THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

TECHNICAL REPORT OF IEICE.

Motion Prediction Based on Eigen-Gestures

Masato NAKAJIMA†, Seiichi UCHIDA†, Akihiro MORI†, Ryo KURAZUME†,

Rin-ichiro TANIGUCHI†, Tsutomu HASEGAWA†, and Hiroaki SAKOE†

† Graduate School of Information Science and Electrical Engineering, Kyushu University
Motooka 744, Nishi-ku, Fukuoka-shi, Fukuoka, 819–0395 Japan
E-mail: †nakajima@human.is.kyushu-u.ac.jp

Abstract Motion prediction is important for the realization of "proactive" gesture-based man-machine interaction systems, which can react in various ways before the end of user's action. In this paper, two motion prediction methods are proposed. The first method is based on a naive extrapolation using a reference gesture. The second method is based on eigen-gestures, which represent typical spatial and temporal variations from the reference gesture. Experimental results show that the second method outperforms the first method because the eigen-gestures are useful to compensate various changes in input gestures.

Key words gesture recognition, prediction, eigen-gesture, principal component analysis

1. Introduction

Motion prediction is an important task for sophisticated gesture-based man-machine interaction systems. If a system predicts user's subsequent posture, the system can react in advance to the end of user's action. This "proactive" reaction will help users in various ways. For example, the user can stop her/his action after the system determines its prediction and start appropriate reaction. In addition, the system can provide intelligent advice for the user if the prediction result indicates user's failure.

In this paper, two recognition-based motion prediction methods are proposed. Both of those methods are commonly based on *early recognition* of gestures. The early recognition is a technique for determining the category of a current gesture in its beginning part. For example, the category "hurrah" is determined at the moment that both hands begin to rise.

The first motion prediction method is a simple method where the subsequent posture is predicted by a naive extrapolation. The second motion prediction method is a more sophisticated method where the subsequent posture is predicted as a linear combination of several principal gestures, called eigen-gestures, which represent spatial and temporal variations of input gestures.

The second motion prediction method is related to the past trials on the principal component analysis (PCA) of gestures (e.g., [1], [2]). In addition, it is related to the past image interpolation techniques using PCA, since our prediction problem can be considered as a kind of interpolation/extrapolation problem. Amano [3] has proposed PCA-based image interpolation methods, called BPLP and kBPLP, for recovering missing pixels on an image. Naster et al. [4] have proposed an image warping method with PCA-based constraints and applied it to an image interpolation problem. The latter method can be considered as an extension of those PCA-based image interpolation methods to motion prediction.

2. Early recognition of gestures – outline

All of the proposed motion prediction methods are based on gesture recognition. Let a vector sequence $\boldsymbol{r}_c = r_{c,1}, \ldots, r_{c,t}, \ldots, r_{c,T_c}$ be the reference gesture of the category c where $r_{c,t}$ is a D-dimensional feature vector representing the posture of a user at the frame t. For example, $\boldsymbol{r}_c$ represents a typical posture sequence of a gesture "bye". Similarly, let a vector sequence $\boldsymbol{x} = x_1, x_2, \ldots, x_\tau, \ldots, x_T$ be an input gesture[1]. A conventional gesture recognition algorithm can estimate the category $c = c^*$ of the input gesture by matching the *entire* sequence of $\boldsymbol{r}_c$ to the *entire* sequence of $\boldsymbol{x}$ under nonlinear frame alignment to compensate temporal difference between two gestures. In this paper, we use a dynamic programming (DP) matching algorithm [5], [6] for nonlinear frame alignment.

[1] For simplicity, we assume that $\boldsymbol{x}$ comprises a single gesture, although we can easily extend the proposed method to deal with gesture sequences.

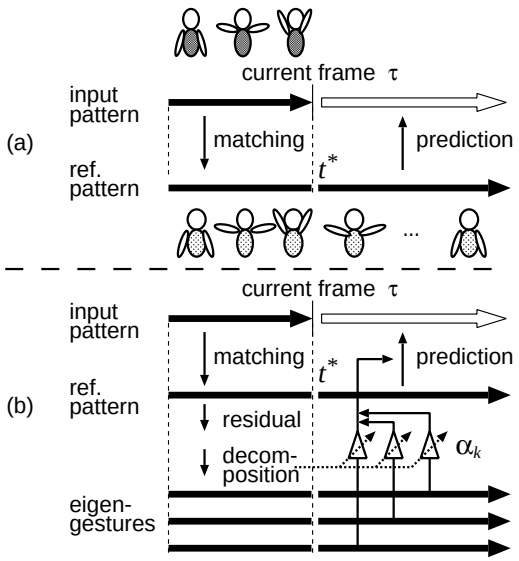

Figure 1 Illustration of two motion prediction methods: (a) Prediction based on simple extrapolation. (b) Prediction based on a linear combination of (spatial) eigen-gestures.

