Adaptive Score-Following System by Integrating Gaze Information

Kaede Noto
Graduate School of
Future University Hakodate
g2118030@fun.ac.jp

Yoshinari Takegawa
Graduate School of
Future University Hakodate
yoshi@fun.ac.jp

Keiji Hirata
Graduate School of
Future University Hakodate
hirata@fun.ac.jp

ABSTRACT

In actual piano practice, people of different skill levels exhibit different behaviors, for instance leaping forward or to an upper staff, mis-keying, repeating, and so on. However, many of the conventional score following systems hardly adapt such accidental behaviors depending on individual skill level, because conventional systems usually learn the frequent or general behaviors. We develop a sc ore-following system that can adapt a user's individuality by combining keying information with gaze, because it is well-known that the gaze is a highly reliable means of expressing a performer's thinking. Since it is difficult to collect a large amount of piano performance data reflecting individuality, we employ the framework of the Bayesian inference to adapt individuality. That is, to estimate the user's current position in piano performance, keying and gaze information are integrated into a single Bayesian inference by Gaussian mixture model (GMM). Here, we assume both the keying and gaze information conform to normal distributions. Experimental results show that, taking into account the gaze information, our score-following system can properly cope with repetition and leaping to an upper row of a staff, in particular.

1. INTRODUCTION

The goal of our study is to build a score-following system that adapts a user's individuality. Score-following is one of the important topics in MIR and is a fundamental technique used in many applications including automatic accompaniment, estimation of current position from audio [4, 5, 19], and estimation from symbols [2, 3, 16, 17]. In reality, a score-following system may often face problematic situations, in which current position leaps forward or backward freely. Such leaps occur because of repetition and wrong keying, and during practice. Therefore, researchers of score-following attempt to propose systems and/or algorithms for reacting to or tracking a current position which includes occasional leaping forward or backward [10–12, 18]. For a score without repeated and/or iterated phrases, conventional systems and algorithms can estimate current position almost correctly.

Copyright: ©2019 Kaede Noto et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Gaze can give us significant information for score-foll owing, when, for instance, starting performance from an arbitrary position [13]. Our system can use gaze information to estimate current position correctly to some extent, even when keying information is unavailable. There is, however, a crucial issue to be considered which is called eye-hand-span (EHS). EHS means the distance between the point on the score at which a player looks to obtain, in advance, information of notes to be played, and the actual current keying position [14]. Usually, during performance, a pianist looks at a point on the score approximately a phrase or a few notes ahead of current keying position. Since the length of EHS depends on, for instance, individuality, the structure of the melody, the degree of proficiency, and tempo, our previous system takes into account the average length of EHS obtained from experimental data and estimates the current position from both gaze and keying information multiplied by fixed weights. Thus, the system unfortunately neither conducts individual EHS estimation nor assigns the optimum weights to both gaze and keying information for each pianist.

This paper proposes a score-following system, which a dapts a user's individuality in piano practice. Due to the difficulty of collecting a large amount of individual's performance data to learn, we adopt the Bayesian inference, which has the advantage that it can be used even if only a small amount of learning data is available. Thus, we propose a method for treating gaze (EHS) probabilistically and integrating gaze and keying information within the B ayesian inference framework. First, we assume the length of EHS follows the normal distribution and that the distribution is updated dynamically by observed data. We define the distribution of keying information in the same way. Next, we integrate gaze and keying information, using Gaussian Mixture Model, to be used as the likelihood function in the Bayesian inference. Advantages of the m ethod include improvement of the accuracy of estimating current keying position and the ease of adding other new features which reflect the user's individuality and/or thinking.

2. RELATED WORK

2.1 Gaze Information and Individuality

A performer perceives music while forming 'chunks', wh ich are units to recognize a sequence of pitch events as a pattern. The size of EHS is related to the size of a chunk. Weaver revealed that professional performers perceive the

notes on a score as horizontal and vertical groupings [20]. Furneaux et al. conducted experiments to identify the differences in EHS between professional and amateur pianists [7]. In this research, EHS was defined as the number of notes between the note(s) being played and the note(s) upon which the player's gaze was fixed (performance point and gaze point). The professional's EHS (approx. four notes) was larger than the amateur's (approx. two notes). From the point of view of melody, Kobori and Takahashi compared the eye movements when professionals and amateurs play a melody on piano and guitar [9]. This research suggested that the individuality of performers, the difficulty of music pieces and the performer's knowledge on music pieces were the crucial factors which influenced eye movement.

