MIDI-Assisted Egocentric Optical Music Recognition

Liang Chen
Indiana University
Bloomington, IN
chen348@indiana.edu

Kun Duan
GE Global Research
Niskayuna, NY
kun.duan@ge.com

Abstract

Egocentric vision has received increasing attention in recent years due to the vast development of wearable devices and their applications. Although there are numerous existing work on egocentric vision, none of them solve Optical Music Recognition *(OMR) problem. In this paper, we propose a novel optical music recognition approach for egocentric device (e.g. Google Glass) with the assistance of MIDI data. We formulate the problem as a structured sequence alignment problem as opposed to the blind recognition in traditional OMR systems. We propose a linear-chain* Conditional Random Field *(CRF) to model the note event sequence, which translates the relative temporal relations contained by MIDI to spatial constraints over the egocentric observation. We performed evaluations to compare the proposed approach with several different baselines and proved that our approach achieved the highest recognition accuracy. We view our work as the first step towards* egocentric optical music recognition*, and believe it will bring insights for next-generation music pedagogy and music entertainment.*

1. Introduction

Egocentric vision becomes an emerging topic as *first-person camera* (e.g. GoPro, Google Glass) has gained more and more popularity. These wearable camera sensors have attracted a lot of computer vision researchers due to its wide range of applications [3]. Building these applications is, however, challenging due to various reasons such as the special observation perspective, blurs caused by camera motion and real-time computation request.

In the recent few years, egocentric applications have extended to many areas such as object recognition [8, 14, 19], video summarization [21] and activity analysis [7, 16, 22]. Similar with [8], we assume weak supervision is available to the recognition system. More specifically, we assume note sequences from the corresponding MIDI file is given, which provides useful information to direct the recognition process.

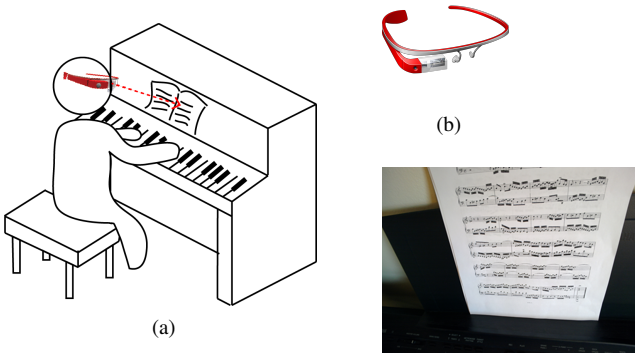


Figure 1: (a) Piano player with egocentric score reader; (b) Wearable camera; (c) First-person view score image captured by the device

To the best of our knowledge, there is no existing work on music recognition using egocentric cameras. One possible reason is the limitation of existing Optical Music Recognition (OMR) systems [15]. An fully automatic system with consistently high accuracy is not realistic in practice [18]. Therefore, it is difficult to directly apply any previous OMR softwares to this challenging egocentric problem. Moreover, the inconstant view point angles make egocentric images much more distorted compared with printed music pieces (the default input for most OMR softwares), placing even more difficulty to the problem. In order to bypass these difficulties, we propose a novel framework that uses the MIDI data as a guidance of recognition. MIDI is an easily accessible music symbolic format, and also rather easy to parse. There have already been numerous audio-to-score alignment applications [6, 13, 17], which mainly focused on matching MIDI with audio. Different from these applications, our system applies a graphical model to incorporate MIDI into OMR system and focuses on egocentric recognition.

Fig. 1 shows a sample use case of our system, where the

human subject sits almost still in front of a piano. This allows to simplify the problem from processing entire video to individual frames. In addition, music scores are highly structured according to their symbol-level semantics. The temporal information contained by MIDI data implies the exact spatial order of notes appearing on the image. Moreover, we can explore interesting relationships between the Inter-Onset Intervals (IOI) of adjacent notes and their spatial distances in notation. Given the above observations, we feed the MIDI data into the recognition process and constrain the search space, such that the outputs are musically meaningful. Once the structure is determined, the corresponding score can be represented by a connected graph and the problem can be formulated as graphical model inference.

Yi *et al*. [23] proposed an interesting egocentric Optical Character Recognition (OCR) framework to assist blind persons. They applied an Adaboost model for text region localization and then used off-the-shelf OCR engines to perform the recognition. Analogously, we also propose a pipeline for egocentric OMR system. More specifically, we decompose our system into three steps. In the first step, we localize the score region based on foreground-background segmentation. In the second step, we propose to automatically discover the staff lines using Random Sample Consensus (RANSAC). In the third step, we use a linear chain Conditional Random Field (CRF) to model the note sequence and search for the optimal sequence that best aligns with the observation by incorporating MIDI information.

