. The HMM approach to speech recognition is one of the most famous and successful examples of this kind, combining top-down modeling with computationally powerful dynamic programming (DP) techniques for search and training. However, the real-world recognition problems admitting feasible top-down approaches appear to constitute a small minority. All OMR approaches we know, except [3, 6], instead proceed *bottom-up* — beginning with the image data, gradually trying to piece together progressively higher-level constructions, ultimately concluding with the overall interpretation of the data. The preference for bottom-up strategies by nearly all OMR researchers (including ourselves) stems from their computational feasibility. The hallmark of a bottom-up approach is a series of *intermediate* and *irreversible* decisions as one climbs the ladder connecting the image data and its interpretation. The Achilles’ heel of the bottom-up paradigm is the inevitable incorrect intermediate decision constituting a blind alley that cannot be retraced. In OMR, the most obvious example would be an incorrect segmentation of the data leading to unrecognizable symbols, though many others exist.

The greatest virtue of the top-down recognition approach is its simultaneous focus on recognition and interpretation — primitive hypotheses such as “note head here” are only considered when they fit into a meaningful interpretation of the scene (say measure) at hand. While it seems too optimistic to hope to formulate OMR in an entirely top-down manner, there are many sub-problems where one can employ this philosophy. We do this whenever possible. While we begin bottom-up, seeking various self-contained objects without regard for their overall organization, each individual search procedure is itself top-down. For example, we find isolated chords (including single notes) by exploring candidate stem locations through grammar-induced DP strategies that consider *every* meaningful chord presentation and result in a globally optimal interpretation. Similarly, we recognize a beamed group by building a grammar of the possible presentations and computing the globally optimal meaningful structure. Further details can be found in [7].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

© 2012 International Society for Music Information Retrieval.

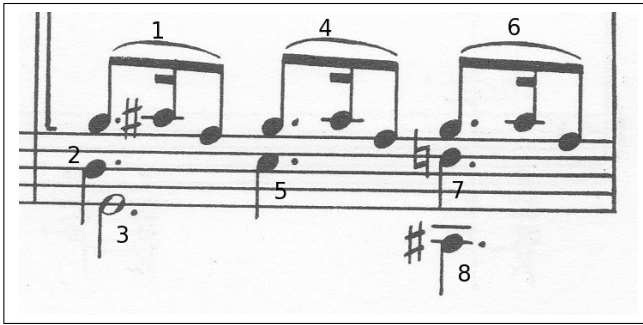


Figure 1. Numbering of symbols for polyphonic rhythm decoding.

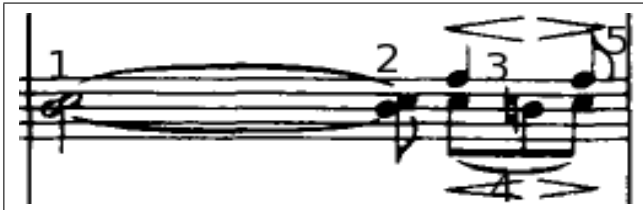


Figure 2. Example of a voice split.

The result of this process is a collection of mutually inconsistent and overlapping hypotheses that share “body parts” in impossible ways — this is a typical pitfall of a bottom-up approach where it is difficult to formulate the concept of a hypothesis’ unique “claim” to a particular image pixel. As described in [7], we resolve these conflicts by a phase seeking non-overlapping variants of the identified objects, completely discarding some of them, resulting in a collection of separate symbols that best explain the image data. This is where our present discussion begins.

An OMR system may attempt many different levels of music understanding. The most superficial approach would only record the primitive symbols (note head, stem, ledger line, etc.) found on the page, “punting” on any deeper interpretation of their meaning. Many levels of increasing depth could conceivably be added to this baseline. While we remain uncertain about the right depth of interpretation for our OMR system, it is hard to imagine a useful approach that does not understand rhythm and pitch. Without such interpretation, we cannot even play back the music, significantly limiting the value of the resulting symbolic data. Of these problems rhythm is, by far, the more challenging one.

