The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Generating Natural Hand Motion in Playing a Piano

Kazuki Yamamoto, Etsuko Ueda, Tsuyoshi Suenaga, Kentaro Takemura,
Jun Takamatsu and Tsukasa Ogasawara

***Abstract*— Generating natural motion of an articulated object with higher DOF (*e.g.*, humanoid robot and robot hand) is a crucial issue in robotics and computer graphics fields. Use of the motion capture data is one of the solutions, but it requires expensive device and time-consuming measurement. In this paper, we propose a method for generating natural hand motion to play a piano from the inputted music score. The proposed method uses inverse kinematics while considering naturalness of hand poses. We revisit background of the inverse kinematics based on the maximum likelihood estimation and use the prior model of the hand pose to achieve the naturalness. We evaluate the effectiveness of the proposed method using voluntary survey.**

I. Introduction

Generating natural motion of an articulated object with higher DOF (*e.g.*, humanoid robot and robot hand) is very important. For example, visual improvement in computer graphics (CG) provides us with an alternative method to create animation. However, difficulty in generating natural motion seriously deteriorates quality of the created animation. The robotics technology, such as inverse kinematics and dynamics, is employed to overcome such a problem.

Inverse kinematics is a method to calculate the configuration of an articulated object, which satisfies several conditions represented in the world coordinate system. This is usually equivalent to solve redundant system of simultaneous non-linear equalities. The Newton method [7] is employed to solve. For example, Yamane and Nakamura [10] proposed the pin-drag interface to generate natural human poses by employing the method. Human interaction achieves naturalness of the generated poses. Komura *et al.* [3] proposed to define the metric in the configuration space to realize the naturalness. Sentis and Khatib [8] progressively reduced the redundancy by adding the constraints when generating motion of a humanoid. They mainly consider physical consistency and naturalness of the motion is out of their scope.

Use of the motion capture data is simple but powerful for generating natural motion. One disadvantage is to measure all the necessary motions using the special equipment, *i.e.* motion capture system. To reduce the cost in measurement, the *motion graph* technique (*e.g.* [4]) was proposed. To generate the target motion, this technique first rearranges fragments of the captured motions and next interpolates gaps between the fragments. However, due to the limitation of the interpolation, any rearrangements and any fragments are not permitted. Thus, variations of the generated motions are limited.

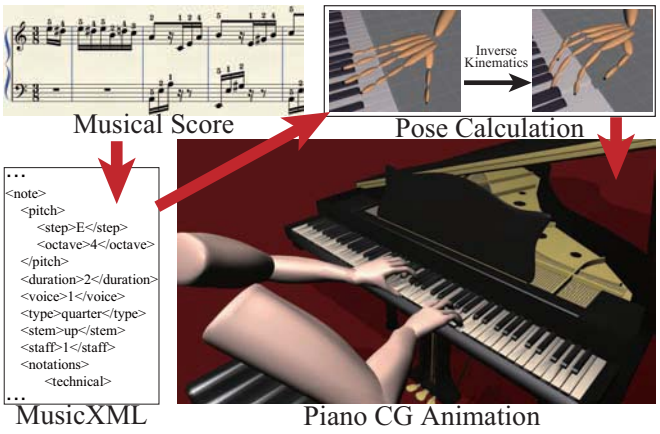

Fig. 1. Overview of the proposed method. The method only uses a music score as input and outputs CG animation of playing a piano. The method uses the key-frame technique. The key is how to calculate the natural key pose at each key frame.

Recently, the framework to combine dynamic simulator with the motion capture data is focused on. Pollard and Zordan [6] extracted PD-control parameter of human grasping from the motion capture data. Given the contacts between a hand and an object, this PD-controller generates grasping motion. Many proposed methods such as [12], [9] concentrated on generating the motion after physical interaction occurs. These methods mainly consider physical consistency and expect that the naturalness originates from the motion capture data.