2.2 Automatic Score Following

A central issue in score following is correctly estimating performance position in response to a variety of uncertain events such as leaping forward, mis-keying and repeating. Nakamura et al. developed a score-following system, Eurydice, that estimated the user's current position using improved HMM and Viterbi algorithm [17]. Since the weight was calculated based on distance, Eurydice tended to estimate a position near the previous current position. To some extent, Eurydice achieved accurate estimation in performance containing unexpected movements such as mistakes, repeats, and skips. However, since Eurydice used only keying information, it was difficult to identify the phrase being played in the case of a melody containing many repeated phrases.

Terasaki et al. proposed a score-following system that was hardly affected by unexpected movements [13]. The system introduced gaze likelihood to the cost of DP matching. Gaze likelihood was obtained by HMM, which was employed for predicting gaze and removing noise from the raw data of eye movement. The system could correctly estimate current position, even when a player started at a point different from the previous point at which he/she had stopped playing, or a beginning point of a repeated phrase. Terasaki et al. evaluated the estimation error rate for musical scores including repeated phrases and found that the correct answer rate was improved by 1.2 times (from 72.7% to 85.2%). However, the system unfortunately could not cope with the individuality of EHS.

Grubb et al. introduced the Bayesian inference into an automatic accompaniment system [6]. They defined the probability distribution with respect to note number i as a random variable, which was updated every time data was observed. Parameter d stands for the estimated distance from a pre-estimated position, v the observation value, and j the performer's position at the previous timing. The estimated position was updated by the most recent observation. First, based on the previous position and estimated distance, the current position is estimated, as follows.

$$f_{I|D}(i|d) = \int_{j=0}^{\|Score\|} f_{I-J|D}(i-j|d) \cdot f_{\text{Source}}(j)\partial j$$

Next, this estimated value is further updated to take into account the observation using the Bayesian inference, as follows.

$$f_{I|D,V}(i|d,v) \propto f_{V|I}(v|i) \cdot f_{I|D}(i|d)$$

$f_{V|I}(v|i)$ is made from observation, and is considered as a likelihood. In an experiment on recorded data and real-time data, a latency of 159 ms occurred on average. This value falls below the limit of latency that humans can perceive, which is 10$\sim$ 100 ms [5]. Thus, Grubb's system did not interrupt piano performance, but users reported that they felt some discomfort.

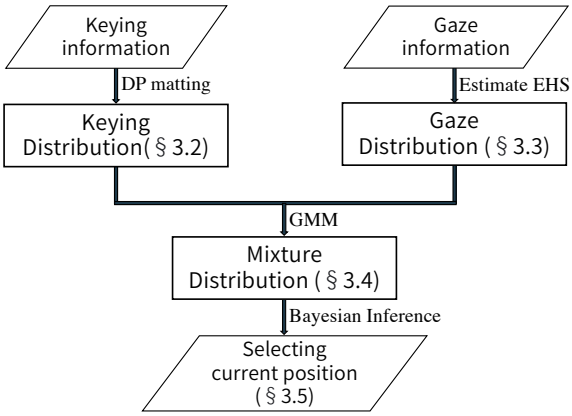

Figure 1. System Configuration

3. A SCORE-FOLLOWING SYSTEM USING KEYING AND GAZE INFORMATION

This section describes the score-following model which deals with gaze, and shows how the input data are converted to distributions and how they are combined.

3.1 Combining Gaze information with Keying Information

Fig.1 shows the processing flow of the proposed system. The model takes two input data at the same time: keying information and gaze information. The keying information as MIDI numbers and the gaze information measured by an eye-tracking device are entered into the system. The system fits the input data to normal distributions, and integrates the distributions using GMM. After integrating the distributions, by Bayesian inference, the system estimates a note number on a score as current position ($i \in \{1, 2, 3, \cdots I\}$).