In the simplest case — single voice music with no unmeasured notes — rhythm interpretation is rather straightforward: symbols can be clearly ordered from left to right, with the onset time of each symbol within a measure given as the sum of all preceding symbol durations. This situation quickly breaks down when the music uses multiple voices on a single staff as in Figure 1. Identifying the onset times here requires that we partition the symbols into three simultaneous voices, thus allowing the application of our monophonic strategy to each voice. Unfortunately, the number of voices is not known *a priori* and frequently changes throughout the duration of the measure as in Fig-

ure 2. In this work we propose a method of rhythmic interpretation that understands the music in terms of voices, allowing voices to be added or dropped anywhere in the bar.

Of course, this rhythmic understanding is an essential part of the symbolic data we seek to create, thus valuable in its own right. However, the process of understanding rhythm can be *combined* with the recognition process to improve the accuracy of our recognized results. The simplest example of this idea would leverage the “time signature constraint” — the note values in a voice must add up to the time signature when viewed as a rational number. For example, [8] has used this idea as a post-process to identify potential recognition errors. In this context we consider *multiple* hypotheses for each recognized symbol, choosing the best scoring *overall* measure interpretation obeying this constraint. More generally, we present a model for the possible polyphonic presentations of a measure, seeking the best scoring interpretation generated by the model, given our recognized symbols. The most significant contribution of our present work is a top-down approach for rhythm interpretation that *integrates* recognition with higher-level understanding.

2. RHYTHM DECODING

2.1 Monophonic Rhythm Decoding

The basic processing unit of our system is the measure, whose identification is discussed in [7]. In order to recognize the contents of a measure, we must both correctly segment the measure into meaningful pieces while interpreting the meaning of each piece. We first treat the case of *monophonic* music, here meaning that the notes and rests form a single stream of events. The most typical example would be music played by an instrument that produces a single note at a time, though our approach also applies to sequences of chords, as long as the notes of each chord share a stem.

Suppose we have partitioned the monophonic measure into a sequence of K symbols that can be unambiguously ordered from left to right. These objects could be rests, isolated notes, beamed groups, as well as objects without associated musical time such as clefs. If extraneous symbols, not corresponding to actual document symbols, have been (mis)recognized, it won’t matter how these symbols figure in this ordering.

We let S_k be a collection of possible *interpretations* for the k th object. For instance, for a rigid isolated symbol such as a rest, we consider all possible position and label hypotheses, retaining the best scoring position for each label (quarter rest, eighth rest, etc). In the case of an isolated note, our recognition result may involve a closed note head, though an open note head may match nearly as well. Or perhaps we were uncertain about the number of augmentation dots or flags attached to the note. We revisit the DP analysis of the note, modifying the trace-back phase to create the “N-best” interpretations of the object [9]. Thus our isolated note analysis produces a list of possible inter-

pretations along with scores measuring the quality of fit to the image data.

Similarly we construct an N-best list for the interpretations of a beamed group. These hypotheses may differ in the number of beams that connect any pair of adjacent notes, existence of partial beams, or number of augmentation dots attached to a note. In summary, the input to our rhythm recognizer is an ordered list of K objects, each with collection of possible hypotheses, S_k . For each $s_k \in S_k$ we let $D(s_k)$ denote the musical time consumed by the hypothesis, with recognition score $H(s_k)$. Our convention for musical time gives a quarter note duration $\frac{1}{4}$, and eighth note $\frac{1}{8}$, with similar rational numbers for other notes, rests, or beamed groups. In each collection, S_k we include the “null” interpretation with duration and score 0, corresponding to the case of a false positive recognition error.

In many cases we find that the best scoring hypotheses collectively make rhythmic sense. That is, we find that $\sum_k D(\hat{s}_k) = T$ where $\hat{s}_k = \arg \max_{s_k \in S_k} H(s_k)$ and T is the measure’s time signature represented as a rational number (e.g. $T = \frac{3}{4}$ for 3/4 time). In such a case there would be no reason to consider any other possible interpretation of the symbols. However, it is common to encounter scenarios where the best scoring hypotheses do *not* “add up,” while the correct interpretations of some symbols are found “further down” in the hypothesis list. In such a case it makes sense to look for the sequence s_1^*, \dots, s_K^* with $s_k^* \in S_k$ given by

$$s_1^*, \dots, s_K^* = \arg \max_{\sum_k D(s_k) = T} \sum_k H(s_k)$$

where the maximum is taken over all sequences s_1, \dots, s_K with $s_k \in S_k$ for $k = 1, \dots, K$.