In this paper, we propose the method for generating natural hand motion to play a piano from the inputted music score. Though we also employ motion capture data supplementarily, we interpret them to extract aspects of the naturalness and combine the interpretation with inverse kinematics. This enables the proposed method to remove the restriction on types of music scores.

Though the target task seems to be niche, there are many important aspects from robotics perspective. First, the target motion is very quick, but complex and dexterous. Second, hand pose at a specific moment must satisfy several constraints simultaneously (*e.g.*, in the case to play some chord). Since these two aspects are observed in other dexterous motions, we believe the applicability of the proposed method to other types of natural motions.

K. Yamammoto, T. Suenaga, K. Takemura, J. Takamatsu, and T. Ogasawara are with Graduate School of Information Science, Nara Institute of Science and Technology, Japan {kazuki-y, tsuyo-s, kenta-ta, j-taka, ogasawar}@is.naist.jp

E. Ueda is with Department of Control Engneering, Nara National College of Technology, Japan ueda@ctrl.nara-k.ac.jp

Figure 1 shows the overview of the proposed method. The method only uses a music score as input and outputs CG animation where a visually realistic hand model plays a piano, as shown in the right bottom of the figure. The method uses the key-frame technique. First, it extracts the necessary information from the score. Next, key frames are defined from this information. Then, a natural key pose is calculated at each key frame. The animation is generated by interpolating these key poses.

We assume that it is previously known which finger strikes the target key. The research method to solve this fingering issue has already been proposed, such as [11]. Since small and lightweight objects only move in playing a piano, we assume that dynamic non-linearity property is not dominant. Thus, we mainly concentrate on kinematic issues in this paper.

II. Preliminary: inverse kinematics

Issue in the inverse kinematics is generally formulated as Eq. (1):

$$\mathbf{p} = \begin{pmatrix} f_1(\mathbf{q}) \\ \vdots \\ f_n(\mathbf{q}) \end{pmatrix}, \tag{1}$$

where n is the number of constraints, the function f_i returns configuration of some link in the world coordinate system when giving robot's configuration $\mathbf{q}$ as input, and $\mathbf{p}$ ($\in R^n$) represents the desired values of the configurations in the world coordinate system, such as location of a finger tip in playing.

Since this equation is non-linear and thus difficult to analytically solve, the Newton method [7] is usually employed. Eq. (1) is approximately represented around some given configuration $\mathbf{q}_0$ as

$$\mathbf{p} = \mathbf{J}(\mathbf{q} - \mathbf{q}_0) + \mathbf{p}_0, \tag{2}$$

where $\mathbf{p}_0 = (f_1(\mathbf{q}_0), \ldots, f_n(\mathbf{q}_0))^T$ and $\mathbf{J}$ is the Jacobian matrix defined as

$$\mathbf{J} \triangleq \begin{pmatrix} \frac{\partial f_1}{\partial \mathbf{q}}^T \\ \vdots \\ \frac{\partial f_n}{\partial \mathbf{q}}^T \end{pmatrix}.$$

Eq. (2) is solved using the pseudo-inverse matrix $\mathbf{J}^+ (\triangleq \mathbf{J}^T(\mathbf{J}\mathbf{J}^T)^{-1})$ as

$$\mathbf{q} = \mathbf{J}^+\dot{\mathbf{p}} + (\mathbf{I} - \mathbf{J}^+\mathbf{J})\mathbf{k} + \mathbf{q}_0, \tag{3}$$

where $\mathbf{I}$ is the identity matrix, $\dot{\mathbf{p}} \triangleq \mathbf{p} - \mathbf{p}_0$, and $\mathbf{k}$ is a free parameter vector.

