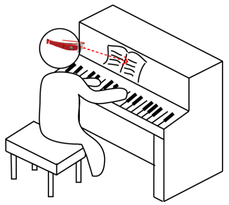
MIDI-Assisted Egocentric Optical Music Recognition

Liang Chen[1], Kun Duan[2]
[1]School of Informatics and Computing, Indiana University, Bloomington, IN
[2]GE Global Research, Niskayuna, NY

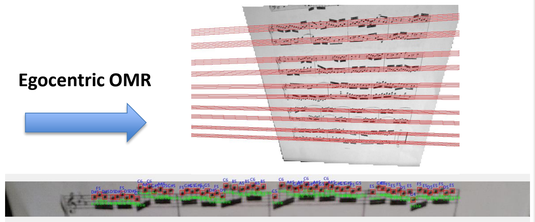
INDIANA UNIVERSITY

GE Global Research

1, Overview



Egocentric OMR



Automatic Score Reading with Egocentric Cameras, e.g. **Google Glass**

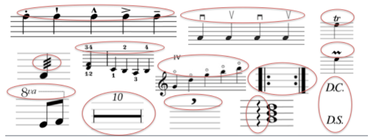
Interesting Applications:
1, Assistive sight-reading
2, Interactive egocentric games involving music

2, Challenges

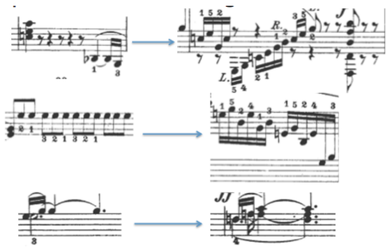
1, Image Distortion due to first-person view perspective
2, Degradation and Blur caused by camera motion
3, Music Interpretation Problem (OMR is **challenging**)
To summarize, egocentric OMR suffers from the challenges from both egocentric vision and offline OMR.

Why can't we directly apply off-the-shelf OMR software to solve the problem?
Given the state-of-the-art, OMR is still an **unsolved** problem. Both symbol recognition and music interpretation are **difficult**.

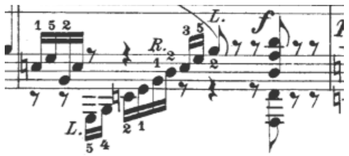
Long tail of uncommon symbols



General notation rules were always violated somewhere



Lack of paradigm to model 2D layout of music notation

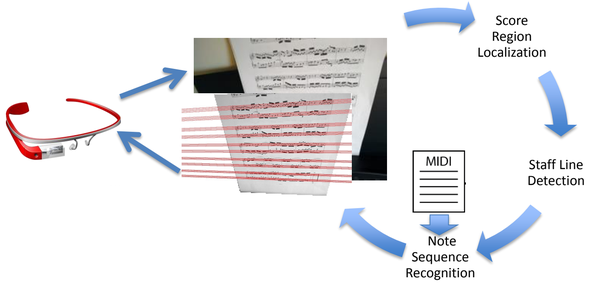


Ambiguous rhythm interpretation

3, Our Idea

Instead of blindly recognize and interpret music symbols, we assume the symbolic data (MIDI) is already known, and try to solve the problem by incorporating MIDI data into the recognition process.

Benefits:
1, Release the burden: recognition problem converted to alignment problem, which we have well-established paradigm to solve.
2, Reasonably constrain the output space: inference result is always musically meaningful.
3, Once the observation and MIDI segment is correctly matched, music semantics will be automatically parsed.



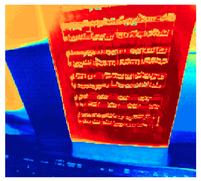
4, Preprocessing: Score Region Localization

1, Score boundary was modeled as straight lines – easy to be parameterized.
2, Data likelihood was learnt unsupervisedly with Gaussian Mixture Model (GMM).