3.2 Distribution of Keying

To formalize the keying information as a normal distribution, we need to define the average and the variance. Firstly, to obtain the average, we use DP matching. DP matching finds the degree of similarity between notes on a score and notes being played, called the best match, and chooses the score position having the best match as a performance position [12]. However, it is difficult for DP

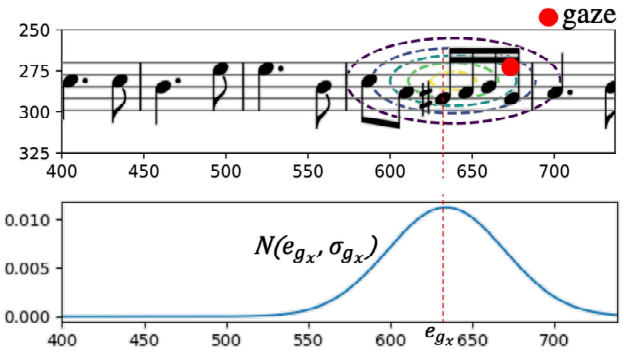

Figure 2. Create the distribution of gaze, with note estimated from EHS as μ

matching to follow a performance which includes backward leaps and repeats. As a simple solution, we employ an exhaustive search of all the played notes over an entire score, as in our previous research [13]. Then, we can identify the note number $e_{DP} \in \{1, 2, 3, \cdots I\}$ that is most probably the average value. Next, we define a value corresponding to the variance. For all notes, we calculate the average distance $\sum |x_i - x_{i+1}|/I$ between every adjacent note in the horizontal direction. Let us regard the value obtained as the standard deviation of the distribution of keying. Then, we define the keying distribution as follows:

$$p_{DP}(i) \sim N(\mu_{DP}, \sigma_{DP})$$

where μ_{DP} is equal to e_{DP} and σ_{DP} is $(\sum |x_i - x_{i+1}|/I)^2$.

3.3 Distribution of Gaze

To introduce the gaze information, we define the distribution of gaze considering eye-hand-span (EHS) in Fig.2.

EHS consists of the horizontal and vertical spans $x_i - g_x$ and $y_i - g_y$, where (x_i, y_i) means the note number of the key being played and (g_x, g_y) the gaze point on a display (where, on the display, the user is looking). Here, concerning the size of EHS, we assume it follows the normal distribution:

$$p_x(x_i - g_x) \sim \mathcal{N}(x_i - g_x | \mu_{g_x}, \sigma_{g_x}) \quad (1)$$

$$p_y(i_y - g_y) \sim \mathcal{N}(i_y - g_y | \mu_{g_y}, \sigma_{g_y}) \quad (2)$$

Terms $x_i - g_x$ and $x_i - g_y$ are the lengths of EHS in the horizontal and vertical directions, respectively. Variables μ_g and σ_g represent the average length of EHS and the variance, respectively. Since the Gauss-gamma distribution, as prior distribution, is used in the Bayesian inference, μ_g and σ_g can be analytically calculated.

According to the length of EHS learned, we estimate the user's current position from gaze point (g_x, g_y) and assign g_x and g_y to Equations(1) and (2). Since, at this moment, more than one candidate note is obtained, to determine the user's current position, we choose a note by calculating the likelihood of each note, considering EHS, as follows:

$$e_g = \arg\max_{i \in I} [p_x(i) p_y(i)]$$

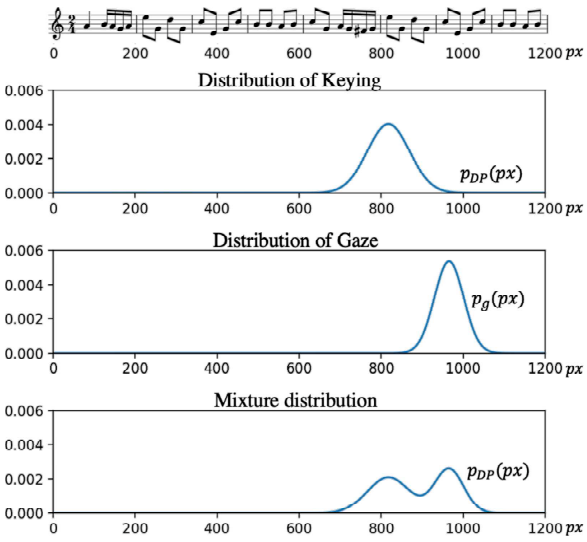

Figure 3. Integration of distributions by GMM

Then, using the above variables, we define the gaze distribution, $p_g(i)$, with e_g as the average and σ_{g_x} as the variance (Fig.2).