We need to assign the appropriate value to $\mathbf{k}$ in order to generate natural poses. In $\mathbf{k} = \mathbf{0}$, the calculated $\mathbf{q}$ minimizes $\dot{\mathbf{q}}^T\dot{\mathbf{q}}$, where $\dot{\mathbf{q}} \triangleq \mathbf{q} - \mathbf{q}_0$. Instead of deciding the appropriate $\mathbf{k}$, several methods used a weighted full-rank matrix $\mathbf{W}$ and solved Eq. (2) by minimizing $\dot{\mathbf{q}}^T\mathbf{W}\dot{\mathbf{q}}$, not $\dot{\mathbf{q}}^T\dot{\mathbf{q}}$. For example, Komura *et al*. [3] decided the weighted matrix $\mathbf{W}$ from the motion capture data. This solution is obtained by using $\mathbf{J}_w^+$ in place of $\mathbf{J}^+$, where

$$\mathbf{J}_w^+ \triangleq \mathbf{W}^{-1}\mathbf{J}^T(\mathbf{J}\mathbf{W}^{-1}\mathbf{J}^T)^{-1}. \tag{4}$$

III. Generating natural hand pose

In this section, we describe a method for generating natural hand poses, since motion can be regarded as a sequence of the poses.

A. Revisit background of inverse kinematics

We propose a method for deciding $\mathbf{k}$ in Eq. (3) based on naturalness of hand pose. To achieve this, we give another insight to the inverse kinematic based on the maximum likelihood estimation and use the prior model to decide natural hand poses.

Given the condition $\mathbf{p}$, the desired configuration $\hat{\mathbf{q}}$ is obtained by maximizing Eq. (5) based on the maximum likelihood manner.

$$\hat{\mathbf{q}} = \underset{\mathbf{q}}{\operatorname{argmax}}\, p(\mathbf{q}|\mathbf{p}). \tag{5}$$

By substituting $p(\mathbf{q}|\mathbf{p}) = \frac{p(\mathbf{p}|\mathbf{q})p(\mathbf{q})}{p(\mathbf{p})}$ into Eq. (5), Eq (6) is obtained.

$$\hat{\mathbf{q}} = \underset{\mathbf{q}}{\operatorname{argmax}}\, p(\mathbf{p}|\mathbf{q})p(\mathbf{q}). \tag{6}$$

Note that we assume $p(\mathbf{q})$ follows the uniform distribution. The first term $p(\mathbf{p}|\mathbf{q})$ in the right side represents the degree of satisfaction to the condition and the second term $p(\mathbf{q})$ represents the prior model (*i.e*., easiness) of hand configuration.

Since location of a finger tip is expressed in deterministic manner, the condition $p(\mathbf{q}|\mathbf{p})$ is defined as the Dirac delta function. Roughly speaking, when the condition represented in Eq. (1) does not holds, $p(\mathbf{q}|\mathbf{p})$ is zero, *i.e*., these $\mathbf{q}$ can be removed as candidates for the maximization. Considering this, Eq. (6) is equivalently transformed to:

$$\hat{\mathbf{q}} = \underset{\mathbf{q}}{\operatorname{argmin}}\, E(\mathbf{q}) \text{ suppose to Eq. (1)},$$

where $E(\mathbf{q}) = -\log p(\mathbf{q})$. This shows that the prior model of hand pose is one of the solutions to achieve the naturalness.

B. Constructing prior model

We measure a person playing a piano using motion capture system to construct the prior model. Figure 2 shows positions of the infrared reflection markers attached to the player, who played Beethoven's *Für Elise* and Mozart's *Piano Sonata No. 16 in C major, K. 545*, while keeping a tempo by a metronome. He plays a piano for 18 years and is equivalent to the semipro level.

Figure 3 shows histograms of joint angles of elbow, wrist, MP, PIP, DIP joints in middle finger, and MP and IP joints in thumb. This figure reveals the following two things: 1) each histogram has a peak. 2) variance (*i.e*., width of histogram) is different from each other; for example, variances in elbow and wrist are smaller than those in the others. The former indicates existence of the pose in rest and the latter indicates

Fig. 2. Positions of the attached infrared reflection markers. We attached the marker on each joint to simplify estimating trajectories of joint angles.

easiness of move. We formulate the prior model as the Gaussian distribution to simplify the calculation and its parameters are decided from that observation and the piano textbook [2].