Summary of contributions. Our contribution in this paper is three fold. Firstly, we are the first to propose the problem of egocentric optical music recognition, which has important applications for education and entertainment purposes. Secondly, we propose a novel MIDI-assisted egocentric OMR system that recognizes music symbols, and aligns them with the structured MIDI data using a CRF model. Lastly, we collect the first egocentric OMR dataset using a Google Glass, and perform systematic benchmark experiments. We show that our approach is accurate compared to several baseline methods.

2. Related Work

Image segmentation. Segmentation plays an important role in many computer vision systems by serving as preprocessing step. Ren *et al*. [19] proposed a bottom-up approach for figure-background separation, jointly using motion, location and appearance cues. Fathi *et al*. [7] segmented the foreground and background at super-pixel level, and model the temporal and spatial connections with a MRF. Serra *et al*. [20] combined hand segmentation and activity recognition to achieve higher accuracy. The objective of our paper, however, is different with segmenting such foregrounds (e.g. human hands or natural objects). Our goal is to separate the document out of a natural scene. Some primitive methods has been proposed in [11], but it's not directly applicable to the much more complex egocentric environment. In our experiment, we make use of the shape prior of the music scores and a probabilistic color model to identify the foreground region.

Staff line detection. Staff detection or removal is always one of the key steps in OMR. The performance of the pitch recognition is highly dependent on the staff detection accuracy. Therefore, in order to assign the location of notes to their correct pitch index, we need to find staff lines at first. Cardoso *et al*. [5] modeled staff finding problem as a global search of stable path, which is not a computationally cheap design. Fujinaga *et al*. [10] uses projection-based approach to remove staff lines and keep the most of music symbols. Our task is more challenging in that the staves don't share the same angles due to the multidimensional page distortions. Further, the observation is much more blurry than printed version, and we have higher efficiency request than offline systems. To overcome all these new difficulties, we choose to apply a bottom-up approach to propose and select plausible staff-line models. The popular RANSAC [9] framework proved success in various real-time systems [1, 2]. Our method is inspired by these sampling-based methods.

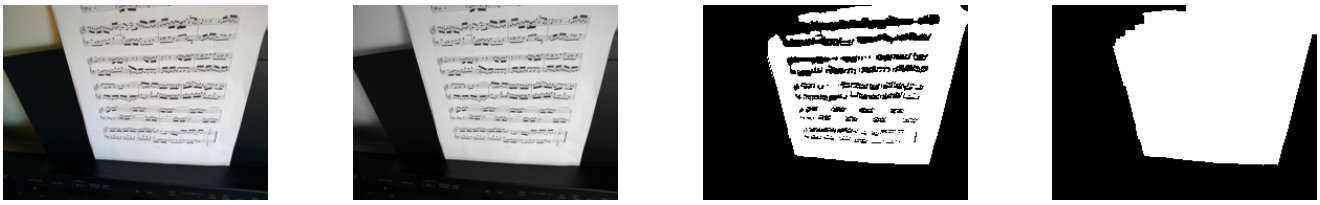
Optical music recognition (OMR). There have been a lot of progress of OMR studies but the current state-of-the-art still leave many open questions [2, 4]. These offline systems heavily rely on human labors for error corrections, and thus it's impractical to apply them directly in egocentric scenarios. The traditional OMR takes on the responsibility to identify symbols from scratch, without any assistive information. This proved to be a challenging problem since even if all the musical symbols have been correctly identified, the higher-level interpretation is still non-trivial [12]. Our approach, on the contrary, embeds useful music information of MIDI to the deepest heart of the system, and use it to direct the whole recognition process.

In the following sections, we will explain the technical details. We first describe our approach for localizing the sheet music in the captured image in Section 3.1, and then discuss our staff line detection algorithm in Section 3.2. We then introduce our inference algorithm for aligning music symbols and MIDI data in Section 3.3. Experimental studies are explained in Section 4.

3. Approach

3.1. Sheet Music Localization

Modeling the Sheet Music Region. The score region has a strong shape prior due to the viewpoint of the observer and the rectangular boundaries of the original score documents. We treat the sheet music localization as a parameter-



(a) (b) (c) (d)

Figure 2: Proposing candidate score region: (a) color image down-sampled to $1/10$ its original size; (b) converted to grayscale; (c) thresholding and binarization; (d) morphological smoothing and hole-filling.

ized boundary identification problem, which can be formulated as the optimization of these boundary parameters.

$$\Theta^* = \arg\max_{\Theta} \sum_{(i,j)\in R_\Theta} D(p(i,j)) \tag{1}$$

$D(p(i,j))$ is the data term for pixel $p(i,j)$. The region parameter for region R_Θ, $\Theta = \{\Theta_l, \Theta_r, \Theta_t, \Theta_b, \Theta_I\}$, contains five components respectively representing the left, right, top, bottom boundaries and the support of image. Θ_I is one scalar parameter; each of the rest contains two variables: the angle and intercept: $\Theta_{l,r,t,b} = (\theta_{l,r,t,b}, int_{l,r,t,b})$. The inference was performed in the parameter space S_Θ, constrained by the shape prior (reflected in angles) and the minimum width/height of the foreground region. The image was down sampled in this step for sake of computational efficiency.