$$p_g(i) \sim N(i | e_{g_x}, \sigma_{g_x})$$

3.4 Integration of Multiple Information

We use GMM to integrate the keying and gaze distributions (Fig.3). To use GMM, we convert the random variable of keying distribution into a coordinate value on the x axis on a score (in units of pixels). The integrated probability distribution by GMM is given by Equation(3) with mixture ratio π_k :

$$P_{GMM}(i) = \sum_{k=1}^{2} \pi_k N(i | \mu_k, \sigma_k) \quad (3)$$

$$\sum_{k=1}^{K} \pi_k = 1$$

where μ_k = $[e_{DP}, e_g]$ and σ_k = $[\sigma_{DP}, \sigma_{gx}]$.

Since we need to decide the significance of each type of information, we use the EM algorithm to update mixture ratio π_k.

$$r(Z_{nk}) = \frac{\pi_k N(x | \mu_k, \sigma_k)}{\sum_{i=1}^{2} \pi_i N(x | \mu_i, \sigma_i)}$$

$$\pi_k = \frac{\sum_{n=1}^{N} r(Z_{nk})}{N}$$

Here, x represents an observation, which is assumed to have been sampled from the normal distribution of average μ_i and variance of σ_i. To estimate π_k, we calculate the $r(Z_{nk})$ which represents the ratio of the distribution of each k $(k = 1, 2)$ at the values of the density function of the mixed distribution at pixel x. Since we cannot know what a player is thinking, in principle, we cannot know the true point at which he/she is playing.

To obtain as accurate a value of π_k as possible, we instructed subjects to play notes in the order indicated by a score, that is, linearly from the beginning to the end.

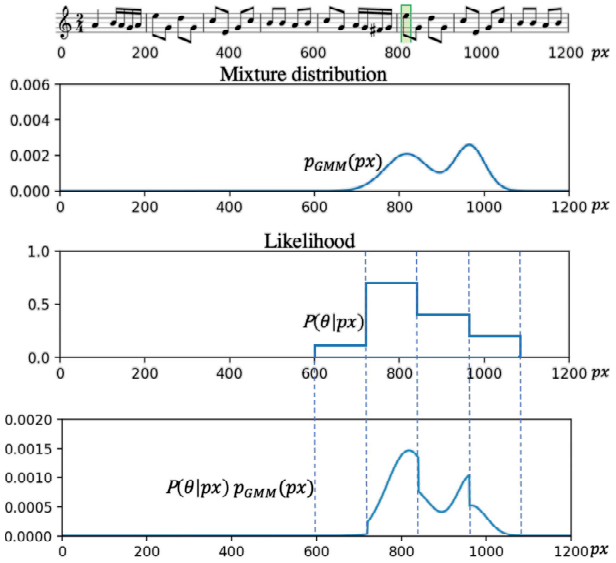

Figure 4. Estimating Note by Multiplication of Likelihood and Mixture Distribution

Mixture ratio π_k represents the maximum likelihood of the mixing ratio. In E step, we calculate the expected value of P_{GMM} when the mixture ratio is π_k. In M step, π_k is optimized by the data obtained by sampling for each note from a normal distribution that assumes $\mu = i_x$, $\sigma_g = (\sum |x_i - x_{i+1}|/I)^2$ by the maximum likelihood estimation. From the obtained mixture ratio, we define the mixed distribution that is the weighted summation of the keying and gaze information (Fig.3).

3.5 Estimating Keying Position based on Bayesian Inference

The current position is estimated from the mixture distribution and likelihood. The relationship between a note and the variables can be represented by the occurrence probability of the i-th note, $P(i)$, and the value distribution of variables, $P(\theta)$. To estimate $P(i|\theta)$ using Bayesian estimation, it is necessary to know $P(i)$ and $P(\theta|i)$ in advance.

However, it is difficult to uniquely determine $P(i)$ because a performance includes errors and leaps forward and backward. Thus, we substitute mixed distribution for $P(i)$. Accordingly, we estimate $P(\theta|px)$, where px means the discrete frequency distribution of θ when note i occurs. The parameter θ represents a random variable about the combination of gaze and keying information.