C. Inverse kinematics with naturalness

Since the prior model is formulated as the Gaussian distribution, $E(\mathbf{q})$ in Eq. (5) is simply represented as

$$E(\mathbf{q}) = (\mathbf{q} - \mathbf{q}_b)^T \Sigma^{-1} (\mathbf{q} - \mathbf{q}_b), \tag{7}$$

where $\mathbf{q}_b$ is the pose in rest and Σ is the covariance matrix. Using Eq (3),

$$\begin{aligned} \mathbf{q} - \mathbf{q}_b &= \mathbf{J}^+ \dot{\mathbf{p}} + (\mathbf{I} - \mathbf{J}^+ \mathbf{J})\mathbf{k} + \mathbf{q}_0 - \mathbf{q}_b \\ &= \mathbf{A}\mathbf{k} - \mathbf{b}, \end{aligned} \tag{8}$$

where $\mathbf{A} = (\mathbf{I} - \mathbf{J}^+\mathbf{J})$ and $\mathbf{b} = \mathbf{q}_b - \mathbf{q}_0 - \mathbf{J}^+\dot{\mathbf{p}}$. From the least square minimization, the appropriate $\mathbf{k}$ should satisfy

$$\frac{\partial E(\mathbf{q})}{\partial \mathbf{k}} = 0.$$

From this equation and Eq. (7), Eq. (9) holds:

$$\mathbf{A}^T \Sigma^{-1} \mathbf{A}\mathbf{k} = \mathbf{A}^T \Sigma^{-1} \mathbf{b}. \tag{9}$$

This equation is solved by the singular value decomposition [7].

IV. Motion to play a piano

The proposed method employs the key-frame technique to generate hand motion in playing a piano. This technique generates the motion by interpolating key poses defined in several key frames. At least, these two issues should be considered:

- Assign appropriate key frames
- Generate natural key poses

In this paper, we employ liner interpolation to generate the motion. Since motion in playing a piano is very quick, difference of interpolation is peripheral.

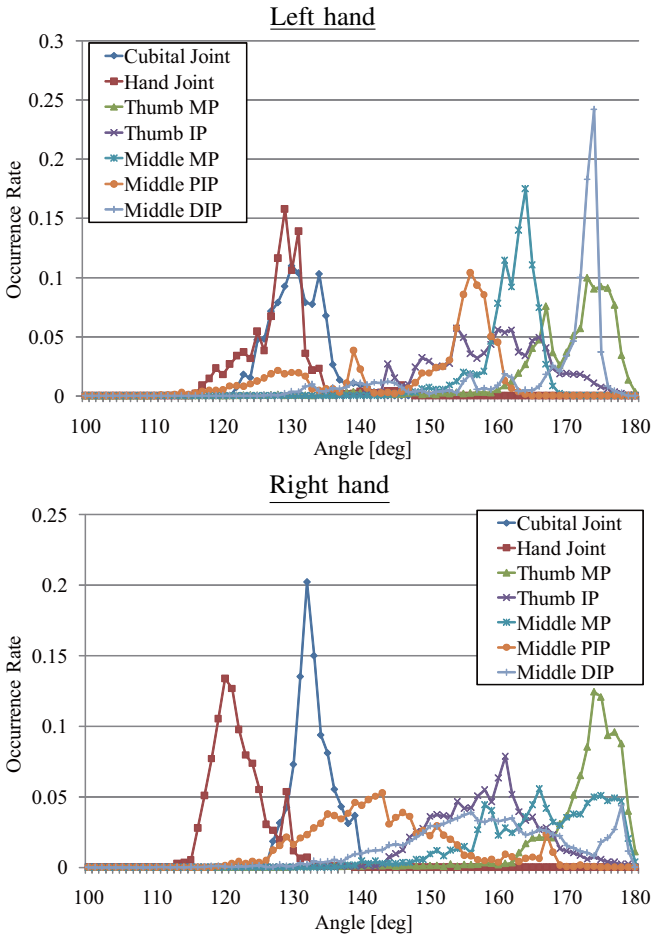

Fig. 3. Occurrence rate of joint angles in each hand. This figure shows that 1) each histogram has a peak. 2) variance (*i.e.*, width of histogram) is different from each other; for example, variances in elbow and wrist are smaller than those in the others.