Data Likelihood. We learn the data model in Eqn. 1 in an unsupervised way, which adapts to different illumination conditions. We first convert the down-sampled RGB image to grayscale and apply a threshold to obtain pixels with high intensities. We smooth these *seed* regions and learn the probabilistic representation for the foreground with r,g,b components of the colored version inside this smoothed candidate region using Gaussian Mixture Models (GMM): $G = \sum_{1\le i\le N} \alpha_i Norm(m_i, \sigma_i)$. We learn the background GMM model analogously outside the smoothed candidate region. The smoothing process is illustrated in Figure 2.

Note that N is the number of the components in the model, m and σ are the mean and standard deviation for each component. We set $N = 3$ for both G_{fg} and G_{bg}, and learnt the parameters via several iterations of standard Expectation-Maximizaiton(EM) process. We use the log ratio of these two distributions to represent the data likelihood (Eqn. 2):

$$D(p(i,j)) = \log \frac{G_{fg}(p(i,j))}{G_{bg}(p(i,j))} \tag{2}$$

Figure 3 shows us the *foreground heat map* generated from the proposed data model. The higher the value is, the more possible it belongs to the score region. The inference will then be performed over this *heat map*.

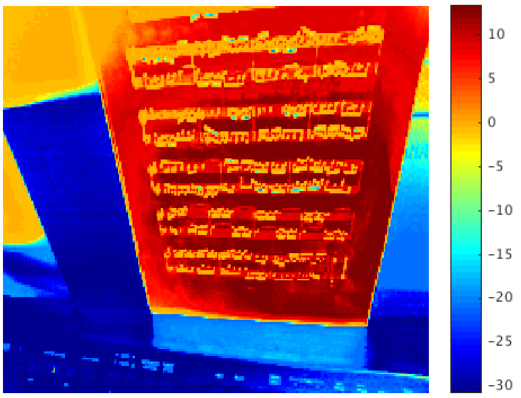


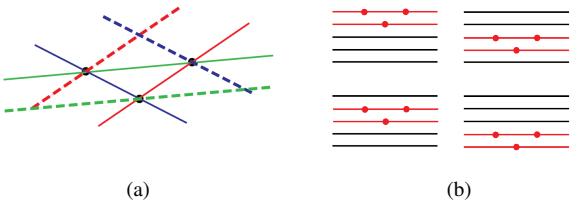
Figure 3: Foreground heat map for score region localization.

3.2. Staff Detection

Staff lines in egocentric scores are oftentimes *skewed*. More importantly, they appear with very different angles. A top-down model for staff detection on the whole page requests excessive computation, so we resort to a more efficient bottom-up RANSAC approach. The algorithm proposes plausible local models and evaluate them by global votes.

We model the staff as groups of five parallel lines. The model is composed of a parameter tuple (α, β, Δ), where α and β represents the slope and intercept of the first staff line, and Δ is the gap between two adjacent lines. We propose a constant number of local models based on a group of three sampled pixels from the binarized score region. We call one such sampled group as a pixel triplet; each triplet proposes $3 \times 4 = 12$ local models (see Figure 4).

We prune the least voted hypothesized models and only keep those satisfying two different criteria through non-



(a) (b)

Figure 4: Staff model proposal: (a) three possible directions of adjacent two staves based on the sampled triplet; (b) four possible locations of adjacent two staves on the complete staff.

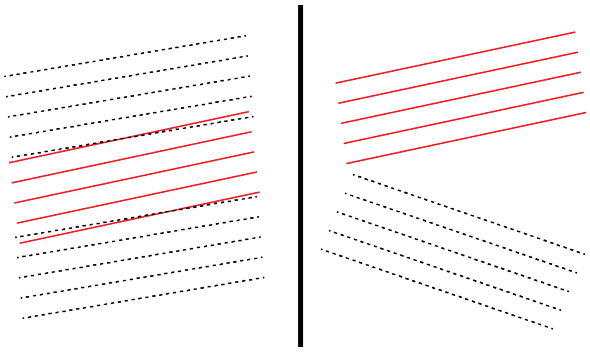


Figure 5: Non-Maxima-Suppression for staff identification with two different constraints. Left: non-overlapping constraint; Right: neighborhood slope similarity constraint. Solid Red: local optimal model; Dashed Black: eliminated models which violates these two constraints.

maximum suppression (Figure 5): *neighborhood slope similarity* (the neighbor staves should have close slopes) and *non-overlapping* (staves should not conflict with each other) constraints. The thinned staff models were accepted as the final interpretation of the whole-page staff structure.