The identification of this optimal configuration is a simple exercise in dynamic programming. To this end we let P_k denote the possible measure positions for the k th object:

$$P_k = \left\{ \sum_{k'=1}^k D(s_{k'}) : s_{k'} \in S_{k'} \right\}$$

for $k = 1 \dots, K$, with $P_0 = \{0\}$. These are the “states” of the DP calculation. Then we initialize $M_0(0) = 0$ and recursively define

$$M_k(p_k) = \max_{\substack{p_{k-1} \in P_{k-1}, s_k \in S_k \\ p_{k-1} + D(s_k) = p_k}} M_{k-1}(p_{k-1}) + H(s_k) \quad (1)$$

for $k = 1, \dots, K$ and $p_k \in P_k$. The optimal path, s_1^*, \dots, s_K^* , has score $M_K(T)$ — it is a simple exercise to recover the path that generates the optimal score $M_K(T)$.

By enforcing the time signature constraint on our interpretation we guarantee that the result makes rhythmic sense and fix recognition errors in the process, analogous to the decoding of an error-correcting code.

2.2 Recognizing System Rhythm

A system groups together a collection of staves that are played simultaneously. Usually systems align symbols oc-

curing at the same musical time to the same horizontal position. For instance, corresponding bar lines of a system generally occur at a common horizontal position — in fact, our system recognizer identifies systems by partitioning the staves into groups having common bar line positions. As always with music notation, there are exceptions to this general rule, such as when whole rests are centered rather than “left aligned,” or when symbols must be offset from their idealized positions to avoid overlap, as with unison whole notes.

This alignment convention can be used to extend the idea of the preceding section by adding a term to the score function penalizing misalignment of simultaneous events. Suppose we begin with a system of L staves and write s_1^l, \dots, s_K^l with $s_k^l \in S_k^l$ for an interpretation of the l th staff. As a minor abuse of notation, we write $P(s_k^l) = \sum_{k'=1}^k D(s_{k'}^l)$ for the measure position of s_k^l , though clearly $P(s_k^l)$ depends on the entire history leading to s_k^l . For every pair of simultaneous rhythmic events in a system measure — that is, $s_k^l, s_{k'}^{l'}$ with $P(s_k^l) = P(s_{k'}^{l'})$, we penalize their misalignment by $Q(|X(k, l) - X(k', l')|)$, with some non-decreasing function, Q , where $X(k, l)$ gives the horizontal location of the k th event in the l th staff. For a rest, we would take this location to be the horizontal component of its center, while for an isolated note would take the horizontal component of the note head center, since this is normally what the layout tries to align. For beamed groups we simply use the horizontal component of the first note head center.

We now optimize the criterion:

$$J = \sum_{l=1}^L \sum_k H(s_k^l) - \sum_{\substack{l, l'=1 \\ l \neq l'}}^L \sum_{P(s_k^l) = P(s_{k'}^{l'})} Q(|X(k, l) - X(k', l')|) \quad (2)$$

subject to the usual time signature constraint on each measure. Due to the very large state space that would ensue, it may not be feasible to perform simultaneous optimization over all $\{s_k^l\}$ variables by DP. A computationally more tractable approach would be the familiar Gauss-Seidel or “coordinate-wise” optimization. That is, we first recognize each staff measure independently according to the technique of the previous section. Then we iteratively revisit each staff in turn, holding the interpretation of the other staff measures fixed while optimizing over the current staff measure. This calculation is possible since, when considering staff l' , the measure positions, $P(s_k^l)$, $l \neq l'$, are known since the s_1^l, \dots, s_K^l are fixed, while for $l = l'$, $P(s_k^l)$ is the DP state.