A. Key frames

In order to avoid the limitation of maximum velocity of finger tip, motion to prepare keying a note (referred to as *pre-keying motion*) appears. Figure 4 shows one example of the pre-keying motion; the index finger approaches the target key, while the thumb keeps putting on the key.

We regard three focusing moments, *pre-keying*, *keying*, and *releasing*, as key frames. The last two key frames are directly derived from a music score. The first key frame is assigned a few handled milliseconds (in this paper, 100 [ms]) before keying the note. The duration is decided from observation of a person playing a piano.

B. Key poses

Positions of finger tips at pre-keying, keying, and releasing moments are derived from a music score. Once given the positions, we generate hand poses using the method proposed in Section III.

C. Examples

Consider the case to play the music score in Figure 5. Let y_n^{over} (3 millimeter higher then the key) and y_n^{key} be the

Fig. 4. Example of pre-keying motion. The index finger approaches the target key, while the thumb keeps putting on the key.

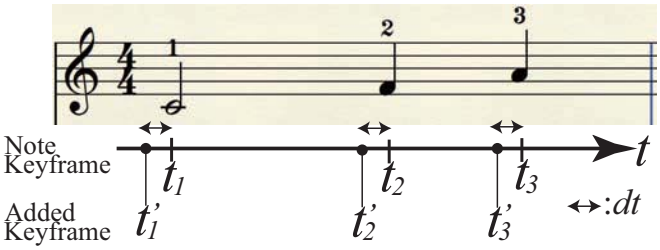

Fig. 5. Key frames obtained from the score. Considering pre-keying motion, three pre-keying key-frames at time $t_1^{'}$, $t_2^{'}$, $t_3^{'}$ are added.

heights of finger n's tips (1: thumb, 2: index finger, in this example) at pre-keying and keying moments, respectively. The height of thumb's tip is set to y_1^{over} at time $t_1^{'}(= t_1 - dt)$ and is set to y_1^{key} at time t_1, respectively. At time $t_2^{'}$, which corresponds to pre-keying moment, the heights of thumb's tip and index finger's tip are set to y_1^{key} and y_2^{over}, respectively. At time t_2, which corresponds to keying moment, the heights of thumb's tip and index finger's tip are y_1^{over} and y_2^{key}, respectively. Figure 6 summarizes all the key poses of the score.

V. Experiment

A. Input

e

As mentioned above, we employed files written in the *MusicXML* format [1], which is tag-based format like XML. Software for music composer, such as *Finale*, outputs such files.

First, we extract the following information from the file:

- Tempo
- Pitch
- Duration
- Finger ID

Pitch and Finger ID are employed for calculating hand pose. Tempo and duration are employed to decide moment of key frame. As described above, ID of the keying finger is previously assigned.

B. Hand model

Figure 7 shows the hand model used in this experiment. The position of the shoulder is fixed. All the joints are rotational in order to unify the unit. This figure shows the model where all the joints are set to zero. The arrows indicate directions of rotation axes. The model has 27 DOF in each hand, *i.e.*, 54 DOF in total. Note that we designed link lengths and range of joint angles from several anatomical textbooks.