3.3. Music Recognition

We model egocentric optical music recognition as a note sequence alignment problem between the egocentric observation and MIDI data. We focus on the note head symbol as the important anchor for this alignment task considering the unique correlation between note events in MIDI and their corresponding note heads on the image. There are occasionally exceptions breaking this bijective MIDI-to-Notehead mapping, such as in trills, grace notes, and tied notes, or due to different notational conventions, but it doesn't undermine the ground of selecting note head as the alignment anchor rather than any other symbols like stem, beam, rest, flag, etc. since the others carry much more variance across different notations.

We extract three important music attributes for each music event from MIDI data: *onset*, *pitch*, and *end of measure*. Table 1 shows the details of the extracted 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

Table 1: Sequence of note events parsed from MIDI (Bach Invention in C major (No. 1), the 1st measure).

Given image data X and the locations of a certain staff line l, we want to estimate the optimal measures aligned to the current staff. Let $\mathcal{S}$ represent the state space over which we search for the optimal alignment. State s is composed of (n, x, y, a), the note event n extracted from MIDI, the location (x, y) of its note head on the page and its latent music attribute a. n contains the pitch and onset of the note, and a is a variable taking the implicit music information that is not directly contained by symbolic data. In our experiment setting, we specifically infer the *clef* associated with the current note to unveil the missing semantics.

The inference problem thus can be formulated as:

$$S^* = \arg\max_{\{s_i\}} E(s_i|X,l) + E(s_i, s_{i+1}|X,l) \tag{3}$$

$$= \arg\max_{\{s_i\}} E(n_i, x_i, y_i, a_i|X,l) + E(s_i, s_{i+1}|X,l) \tag{4}$$

Once we have the note's information, staff locations and its associated clef, the note's vertical position becomes a deterministic function of its horizontal coordinate:

$$y = f(x|n, a, l) \tag{5}$$

The pairwise term in Eqn. 4 serves as a hard spatial constraint. It penalizes the impossibly small distance between adjacent notes if they have large Inter Onset Interval (IOI). We use a small quantization value as the IOI threshold (ϵ), and a predefined number of space units (staff gap σ) as the minimum note distance. This constraint sets reasonable minimum distance for ordinary note pairs while allowing for occasional violations caused by small notes like trills or grace notes.

$$\begin{aligned} &E(s_i, s_{i+1}|X,l) \\ &= E(\|x_i - x_{i+1}\||X,l,n_i,n_{i+1}) \\ &= E(\Delta_{i,i+1}|X,l,n_i,n_{i+1}) \\ &= \begin{cases} -\inf, & \Delta_{i,i+1} < C \cdot \sigma, IOI_{i,i+1} > \epsilon \\ 0, & otherwise \end{cases} \end{aligned}$$

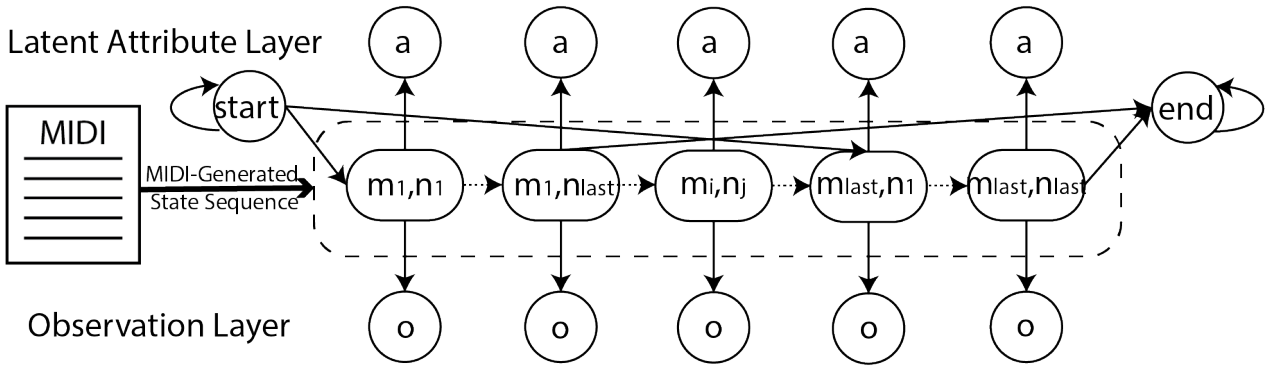


Figure 6: Graphical model for MIDI-assisted Optical Music Recognition. m_i, n_j denotes the j-th note event of measure i. We omitted the state transitions to white space for a more straightforward illustration.

We assume all the note events have the same prior probability. Now that the pairwise energy does not correlate to the scale of unary's, the unary term can be rewritten as $E(x_i, a_i|n_i, X, l)$.