Instead of the above conventional algorithm, an *early recognition algorithm* is developed for the motion prediction methods which are discussed later. The early recognition algorithm can estimate not only the category $c = c^*$ but also the frame $t = t^*$ corresponding to the current input frame τ by matching the *beginning part* of the reference $r_{c,1}, \ldots, r_{c,t}$ to the *beginning part* of the input $x_1, \ldots, x_\tau$ (Fig. 1(a)). An early recognition algorithm, which is realized by modifying the conventional DP-based matching algorithm, is detailed in Appendix.

3. Motion prediction based on simple extrapolation

A simple motion prediction method is described in this section. The method utilizes the result of the early recognition algorithm, i.e., (c^*, t^*) and predicts the subsequent posture at the frame $\tau + \delta$ as

$$\widehat{x}_{\tau+\delta} = r_{c^*,t^*+\delta}. \tag{1}$$

Figure 1(a) illustrates this prediction method. The prediction method employs the extrapolation where the reference gesture itself is used as the predicted posture. *If there is neither spatial variation nor temporal variation*, we can expect good accuracy by the prediction method. Conversely, if an input gesture varies from the reference gesture spatially and/or temporally, the prediction accuracy will be degraded.

4. Motion prediction based on eigen-gestures

4. 1 Estimating eigen-gestures

This section describes another motion prediction method based on "eigen-gestures", which are introduced for coping with spatial and temporal variations of gestures. This section describes a procedure for estimating the eigen-gestures. The eigen-gestures are comprised of two components; *spatial eigen-gestures* and *temporal eigen-gestures*.

4. 1. 1 Spatial eigen-gestures

The spatial eigen-gestures represent frequent spatial variations of the gestures of a category. Let $\boldsymbol{y}_{c,n}$ denote the nth $(n = 1, \ldots, N)$ training sample of the category c. Then a spatial difference vector $\boldsymbol{v}^{\mathrm{s}}_{c,n}$ of $\boldsymbol{y}_{c,n}$ from $\boldsymbol{r}_c$ is defined as

$$\boldsymbol{v}^{\mathrm{s}}_{c,n} = \left(v^{\mathrm{s}}_{c,n,1} \cdots v^{\mathrm{s}}_{c,n,t} \cdots v^{\mathrm{s}}_{c,n,T_c}\right), \tag{2}$$

where $v^{\mathrm{s}}_{c,n,t}$ is a D-dimensional vector defined by

$$v^{\mathrm{s}}_{c,n,t} = y_{c,n,\rho_t} - r_{c,t}. \tag{3}$$

The vector $y_{c,n,\tau}$ denotes the τth frame of $\boldsymbol{y}_{c,n}$. The function $\tau = \rho_t$ denotes the nonlinear frame alignment between $\boldsymbol{y}_{c,n}$ and $\boldsymbol{r}_c$ optimized by the DP matching algorithm.

The spatial eigen-gestures are defined as principal components of a set of the spatial difference vectors and thus represent frequent posture variations of the category. Although training samples have different frame lengths, all the spatial difference vectors $\{\boldsymbol{v}^{\mathrm{s}}_{c,n} \mid n = 1, \ldots, N\}$ have the same dimensionality $(T_c \cdot D)$ by the nonlinear frame alignment, $\tau = \rho_t$. This equi-dimensionality of the spatial difference vectors allows having their $(T_c \cdot D) \times (T_c \cdot D)$ covariance matrix $\boldsymbol{\Sigma}^{\mathrm{s}}_c$ and mean vector μ^{s}_c. The spatial eigen-gestures $\{\boldsymbol{u}^{\mathrm{s}}_{c,k} \mid k = 1, \ldots, K\}$ of the category c are then provided as the principal components, i.e., the eigenvectors of $\boldsymbol{\Sigma}^{\mathrm{s}}_c$ with the K largest eigenvalues.

4. 1. 2 Temporal eigen-gestures

A temporal difference vector $\boldsymbol{v}^{\mathrm{t}}_{c,n}$ of $\boldsymbol{y}_{c,n}$ is defined from the nonlinear frame alignment $\tau = \rho_t$, which is already provided for the spatial difference vector. Specifically, the tth element of $\boldsymbol{v}^{\mathrm{t}}_{c,n}$ is the difference $\rho_t - t$. Similarly to $\boldsymbol{v}^{\mathrm{s}}_{c,n}$, the temporal difference vectors $\{\boldsymbol{v}^{\mathrm{t}}_{c,n} \mid n = 1, \ldots, N\}$ have an equi-dimensionality, T_c, even though training samples have different lengths. Thus, we have their $T_c \times T_c$ covariance matrix $\boldsymbol{\Sigma}^{\mathrm{t}}_c$ and mean vector μ^{t}_c. The temporal eigen-gestures $\{\boldsymbol{u}^{\mathrm{t}}_{c,l} \mid l = 1, \ldots, L\}$ of the category c are then provided as the eigenvectors of $\boldsymbol{\Sigma}^{\mathrm{t}}_c$ with the L largest eigenvalues.