2.3 Missing Tuplets and Symbol Overloading

Rhythmic notation allows for various abbreviations that may not literally make sense, but are clear in context. Often the correct interpretation is reinforced by the horizontal alignment of coincident symbols, as in the previous section. For instance, it is common to omit the ‘3’ on

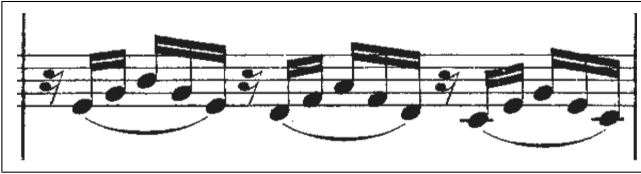


Figure 3. Example of implicit triplets. Our state model requires a triplet to begin on a beat and continue for entire beat before returning to duple rhythm or beginning another triplet.

a beamed group of three notes, when the triplet interpretation is obvious, though this convention also allows for mixing rests and notes implicitly grouped into 3's (or some other tuplet number) as in Figure 3. Another common abbreviated notation uses the half rest or whole rest to denote an empty measure even when the rest doesn't account for the correct bar length. While it may be literally correct to, for instance, write a dotted half rest for a blank measure of 3/4 time, there doesn't seem to be any possibility for misinterpreting the plain half rest, so the shorthand persists.

Examples such as the “overloaded” half rest are easy to treat with the preceding methodology. When the half rest appears as a possible interpretation of a symbol in 3/4 time, we simply add an identically-scored interpretation corresponding to the full length of the measure. The case of the missing triplet on a group of three beamed notes can be handled similarly, allowing both “straight” and triplet interpretations of the group (while in duple meter).

The same ideas can apply in the more complex missing triplets of Figure 3, where the implicit grouping involves several musical symbols. To do this we must multiply our state space by 2 allowing each state to occur in a “straight” and “triplet” version. When we are in a triplet state, all note values count for 2/3 their nominal length. We can leave the triplet state, reverting to the literal interpretation of rhythm, only when the measure position has no factor of 3 in the denominator (i.e. when the triplet is completed). We may also limit the places where triplets can begin (i.e. where we can transition from a non-triplet state to a triplet state) to quarter note or eighth note pulses.

2.4 Polyphonic Rhythm Decoding

As discussed in Section 1, the rhythmic intent of polyphonic notation is often ambiguous, deriving its meaning from implicit use of voices which may appear or disappear at any place within a measure. In this section we present an algorithm for the rhythmic decoding of a measure of polyphonic symbols. For clarity's sake we focus on the simplest statement of the problem, assuming correctly identified symbols, a single staff, and no missing tuplets. However, this technique can be extended using any of the ideas of the previous three subsections. For instance, the ideas of Section 2.1 can be included in an obvious way to cover the case where we have multiple rhythmic hypotheses for each symbol, as in Section 3, with analogous extensions for missing tuplets and staff measures.

We first consider the situation in which the number of voices, V , is known, while the voices persist throughout the entire measure. In such a case, the sum of rhythmic values over all symbols in the measure would be VT . Here the interpretation problem simply separates these symbols into voices, as is necessary for their rhythmic understanding. We begin by numbering the K symbols of the measure from left to right, breaking ties arbitrarily, as in Figure 1: we require only that the resulting sequence of the symbols' measure positions is non-decreasing. We represent a possible interpretation as a sequence of states, one state for each of the K symbols, where a state consists of three quantities for each active voice: the index of the voice's most recent symbol and two rational numbers giving the onset and offset times of the most recent symbol. For instance, the correct state sequence associated with Figure 1 would be:

	voice 1	voice 2	voice 3
1	(1, $\frac{0}{8}, \frac{1}{8}$)	—	—
2	(1, $\frac{0}{8}, \frac{1}{8}$)	(2, $\frac{0}{8}, \frac{1}{8}$)	—
3	(1, $\frac{0}{8}, \frac{1}{8}$)	(2, $\frac{0}{8}, \frac{1}{8}$)	(3, $\frac{0}{8}, \frac{6}{8}$)
4	(4, $\frac{0}{8}, \frac{1}{8}$)	(2, $\frac{0}{8}, \frac{1}{8}$)	(3, $\frac{0}{8}, \frac{6}{8}$)
5	(4, $\frac{0}{8}, \frac{1}{8}$)	(5, $\frac{0}{8}, \frac{1}{8}$)	(3, $\frac{0}{8}, \frac{6}{8}$)
6	(6, $\frac{0}{8}, \frac{1}{8}$)	(5, $\frac{0}{8}, \frac{1}{8}$)	(3, $\frac{0}{8}, \frac{6}{8}$)
7	(6, $\frac{0}{8}, \frac{1}{8}$)	(7, $\frac{0}{8}, \frac{1}{8}$)	(3, $\frac{0}{8}, \frac{6}{8}$)
8	(6, $\frac{0}{8}, \frac{1}{8}$)	(7, $\frac{0}{8}, \frac{1}{8}$)	(8, $\frac{0}{8}, \frac{8}{8}$)

This sequence is “legal” since all voices account for the number of beats expressed by the time signature (9/8), as seen by the 3rd member of each voice in the last row of the table.









Of course, the true state sequence is not known, in practice. We proceed by considering *all* possible state sequences, scoring them according to their plausibility in search of the best scoring candidate. In doing so we generate a search tree where the k th level of the tree treats the k th symbol in our list. At the k th level we expand each branch by adding the k th symbol to all possible voices, while scoring this extension according to several criteria. Perhaps the most important criterion is the degree to which musically coincident symbols align horizontally. When a new symbol enters a voice, we must first consult the state to see if it contains symbols sharing the new symbol's onset time. This is why the symbol's starting position is included as part of the state. For each such coincident symbol in the state, we compute the difference in horizontal position with that of the entering symbol. This is why the state also retains the index of the symbol. The state information can also be used to penalize the addition of a new symbol whose stem direction does not agree with that of the most recent symbol, etc.

The search proceeds over K iterations — one for each incoming symbol, generating a search tree in the process. Each iteration begins by expanding each surviving branch by adding the current symbol to one of the voices, or creating a new voice if available voices exist. These new hypotheses are then scored according to the criteria discussed above. At this point it is possible that we have generated multiple paths to the same state, and, if so, we only retain

the best scoring state. That is, we perform DP cutoffs. In doing so, the particular voice numbering is not considered relevant, so two states that differ only by the labeling of voice numbers are considered identical. After performing DP cutoffs, we may still need to prune the tree further to render the search feasible, retaining only the best scoring B hypotheses after each iteration.

Of course, it is not reasonable to assume *a priori* that we *know* the number of voices. For that matter, the number of voices may change throughout the duration of the measure. The most common instance of this phenomena occurs when a multi-voice measure begins or ends with a rest, in which case it is common to use a single rest for all voices. More generally, it is common to allow voices to come in, or go out, of existence when the resulting notation uses less ink and still suggests the right idea to the reader. Figure 2 shows an example where a voice is added midway through the measure (we regard stems with multiple note heads as a single voice).

We address this problem by adding some flexibility to our state production rules. Regardless of the number of voices, we begin each measure with a single voice. At the beginning of each iteration, any voice is allowed to split into two identical voices, as long as some maximum number of voices has not yet been reached. The incoming symbol is then allowed to extend any currently active voice. Additionally, any two voices sharing the same ending time can merge into a single voice. Since we want to discourage gratuitous use of these kinds of productions, we add a penalty term when they are invoked. The correct state sequence associated with Figure 2 is as follows:

	voice 1	voice 2
1	(1, )	—
2	(2, )	—
3	(2, )	(3, )
4	(4, )	(3, )
5	(4, )	(5, )

3. EXPERIMENTS

We tested the algorithm of Section 2.1 on the 2nd movement of the Mozart Quintet for Clarinet and Strings, K. 581. The original images of the four pages of this movement can be seen at

<http://www.music.informatics.indiana.edu/papers/ismir12>.