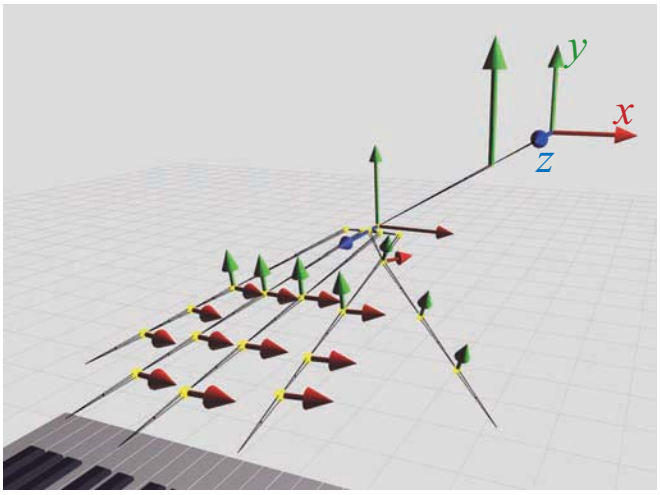

Fig. 7. Kinematic configuration of the hand when joint angles $\theta = \mathbf{0}$.

C. Playing a piano in computer graphics

We generated CG animation to play the following two music pieces:

- Piano Sonata No. 16 in C major, K. 545 (Mozart)
- Für Elise (Beethoven)

Figure 8 shows the animation of K. 545. All the movies can be downloaded from http://robotics.naist.jp/~j-taka/movie/.

D. Evaluation

We verified the effectiveness of the proposed method using voluntary survey. It is worth noting that one of the reasonable methods for evaluating the naturalness is directly compare the generated motion to that recorded by the motion capture system. We did not adopt the method due to hesitating to converge biased evaluation; the observed pose is just one of all the natural poses. As long as we know, there is no method for quantitatively evaluate the naturalness. In conclusion, we rely on the evaluation by voluntary survey.

We generated the following four kinds of CG animation for the comparison:

A. $\mathbf{k} = \mathbf{0}$, $\mathbf{W} = \mathbf{I}$ (Simple)
B. $\mathbf{W}$ is adjusted and $\mathbf{k} = \mathbf{0}$ (*e.g.*, Komura's method [3])
C. The prior model is used and $\mathbf{W} = \mathbf{I}$
D. $\mathbf{W}$ is adjusted and the prior model is used

In this experiment, we used the identity matrix as the covariance matrix Σ in Eq. (7). We also compare animations where pre-keying moment is not used (referred to as *Group a*) or used (reffed to as *Group b*) as key frames. There are eight patterns in total.

Fig. 6. Key poses considering pre-keying motion in the score of Fig. 5.