We train our unary model via linear Support Vector Machine (SVM) and use Histogram of Gradients (HOG) as the image feature. We extracted HOG features for both positive and negative training data and fed these features into the SVM classifier. A small validation dataset is used during the training stage in order to tune SVM parameters. We use the trained model to detect note heads on the test images.

Figure 6 illustrates the graphical model for MIDI-assisted OMR. We parse MIDI messages into a sequence of hidden states in our CRF model, and use this generated graph to infer the optimal MIDI subsequence and align the notes to image observations. For each subsequence hypothesis the inference will estimate $\{n_i\}$, x, y and a simultaneously. As shown in Figure 6, the hidden layer is a Markov chain connecting all the notes in the MIDI sequence, the latent attribute layer takes the clef associated with each note, while the observation layer corresponds to the image data. Once we perform the whole inference via a Viterbi decoder on the target staff, we will locate the optimal measure subsequence and determine the optimal parameters of its containing notes at the same time. .

4. Experiment

We initialized a dataset with the first 5 pieces of **Bach's 15 Inventions** (No. 1 - 5). The dataset contains 54 egocentric images in total, each including 8 to 12 staves. The data was acquired from the online music score repository **IMSLP**[1]. We annotate the staff endpoints and note positions on each image, and manually align the notes to MIDI events as the ground truth.

[1] http://imslp.org/wiki/15_Inventions,_BWV_772-786_(Bach,_Johann_Sebastian)

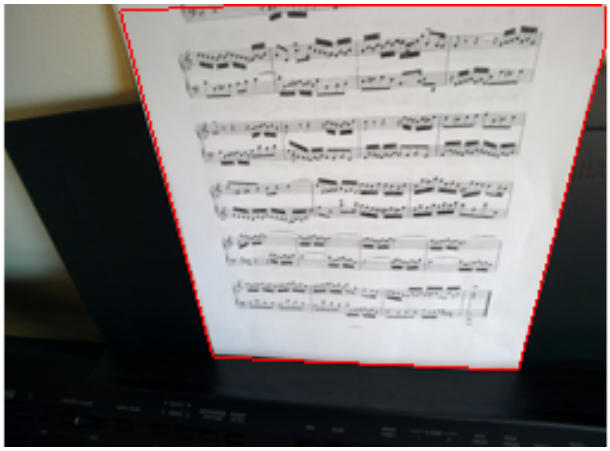


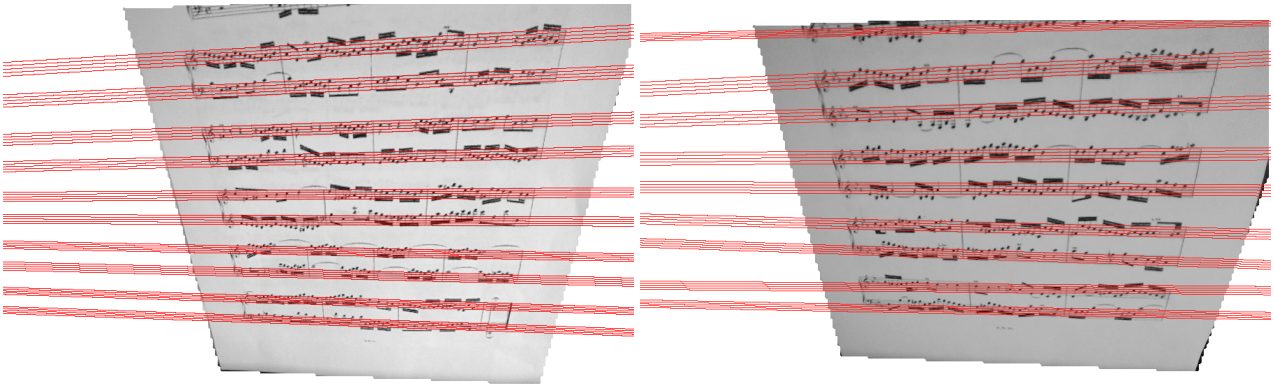
Figure 7: Bach Invention in C major (No. 1): Score region extracted by using the segmentation approach mentioned in Section 3.1.

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

Table 2: Precision, Recall and F-score of staff detection.

Our test set contains 242 independent staves. We evaluate our staff detection accuracy using the mean squared error between the endpoint coordinates of ground truth and estimated staves. We claim a staff is correctly identified if this error is below a small threshold. Table 2 presents the evaluation results for staff detection. Figure 7, Figure 8 and Figure 9 respectively highlights the located score region and detected staves on **Bach Inventions No. 1 - 4**, where all the staves were identified. We have detected 198 true positive staves in total. We will work on these correctly identified staves for later evaluations.