As with all experiments presented here, we begin by finding our best representation of the image data in terms of non-overlapping isolated symbols, isolated chords, and beamed groups. This phase implicitly segments the image into distinct objects. Using the N-best techniques discussed above, we then identify a list of possible interpretations of each symbol or symbol group, thus forming the input to our rhythm decoder. The best scoring hypothesis for each symbol is superimposed in blue in the referenced images. As discussed above, the collection of best scoring individual hypotheses may not make rhythmic sense, thus we seek the best scoring *meaningful* interpretation through our rhythm decoder.

symbol name	Best Score		Rhythm Decoding	
	False+	False-	False+	False-
solid note head	6/898	14/908	5/891	18/908
open note head	1/34	4/37	1/32	6/37
note stem	36/921	10/927	32/913	14/927
1 beam	4/429	9/434	3/427	10/434
2 beam	1/77	4/80	0/76	4/80
3 beam	1/90	2/91	1/91	1/91
aug. dot	113/153	1/39	3/38	4/39
single flag down	0/7	0/7	7/14	0/7
single flag up	0/9	3/12	0/12	0/12
double flag up	0/0	1/1	0/0	1/1
whole rest	27/27	13/13	1/14	0/13
half rest	30/30	0/0	2/2	0/0
quarter rest	12/36	1/25	1/25	1/25
eighth rest	12/30	1/19	2/20	1/19
16th rest	0/1	3/4	0/2	2/4
32th rest	0/6	0/6	1/7	0/6
total	243/2748	66/2603	59/2544	62/2603
decimal	.088	.025	.023	.024

Table 1. False positives and false negatives for each primitive symbol with and without rhythm decoding. The table shows a nearly 4-fold decrease in false positives with essentially no change in false negatives.

Each of these images was hand-marked with ground truth by identifying bounding boxes of the primitive symbols of Table 1, as well as some rhythmically neutral symbols (clefs, accidentals, etc.) that don't appear in the table. As can be seen from the images and the table, the original recognition contained many small false positive symbols such as augmentation dots and whole/half rests. From a statistical point of view, almost any data model will be prone to such "small symbol" errors, due to the higher variability of small-sample estimates. However, many of these unwanted symbols have only marginal data scores and do not appear in the best scoring measure hypothesis subject to the time signature constraint. In fact, the Table 1 shows a nearly 4-fold decrease in false positives with virtually no change in false negatives. False negatives, for the most part, cannot be corrected by our rhythm decoder, since they stem mostly from errors in which the correct hypothesis does not appear anywhere in our input to the algorithm.

The last page of the Mozart Quintet 2nd movement, visible at the website reference above, contains a number of unmarked triplets, as well as several marked ones, as in Figure 3. We tested the algorithm of Section 2.3 which includes unmarked triplets among the hypotheses that are considered. Since a number of the triplets on our page involve two symbols, a rest and two beamed notes, we must modify our state space in the manner described in Section 2.3, giving two versions of each rhythmic position: "triplet" and "straight." We recognized the page using the rhythm decoder of Section 2.1, both with and without accounting for unmarked triplets. When allowing for triplets we correctly recognized all of the triplets on the page, while the larger associated state space caused no additional errors on the measures that did not contain triplets. This is, of course, a small "proof of concept" experiment, rather than a large scale validation.

A final experiment treats the first page of the Rachmaninov *Etudes Tableaux*, op. 33 for piano, also displayed at the aforementioned web page. Piano music is particularly difficult for OMR, due to the higher symbol density, implicit uses of voices, as well as other idiosyncrasies of keyboard notation. However, the frequent use of implicit voices poses an appropriate challenge for our polyphonic rhythm decoder of Section 2.4 — most measures in the right hand of this page contain two voices.

Our goal now is *simultaneously* to choose from the available hypotheses for each object, and to explain the symbols' rhythm in terms of several possible voices. In this way we *integrate* recognition and interpretation, as is consistent with our philosophy, rather than treating them as distinct phases of OMR. While the page uses voices in a consistent manner (two for the right hand and one in the left), we do not assume this knowledge. Rather we assume a maximum of two voices that are allowed to come and go in each measure, as described in Section 2.4. Since we do not yet recognize time signatures, we assume the time signature is known for each measure.