Score Localization as foreground-background separation problem with a strong shape prior. Θ represents boundary parameter set.

$$\Theta^* = \arg\max_{\Theta} \sum_{(i,j)\in R_\Theta} D(p(i,j))$$

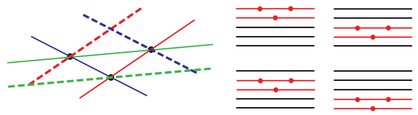
where $D(p(i,j)) = \log \frac{G_{fg}(p(i,j))}{G_{bg}(p(i,j))}$



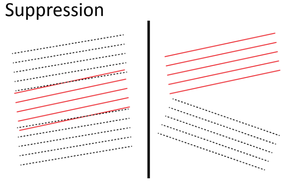
5, Preprocessing: Staff Line Detection

Staff-line detection via Random Sample Consensus (RANSAC)

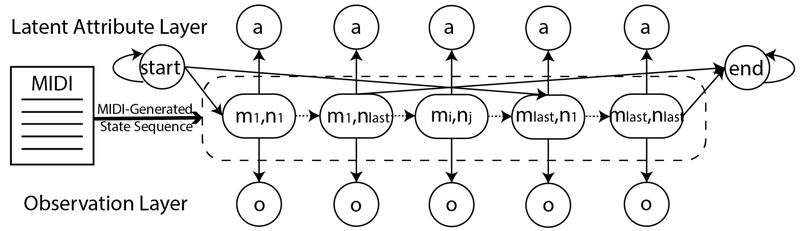
Model Proposal with Sampled Pixel **Triplet**
Model Evaluated with global votes



Model Thinning using Non-Maxima-Suppression



6, Conditional Random Field for Note Sequence Recognition



MIDI parsed into independent note events

Event	Pitch ID (Name)	Onset	End of Measure
1	48 (C3)	2.25	0
2	50 (D3)	2.50	0
3	52 (E3)	2.75	0
4	53 (F3)	3.00	0
5	50 (D3)	3.25	0
6	52 (E3)	3.50	0
7	48 (C3)	3.75	1

We want to infer the optimal states S* that best interprets the observation. Each state s contains one note event in MIDI, its associated notehead's parameters (location on the image) and latent music attribute that models uncertainty of the transform from symbolic to notation.

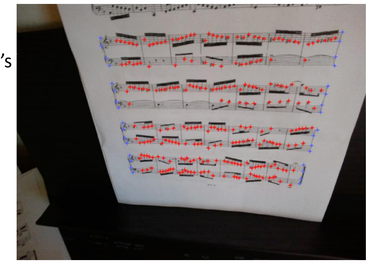
$$S^* = \arg\max_{\{s_i\}} E(s_i|X,l) + E(s_i, s_{i+1}|X,l)$$
$$= \arg\max_{\{s_i\}} E(n_i, x_i, y_i, a_i|X,l) + E(s_i, s_{i+1}|X,l)$$

Unary term of the CRF was learnt via HOG + linear SVM, while pairwise relations models the minimum distance between adjacent notes that have large enough Inter-Onset-Intervals (IOI).

7, Evaluation

Dataset:
54 egocentric images, each including 8 to 12 staves.
The original scores came from the first 5 pieces of Bach's 15 Inventions (No. 1 - 5).
Test set contains **242** independent staves.

Annotations



Staff line detection accuracy

	Precision	Recall	F-Score
Staff Detection	86.1%	81.8%	83.9%

Note sequence recognition evaluation

		Note Detection			MIDI Alignment		
Method	Measure Subsequence Accuracy	Precision	Recall	F-Score	Precision	Recall	F-Score
Greedy	14.1%	42.7%	82.6%	56.3%	27.0%	47.7%	34.5%
CRF	53.0%	85.3%	77.2%	**81.0%**	65.1%	67.1%	66.0%
CRF + Pairwise Constraint	54.0%	80.9%	78.6%	79.7%	68.7%	65.2%	**66.9%**

8, Acknowledgements

We would like to present our special thanks to Prof. David Crandall and IU Computer Vision Lab for providing the wearable device. This work used resources that were supported in part by the National Science Foundation under grant IIS-1253549.

COMPUTER VISION LAB