We evaluate note detection and MIDI alignment accu-



(a) Staff detection on Bach Invention No. 1

(b) Staff detection on Bach Invention No. 2

Figure 8: Detected staves on Bach Inventions No. 1- 2. Background was removed after score region localization.

racy against two other baselines. The first baseline uses a greedy approach to align subsequence notes to the observation. The greedy algorithm also outputs the highest scored subsequence but adds all the detected note's likelihood to the hypothesized subsequence score as long as they don't overlap with each other. This approach ignores both the order and distance constraints of notes. The second one uses the same CRF model but takes off the pairwise distance constraints. In contrast, our approach maintains both the spatial order and constraints.

Figure 10 shows us the experimental results generated by our CRF model. Mapping MIDI events to note heads occasionally causes problems. For instance, there will be multiple note heads detected for a single trilled note since trill is represented by several short notes in MIDI. Also, only one of the tied notes will be recognized since they're merged into one single MIDI event.

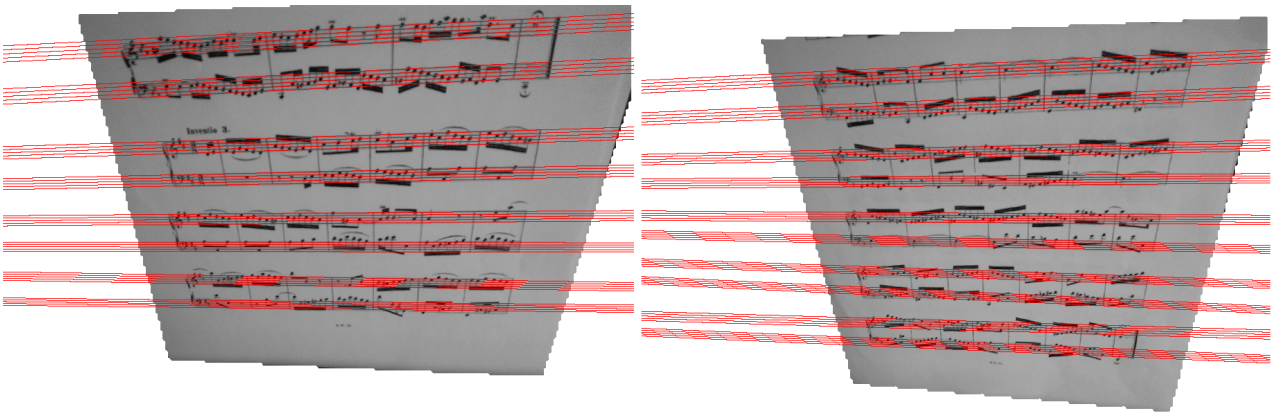
We define two accuracy measurements to evaluate the effectiveness of different approaches. Note detection accuracy measures the portion of detected notes matching the annotated notes at the same locations in the ground truth, while the MIDI alignment metric examines in addition whether the matched notes have the same pitches. We evaluate the accuracy of identified measure subsequence first and based on these matched subsequences we perform note detection and MIDI alignment evaluation. From Table 3 we see that our approach achieved highest accuracy for both subsequence matching and MIDI alignment. Greedy approach tends to detect as many objects as possible, but lost the musical structure otherwise maintained in the CRF model. This explains why there is a significant accuracy decline from its note detection to MIDI alignment. The two CRF models have comparable F-scores; both are significantly higher than that of greedy algorithm. This accuracy improvement is gained by incorporating note sequence structures into the recognition. The note detection rate of CRF without pairwise constraint is slightly higher than the pairwise-constrained CRF, while the constrained one outperforms the other two in the final MIDI alignment evaluation.

5. Conclusion

We presented a optical music recognition approach for egocentric device. Our main idea is to incorporate offline symbolic data into a single joint OMR framework. We extract useful structural information of music symbols from MIDI data to assist the egocentric music score recognition. The proposed approach is shown to outperform several baselines in terms of recognition accuracy.

Our approach provides possibilities to interesting applications that combines music and egocentric vision. After the recognition is performed, the locations for staves, measures and notes will be estimated. The most straightforward application includes playing back the measures of interest to the user or rendering pitches and rhythms on the screen to assist user's score-reading. Other interactive games can be devised by using the data generated from the inference.

One limitation of the proposed approach is that the current system can hardly achieve real-time request since it keeps searching over the complete MIDI data for each estimated staff. We need to design heuristics to prune out impossible measures to improve the processing speed. Another solution is to put the human users into the loop, which will provide additional information to allow real-time computation. It is also desirable to extend the algorithm to process continuous video stream so that we can track the staves and note heads more smoothly and accurately. We leave these interesting challenges as future work.