Thus, the random variables depend on the combination of note numbers and gaze points. To prevent θ from being sparse, the number of random variables was limited by using a gaze range, whereby the musical score was divided into 10 parts in the horizontal direction, instead of gaze points. The bottom part of Fig.4 shows the result of multiplying the mixture distribution by the likelihood. The current performance point e_m is determined so that the multiplication of $P(\theta|i)$ and $P_{GMM}(i)$ is maximized.

$$e_m = \arg\max_i P(\theta|i) P_{GMM}(i)$$

4. EXPERIMENT

4.1 Experimental Description

By properly assigning the parameters, which are gaze distribution, keying distribution and mixture rate, our system can estimate a user's current position correctly. The parameters are drawn from ground truth data which consists of true current position, MIDI key numbers, and gaze points. First, by MIDI data we know current notes that subjects are playing including wrong key strokes. Even if a subject played a wrong key, a subjects are instructed to continue playing without trying to recover wrong key strokes during piano performance as if a subject plays correct notes. Then we align the note that a subject plays with the corresponding gaze data. Thereby, the system acquires gaze distribution, keying distribution, and mixture ratio for each individual subject. After the parameters are identified, we start the experiment.

4.2 Implementation of Proposed Method

If we take a set piece that the subjects already know, it is possible that EHS will be biased due to prior knowledge of the phrases or decreased score reading time. Therefore, we should adopt a piece that none of the subjects knows. We select a piece from Yamaha Music Ability Test (Grade 5 Grade) Sight playing / Improvisation and extract the right-hand part of 27 measures (103 notes) [1]. The set piece contains three identical phrases and there are nine duplicated notes (Fig. 5). On a screen, we show the score image, which is made with MuseScore to remove any musical symbols such as staccato and slur. Thus, the bar lengths of a score on a screen are variable depending on the number of notes and symbols in each bar.

The GUI of the system is implemented in Processing and the model part in Python. The system obtains keying information from the MIDI keyboard and gaze information from a gaze measuring device, Eye-Tribe ET1000, which does not disturb performance because it is small and stationary [15]. The effective range of the distance between the device and user's eyes is 45 to 75cm, and the spatial resolution is the angle of 0.1 degree, which means the resolution of 0.17cm, 50cm ahead. To calibrate the system, sixteen points are showed on a screen one by one. The frame rate of the Eye-Tribe generates sampling data of gaze information at a frame rate of 30 Hz.

The gaze data is transmitted to the model part in Open Sound Control (OSC), which is the protocol for communication among computers, sound synthesizers, and other multimedia devices. After estimating a user's current position by the algorithm described previously, the model part transmits it to GUI. To superimpose the position of the estimated playing note, the melody of the set piece is displayed in the x-axis of note number i and the y-axis of MIDI note number.

4.3 Evaluation Procedure

To obtain experimental data, we asked seven subjects (Subjects A to G) to play a set piece. Five of the subjects (A to E) had experience of learning the piano, and the other two

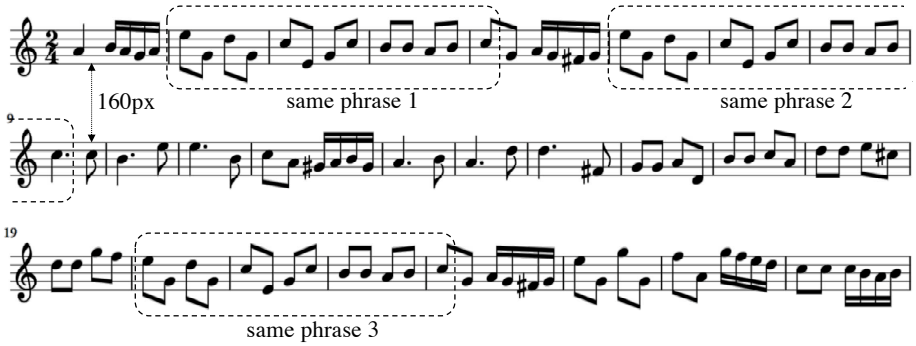

Figure 5. The set piece that contains three identical phrases and there are nine duplicated notes

(F and G) did not but could read a musical score. If a subject plays a melody linearly following a score, of course, the system can estimate almost every note correctly. Next, subjects were instructed to play parts of the piece in reverse order. For instance, subjects played from the 20th to the 23rd measure immediately before they played from the 6th to the 9th measure (Fig. 5). Here, the accuracy rate was defined as the rate at which performance position is correctly estimated with respect to all keystrokes.

