Hearing Sheet Music: [+ a clever backronym I haven't come up with yet]

Stephen Miller
Stanford University
450 Serra Mall Stanford, CA 94305
sdmiller@stanford.edu

1. Introduction

For those of us who are amateurs, learning a new song (particularly on an instrument like piano with multiple parts) often requires slowly stepping through, measure by measure, learning the right hand, then the left hand, then how to combine the two. There eventually comes a moment where the alignment "clicks", and the intended melody becomes clear. From that point on, it's far easier to play.

In this project, I'd like to make a tool to help such learners along, by making an app in which a learner can point at a piece of sheet music at arbitrary perspective in front of a somewhat cluttered background, drag a cursor to designate which measure he/she would like to start from, and will hear it played in realtime.

2. Problem Statement

This is a concatenation of a few visual tasks. One is to locate printed sheet music and its corresponding staff in a real-world image, and estimate an affine transform which will bring it back to a canonical "flat" view. From there, individual detectors (likely SVMs, though the best feature for the task isn't clear yet – maybe shape context?) for particular notes, rests, time signatures, etc must be designed, and combined to form some globally consistent representation of the state of the music (under constraints such as, the sum of the length of the notes in a measure must equal the length of the measure).

These may be too noisy for a single image – but an iPhone gives the ability to use video data. Throwing these into an HMM (or some other temporal model) should improve the chances that a consistent signal will emerge.

Finally, a number of engineering tasks are involved to produce audio and display the cursor advancing over time.

3. Related Work

From a brief literature scan, this seems like a problem which has not been addressed. Academically, it is quite similiar to something which has been addressed, which I intend to research for the project: text recognition. The idea of estimating an affine transform and detecting individual components in sequence is analogous. However, many notes are far more similiar looking than characters, and certain traits (such as the dot after a dotted half note) will likely need to be inferred by global context of the measure, as much as from the image. Furthermore, certain detections (such as the time and key signatures, and the surrounding notes) must inform the others – it is the difference between an F in C and an F# in D, and three quarter notes in 3/4 vs a triplet in 444.

Commercially, tools are available which scan sheet music and convert it to midi form. Yet these rely on having an extremely clean image, subject to no affine transformations. They likely work by fitting very rigid templates, whereas this will need to be far looser to allow for noise / the imperfections of the camera / perspective / etc. Temporal consistency and local context will likely play a much bigger role.

4. Data

What data is used relies, largely, on how I go about it / how sequential the process is. Either way, there will be a dataset of images of sheet music, which I can collect by hand with an iPhone at any music store.

5. Evaluation

There is a clear metric of success here. Ignoring the MIDI aspect (which is crucial to the app but academically simple), the task here is to convert an image or sequence of images into a set of ordered notes. There is already ground truth for that, so it is easy to quantitatively analyze the global success – though something clever might be necessary to align the notes and score missed-detections appropriately. Individual measures for various detection components are also straightforward: accuracy of the affine transform estimated, a confusion matrix for different notes and symbols, etc. Qualitatively, the proof is in the pudding: can you point it at sheet music and hear a song?