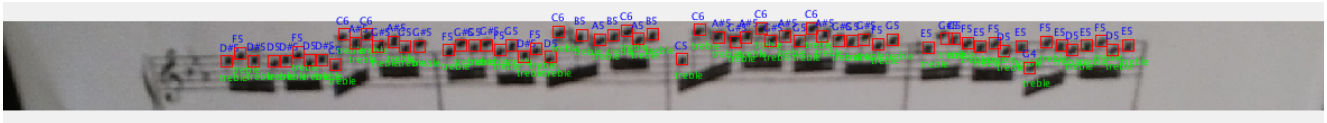
(a) Staff detection on Bach Invention No. 3

(b) Staff detection on Bach Invention in No. 4

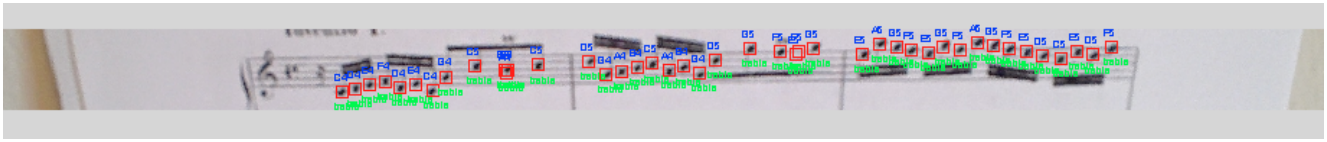
Figure 9: Detected staves on Bach Inventions No. 3 - 4. Background was removed after score region localization.

		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%**

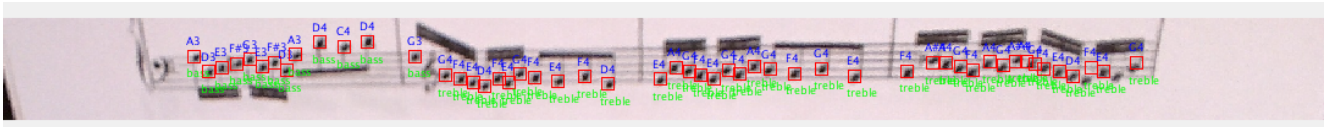
Table 3: Evaluation on the measure subsequence, note detection and MIDI alignment accuracy for (1) greedy algorithm, (2) CRF without pairwise constraint, (3) proposed model.



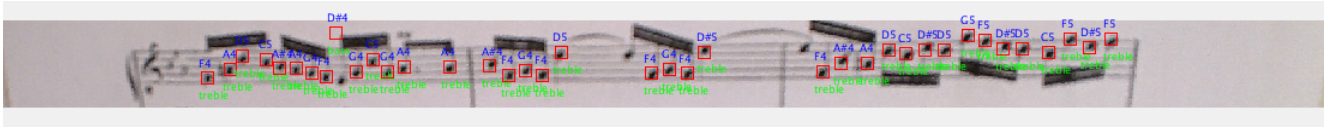
(a) All the notes were correctly identified on Bach Invention No. 5, the 7th staff.



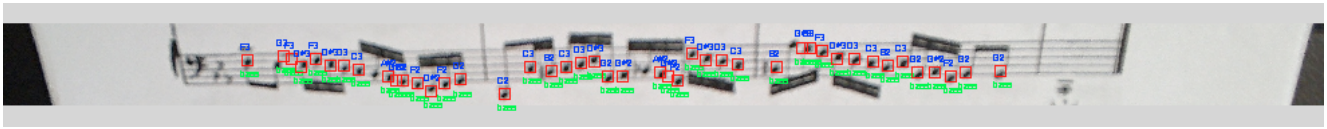
(b) All the notes were correctly identified on Bach Invention No. 1, the 1st staff. Extra notes were detected due to trills.



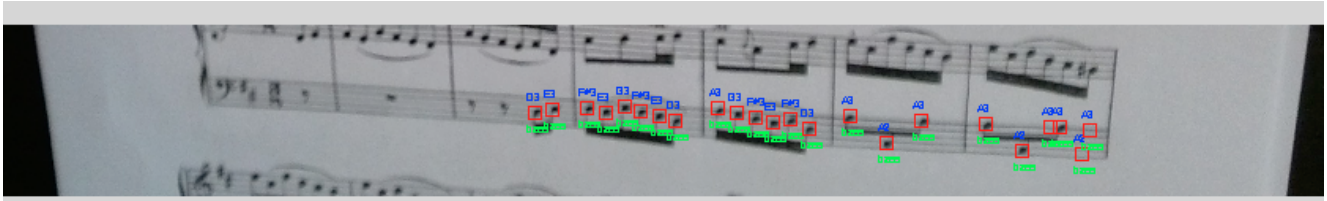
(c) Clef change was correctly identified on Bach Invention No. 1, the 6th staff.



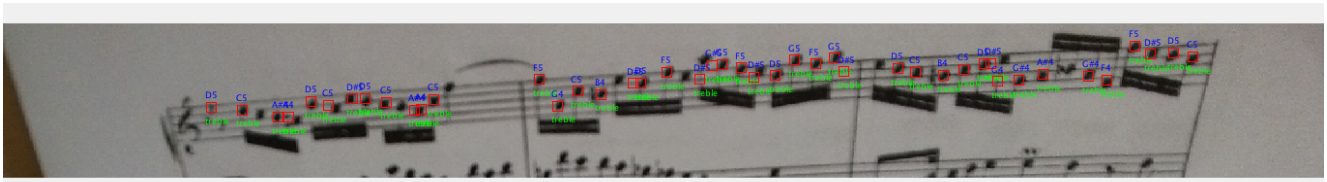
(d) Example of low-level detection error on Bach Invention No. 2, the 13rd staff.