Evaluating OMR in terms of symbol primitives, as in Table 1, is relatively straightforward and common in the OMR literature. We can imagine various useful notation applications based only on such primitive information, justifying a limited place of this kind of evaluation. However, we expect that most uses of OMR will require a higher level of music understanding than that expressed by symbol primitives. One possible approach to OMR evaluation represents each measure as a list of notes (or notes and rests), with each note having several attributes such as position within measure, length, pitch, coordinates of note head, etc. When both ground truth and recognized results are represented in this manner, a false negative can be identified as any note in the ground truth that cannot be “matched” with a note in the recognized results. Here a match requires agreement of *all* attributes of the ground truth note with one of the recognized notes. False positives are computed by reversing the roles of ground truth and recognized results. Since our current emphasis is on rhythm, we evaluate our approach in this manner describing each note in terms of its note head coordinates and rhythmic onset position within the measure. While not explored here, we believe this general evaluation paradigm (with suitable modifications) is serviceable in a wide range of OMR scenarios.

Using this procedure we achieved false negative and false positive rates of 30/402 and 8/380 on the Rachmaninov page. While evaluations in terms of musical quantities such as pitch and rhythm may better measure the usefulness of the OCR results, they don't clearly convey what actually goes wrong in recognition — in contrast, primitive evaluation is quite specific in this regard. On this one-page test, all errors were due to one of three things. One measure simply had misrecognized rhythm, however, the rhythmic result was quite syncopated, suggesting we may be able to further improve by penalizing unusual rhythm. Most of the false negatives were due to the second kind

of error — missing note heads on chords that were otherwise correctly recognized. Our approach cannot possibly recover from such errors. The last type of error results from the unusual figure in the left hand of measures 5-9, in which an eighth and sixteenth are beamed together with a sixteenth rest written in the *interior* of the beamed group. This situation violates our assumption that notes in beamed group are executed in sequence without the possibility of intervening notes/rests from other symbols. We believe this type of error could be corrected with simple modifications of our approach. As before, numerous false positives from recognition are corrected by this procedure.

4. REFERENCES

- [1] G. S. Choudhury, T. DiLauro, M. Droettboom, I. Fujinaga, B. Harrington, and K. MacMillan, (2000), “Optical Music Recognition System within a Large-Scale Digitization Project,” in *Proceedings, International Symposium on Music Information Retrieval*, 2000.
- [2] D. Bainbridge and T. Bell: “The Challenge of Optical Music Recognition,” *Computers and the Humanities* 35: pp. 95-121, 2001.
- [3] L. Pugin, J. A. Burgoyne, I. Fujinaga: “MAP Adaptation to Improve Optical Music Recognition of Early Music Documents Using Hidden Markov Models”: in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, pp. 513-16, Vienna, Austria, 2007.
- [4] D. Blostein and H. S. Baird: “A Critical Survey of Music Image Analysis,” In *Structured Document Image Analysis*, ed. H. S. Baird, H. Bunke and K. Yamamoto, pp. 405-34. Berlin: Springer-Verlag, 1992.
- [5] A. Rebelo and G. Capela and J. S. Cardoso: “Optical Recognition of Music Symbols,” in *International Journal on Document Analysis and Recognition* 13:19-31, 2009.
- [6] G. Kopec, P. Chou, D. Maltz: “Markov Source Model for Printed Music Decoding,” *Journal of Electronic Imaging*, 5(1), 7-14, 1996.
- [7] C. Raphael and J. Wang: “New Approaches to Optical Music Recognition” in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, Miami, USA, pp. 305-310, 2011.
- [8] F. Rossant and I. Bloch: “Robust and Adaptive OMR System Including Fuzzy Modeling, Fusion of Musical Rules, and Possible Error Detection,” *EURASIP Journal on Applied Signal Processing*, vol: 2007.
- [9] R. Schwartz, Y.-L. Chow: “The N-Best Algorithm: An Efficient and Exact Procedure for Finding the Most Likely Sentence Hypotheses,” *Proc. of ICASSP-90*, pp. 81-84, Albuquerque, NM, 1990.