4.4 Results and Discussion

Table 1 shows the number of passed-over keys, that of mistakenly pressed keys, the temporal interval between the restart of the system and the timing of capturing correct position (latency in the table), and the accuracy rate. In regard to the accuracy rate, we see a large gap between two groups: a group that has a rate of more than 90% (Subjects A to D and F) and the other group (E and G). It does not seem that the accuracy rates are related to the numbers of passed-over and mistakenly pressed keys. Concerning these numbers, Subjects F and G reach the largest numbers.

Table 2 shows the weights of the mixture rate. For Subjects A to E who have experience of learning the piano, the weights of DP matching are larger than those of the gaze distribution. In contrast, for Subjects F and G (non-experienced), the weights of gaze are larger than those of DP matching. We think a reason for this is that GMM determines the weights by the maximum likelihood. Hence, during score following, keying information plays an important role for the experienced subjects, while gaze information is important for the non- experienced ones.

Table 1. Accuracy rate and uncertain factors

	Subject	passed-over	missed	latency	acc. (%)
experienced	A	1	0	1	93.3
	B	0	0	2	93.3
	C	0	0	2	93.3
	D	0	0	0	90.0
	E	0	0	1	70.0
non experienced	F	1	2	0	96.6
	G	0	2	-	0.0

Table 2. The Weight of Mixture Rate

	Subject	weight of Gaze	weight of DP
experienced	A	0.13	0.87
	B	0.13	0.87
	C	0.15	0.85
	D	0.26	0.74
	E	0.24	0.76
non experimented	F	0.48	0.52
	G	0.76	0.24

Table 3. The Differences Between Gaze Points Before and After the Experiment

	Subject	Gap for x-axis(px)	Gap for y-axis(px)
experienced	A	1.9	-0.7
	B	29.8	-1.6
	C	-56.0	-70.9
	D	-27.0	-104.7
	E	-12.9	19.0
non-experienced	F	23.2	-143.7
	G	-85.8	-146.4

To examine the performance of the eye-tracking device, we instructed the subjects to look at the same points on the score before and after the experiment. Table 3 shows the average differences of gaze points before and after the experiment, called Gap in the table. The table shows that the absolute values of the experienced subject's gaps are smaller. The absolute value of G's gap is the largest, being shifted left by 85.8 px and upward by 146.4 px. These values correspond to a shift of a half measure to the left and almost 1 staff up in Fig. 5.

While the experimental results show that the proposed method determines the weights of information, to which the user's individuality is adapted, with little data, there are some cases in which position cannot be identified. To examine such cases, let us consider the relationship between

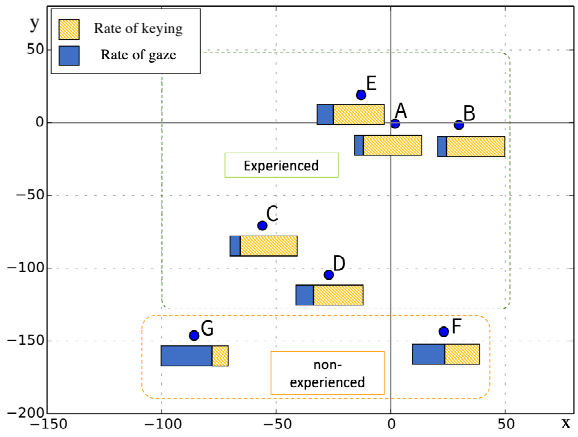

Figure 6. Relationship Between the Gap and the Mixture Rate

misalignment and the mixture rate. Fig. 6 shows the relationships between the gaps and the mixture rates in the scatter plot of the gap data in Table 3. The bars under dots represent the mixture rates for gaze and keying distributions. In the figure, we can see a trend in which the accuracy depends on the mixture ratio. In particular, the gaze mixture rate of Subject F is 0.48, on the other hand, that of Subject G is 0.76. Although the gaze information of subject G is weighted more heavily, G's gaps are also large. We think it is for this reason that the accuracy rate decreases.