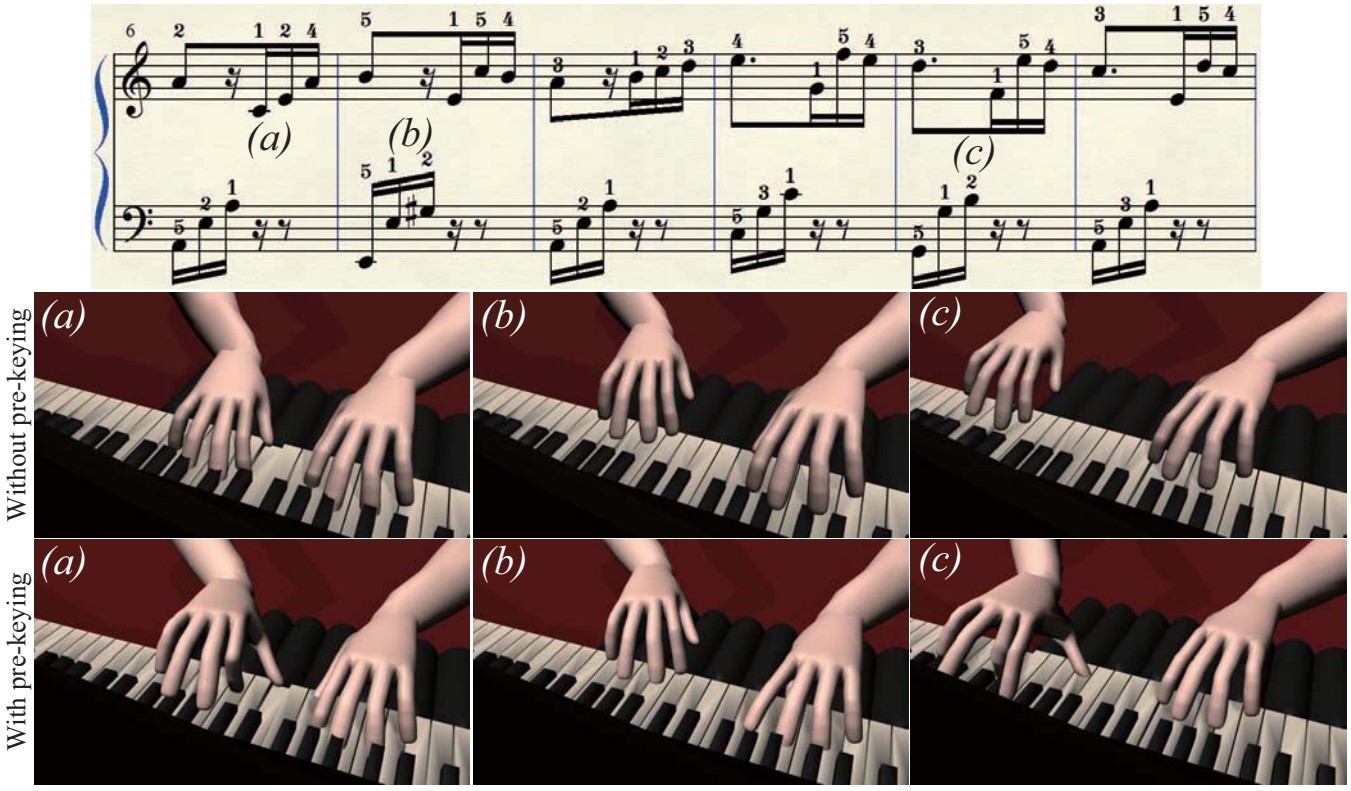

Fig. 10. Comparison of piano-playing CG animations without (the upper row) or with pre-keying motion (the lower row)

For each of the two music pieces, we evaluated all the eight patterns using the Thurstone method [5], a kind of the paired comparison analysis. All the 28 pairs of CG animation are shown to volunteers (26 volunteers in Für Elise and 25 volunteers in K. 545). They select the winner (*i.e.*, more natural CG animation) from the two in each pair.

Figure 9 shows the comparison result. The pattern located on the right side is better. The alphabets, A to D, and a and b, indicate type of the pattern.

Discussion: In all the cases, the eight patterns are roughly classified to two groups: top five patterns (Ba, Ab, Bb, Cb, Db) and bottom three patterns (Aa, Ca, Da). Comparison between Group a and Group b supports the effectiveness of assigning pre-keying key-frames. Only one exception is Ba, which is positively evaluated despite no use of pre-keying key-frames. However, in total, use of pre-keying key-frames derives the better results.

Comparison between Group { Ab, Bb } (not use the prior model) and Group { Cb, Db } (use the prior model) proves that the prior model is useful for generating natural motion. However, differences between Ab and Bb, and between Cb and Db is not so large. Although fine tuning of the weighted matrix $\mathbf{W}$ may improve the effectiveness, it is difficult to search for the appropriate $\mathbf{W}$.

Scores of Cb and Db are nearly best, but scores of Ca and Da are nearly worst. The reason is that the finger moves too quick in order to maintain natural hand poses. Figure 10 shows the visual comparison between Da and Db. The right hand in image (a) of the lower row prepares to key the note E4 while keeping keying the note C4. The similar pre-keying procedure can be seen in image (c) of the lower row. Such a pose cannot be seen at all in the upper row. The image (b) of the lower row shows tricky motion where the thumb of the left hand pass through under the palm. Such motion often appears in human playing. Just considering pre-keying key-frames generates it.