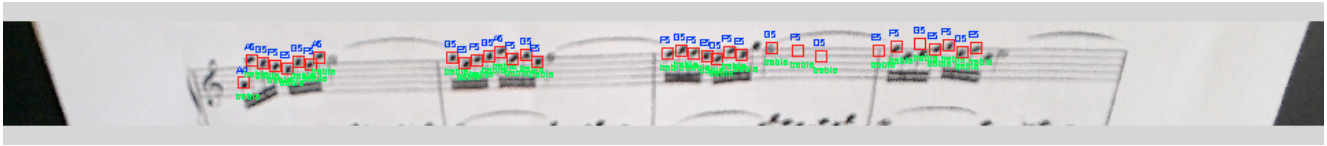
(e) Example of low-level detection error on Bach Invention No. 2, the 18th staff.



(f) Example of low-level detection error on Bach Invention No. 3, the 2nd staff. An extra measure was detected at the end.



(g) Example of high-level detection error on Bach Invention No. 2, the 15th staff.



(h) Example of high-level detection error on Bach Invention No. 1, the 9th staff. The last measure was mis-aligned.

Figure 10: MIDI alignment results. Red: note locations; Blue: pitch names; Green: associated clef.

References

[1] M. Aly. Real time detection of lane markers in urban streets. In *Intelligent Vehicles Symposium, 2008 IEEE*, pages 7–12. IEEE, 2008.

[2] J.-C. Bazin and M. Pollefeys. 3-line ransac for orthogonal vanishing point detection. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 4282–4287. IEEE, 2012.

[3] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg. The evolution of first person vision methods: A survey. 2015.

[4] D. Byrd and J. G. Simonsen. Towards a standard testbed for optical music recognition: Definitions, metrics, and page images. *Journal of New Music Research*, 2015.

[5] J. D. S. Cardoso, A. Capela, A. Rebelo, C. Guedes, and J. P. d. Costa. Staff detection with stable paths. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6):1134–1139, 2009.

[6] R. B. Dannenberg and N. Hu. Polyphonic audio matching for score following and intelligent audio editors. *Computer Science Department*, page 507, 2003.

[7] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012*, pages 314–327. Springer, 2012.

[8] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On*, pages 3281–3288. IEEE, 2011.

[9] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[10] I. Fujinaga. Staff detection and removal. *Visual perception of music notation: on-line and off-line recognition*, pages 1–39, 2004.

[11] U. Garain, T. Paquet, and L. Heutte. On foreground-background separation in low quality color document images. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, pages 585–589. IEEE, 2005.

[12] R. Jin and C. Raphael. Interpreting rhythm in optical music recognition. In *ISMIR*, pages 151–156. Citeseer, 2012.

[13] C. Joder, S. Essid, and G. Richard. An improved hierarchical approach for music-to-symbolic score alignment. In *ISMIR*, pages 39–45. Citeseer, 2010.

[14] S.-R. Lee, S. Bambach, D. J. Crandall, J. M. Franchak, and C. Yu. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 557–564. IEEE, 2014.

[15] V. Padilla, A. Marsden, A. McLean, and K. Ng. Improving omr for digital music libraries with multiple recognisers and multiple sources. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, pages 1–8. ACM, 2014.

[16] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora. Compact cnn for indexing egocentric videos. *arXiv preprint arXiv:1504.07469*, 2015.

[17] C. Raphael. Aligning music audio with symbolic scores using a hybrid graphical model. *Machine learning*, 65(2):389–409, 2006.

[18] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. Marcal, C. Guedes, and J. S. Cardoso. Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval*, 1(3):173–190, 2012.

[19] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3137–3144. IEEE, 2010.

[20] G. Serra, M. Camurri, L. Baraldi, M. Benedetti, and R. Cucchiara. Hand segmentation for gesture recognition in egovision. In *Proceedings of the 3rd ACM international workshop on Interactive multimedia on mobile & portable devices*, pages 31–36. ACM, 2013.

[21] E. H. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference On*, pages 17–24. IEEE, 2009.

[22] L. Xia, I. Gori, J. Aggarwal, and M. Ryoo. Robot-centric activity recognition from first-person rgb-d videos. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 357–364. IEEE, 2015.

[23] C. Yi and Y. Tian. Assistive text reading from complex background for blind persons. In *Camera-Based Document Analysis and Recognition*, pages 15–28. Springer, 2012.