5. CONCLUDION

In this paper, we proposed a score-following system which adapts a user's individuality in piano performance. We fit gaze information and keying information to the normal distribution and integrate them into the Bayesian inference by using GMM. The experimental results demonstrate the relevance of each type of information to the user's individuality. In particular, the keying information is important for an experienced performer, on the other hand, the gaze information is important for a non-experienced one. However, regarding the accuracy of the system, large differences occur among subjects. We think one of the reasons is not considering misalignment of the eye-tracking system. Future work will include developing a robust method to deal with errors related to eye-tracking, and adding other kinds of information which reflect user's individuality and/or mind, such as gesture and blinking.

Acknowledgments

This work has been supported by JSPS Kakenhi 16H01744 and 16K12560, and JST CREST Grant Numbers JPMJCR17A1 and JPMJCR17A3, Japan.

6. REFERENCES

[1] Y. Asano, "Piano Performance Grade 5 Sight playing, Improvisation Variant Example Song Collection ('90 Revised Edition)", Yamaha Music, 1997.

[2] R. B. Dannenberg, "An on-line algorithm for real-time accompaniment", ICMC, pp.193-198, 1984.

[3] R. B. Dannenberg, and H. Mukaino, "New techniques for enhanced quality of computer accompaniment", ICMC, 1988.

[4] M. Dorfer, A. Arzt, and G. Widmer, "Towards score following in sheet music images", ISMIR, 2016.

[5] M. Dorfer, F. Henkel, and G. Widmer, "Learning to Listen, Read, and Follow: Score Following as a Reinforcement Learning Game", ISMIR, 2018.

[6] P. Desain, H. Honing, "Tempo Curves Considered Harmful. In Music Mind and Machine", Amsterdam: Thesis Publishers, pp.25-40, 1992.

[7] S. Furneaux, M. F. Land, "The Effects of Skill on the Eye-Hand Span During Musical Sight-Reading", Proceedings of the Royal Society of London B, 266, pp. 2435-2440.

[8] L. Grubb, R. B. Dannenberg, "Enhanced Vocal Performance Tracking Using Multiple Information Sources", ICMC, pp.37-44, 1998.

[9] S. Kobori, K. Takahashi, "Cognitive Processes During Piano and Guitar Performance: An Eye Movement Study", ICMPC, pp.748-751, 2008.

[10] C. Oshima, K. Nishimoto and M. Suzuki, "A piano duo performance support system to motivate children's practice at home (in Japanese)", IPSJ, 46(1), pp. 157-171, 2005.

[11] S. Sagayama, T. Nakamura, E. Nakamura, Y. Saito, H. Kameoka, and N. Ono, "Automatic Music Accompaniment Allowing Errors and Arbitrary Repeats and Jumps", POMA, vol. 21, 035003, pp. 1-11, 2014.

[12] H. Takeda, T. Nishimoto, and S. Sagayama, "Automatic Accompaniment System of MIDI Performance Using HMM-based Score Following (in Japanese)", IPSJ, 2006.

[13] S. Terasaki, Y. Takegawa, K. Hirata, "Proposal of Score-Following Reflecting Gaze Information on Cost of DP matching", ICMC, pp. 144-149, 2017.

[14] J. Sloboda, "The eye - hand span - an approach to the study of sight reading", Psychology of Music, 2(2), pp. 4-10, 1974.

[15] The Eye Tribe, available from https://theeyetribe.com/ [25 Dec2018].

[16] E. Nakamura, P. Cuvillier, A. Cont, N. Ono, and S. Sagayama, "Autoregressive Hidden Semi-Markov Model of Symbolic Music Performance For Score Following", ISMIR, 2015.

[17] E. Nakamura, T. Nakamura, Y. Saito, N. Ono, and S. Sagayama, "Outer-product hidden Markov model and polyphonic MIDI score following", Journal of New Music Research, vol. 43, no. 2, pp. 183-201, Apr 2014.

[18] B. Pardo and W. Birmingham, "Modeling form for on-line Following of Musical Performances", Proc. of the Twentieth National Conference on Artificial Intelligence, 2005.

[19] V . Thomas, C. Fremerey, M. Muller, and M. Clausen, "Linking Sheet Music and Audio Challenges and New Approaches", Dagstuhl Follow-Ups, 3: pp. 1-22, 2012.

[20] Weaver, H. E. , "Studies of Ocular Behavior in Music Reading", Psychological Mono-graphs, 55(1), pp.1-29, 1943.