It is sure that various aspects to achieve the naturalness should be simultaneously considered. Comparison between

Fig. 8. Generated CG animation of playing a piano

Ba and Bb also shows the bad influence caused by independently applying one of the three techniques, *i.e.*, pre-keying key-frames.

In summary, use of both the prior model and pre-keying key-frame is effective for generating natural hand motion. Although use of the weighted matrix $\mathbf{W}$ may improve the naturalness, the improvement depends on appropriateness of the parameter.

VI. Conclusion

In this paper, we proposed a method for generating natural hand motion to play a piano from the inputted music score. We employed the key-frame technique and proposed the methods to generate natural key poses and to assign appropriate key frames. To achieve the first method, we revisited background of the inverse kinematics based on the maximum likelihood estimation and proposed to use the prior model of the hand pose. We evaluated the effectiveness of the proposed method using voluntary survey. We concluded that combination of all the techniques improves the quality of the

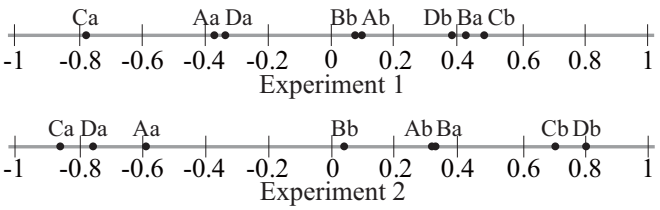

Fig. 9. Score of each method by the Thurstone method. The pattern located on right side is superior. This result shows the animations generated by the proposed method, *i.e.*, Cb and Db, are the best.

created CG animation but apply an individual technique may deteriorate the quality. We confirmed that various aspects should be considered to achieve the naturalness.

One of the future directions is to generate motions under various note's volume, articulation, and emotion. We think that adjusting various parameters, such as timing of pre-keying moment and height of the finger tips, is one of the solutions. Another direction is to represent individuality in playing a piano. To simplify the implementation, we used the simple prior model, where the joint angle is independently moves and each joint angle follows the Gaussian distribution. The elaborate prior model such as hidden Markov model may realize the individuality. Another direction is to apply this framework to the other kind of motion.

References

[1] `http://www.musicxml.org`.

[2] S. Bernstein. *Twenty Lessons In Keyboard Choreography*. Seymour Bernstein Music, 1991.

[3] T. Komura, A. Kuroda, S. Kudoh, C.-L. Tai, and Y. Shinagawa. An Inverse Kinematics Method for 3D Figures with Motion Data. In *Proc. of Computer Graphics International*, pages 266–271, 2003.

[4] L. Kovar, M. Gleicher, and F. Pighin. Motion Graphs. In *Proc. of the ACM SIGGRAPH*, pages 473–482, 2002.

[5] L. L.Thurstone. The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology*, 21:384–400, 1927.

[6] N. S. Pollard and V. B. Zordan. Physically based grasping control from example. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, 2005.

[7] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing (2nd edition)*. Cambridge University Press, 1992.

[8] L. Sentis and O. Khatib. Synthesis of whole-body behaviors through hierarchical control of behavioral primitives. *Int. J. of Humanoid Robotics*, 2(4):505–518, 2005.

[9] T. Shiratori, B. Coley, R. Cham, and J. K. Hodgins. Simulating balance recoverry responses to trips based on biomechanical principles. In *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation*, 2009.

[10] K. Yamane and Y. Nakamura. Natural motion animation through constraining and deconstraining at will. *IEEE Trans. on Visualization and Computer Graphics*, 9(3):352 – 360, 2003.

[11] Y. Yonebayashi, H. Kameoka, and S. Sagayama. Automatic decision of piano fingering based on hidden markov models. In *Proc. of the 20th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 2915–2921, 2007.

[12] V. B. Zordan and J. K. Hodgins. Motion capture-driven simulations that hit and react. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 89–96, 2002.