4. 2 Prediction using eigen-gestures

In this section, a motion prediction method using the

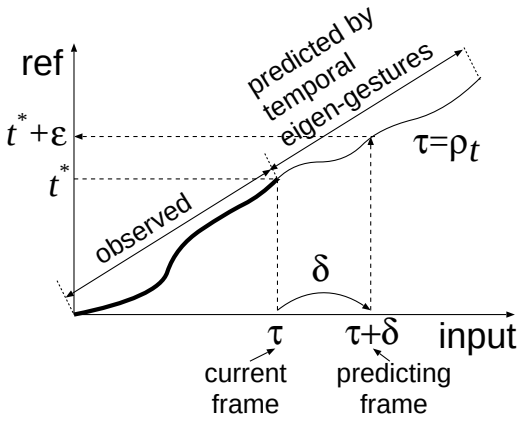

Figure 2 Prediction of temporal variation using temporal eigen-gestures.

spatial eigen-gestures is firstly described in 4. 2. 1 and then another motion prediction method using both the spatial and the temporal eigen-gestures is described in 4. 2. 2.

4. 2. 1 Prediction with spatial eigen-gesture only

Since the spatial eigen-gestures span a subspace of frequent posture variations, any spatial variation can be approximated by a weighted linear combination of the spatial eigen-gestures. Thus, *if there is no temporal variation*, we can predict the posture $\widehat{x}_{\tau+\delta}$ at the current frame τ by

$$\widehat{x}_{\tau+\delta} = r_{c^*,t^*+\delta} + \left(\mu^{\mathrm{s}}_{c^*,t^*+\delta} + \sum_{k=1}^{K} \alpha_k u^{\mathrm{s}}_{c^*,k,t^*+\delta} \right), \quad (4)$$

where $u^{\mathrm{s}}_{c,k,t}$ is the tth frame of the spatial eigen-gesture $\boldsymbol{u}^{\mathrm{s}}_{c,k}$ and α_k is its weight. The prediction method of (4) is comprised of the simple prediction term of (1) and a spatial variation term based on the spatial eigen-gestures. Figure 1 (b) shows an outline of the motion prediction method based on the spatial eigen-gestures.

The weight α_k is determined by fitting the beginning part of the input (i.e., the observed part of the input, $x_1, \ldots, x_\tau$) to the model (4). The optimal fitting is formulated as the minimization problem of the following criterion:

$$I(\alpha_1, \ldots, \alpha_K) = \sum_{t=1}^{t^*} \left\| \left\{ r_{c^*,t} + \left(\mu^{\mathrm{s}}_{c^*,t} + \sum_{k=1}^{K} \alpha_k u^{\mathrm{s}}_{c^*,k,t} \right) \right\} - x_{\rho_t} \right\| \quad (5)$$

where ρ_t denotes the nonlinear frame alignment between $x_1, \ldots, x_\tau$ and $r_{c^*,1}, \ldots, r_{c^*,t^*}$ and is already provided during the early recognition process for the interval $t \in [1, t^*]$. The minimization of I with respect to $\alpha_1, \ldots, \alpha_K$ is a simple least square error problem and thus has a closed form solution.

4. 2. 2 Prediction with spatial and temporal eigen-gestures

Now, we introduce the temporal eigen-gestures into (4) for coping with temporal variations. Under a temporal variation, we cannot assume that the $(\tau + \delta)$th frame of an input gesture will correspond to the $(t^* + \delta)$th frame of the reference pattern; it will correspond to the $(t^* + \epsilon)$th frame ($\epsilon \neq \delta$). Thus, the prediction equation (4) should be modified as

$$\widehat{x}_{\tau+\delta} = r_{c^*,t^*+\epsilon} + \left(\mu^{\mathrm{s}}_{c^*,t^*+\epsilon} + \sum_{k=1}^{K} \alpha_k u^{\mathrm{s}}_{c^*,k,t^*+\epsilon} \right). \quad (6)$$

For the prediction of $\widehat{x}_{\tau+\delta}$ using (6), we should determine ϵ. This determination can be done with the temporal eigen-gestures $\{u^{\mathrm{t}}_{c,l,t} \mid l = 1, \ldots, L\}$. Since the temporal eigen-gestures span a subspace of frequent temporal variations of the category c, their linear combination can represent any temporal variation of $\boldsymbol{x}$ from $\boldsymbol{r}_c$. Considering the frame correspondence illustrated in Fig. 2, the following equation holds:

$$\tau + \delta = \mu^{\mathrm{t}}_{c^*,t^*+\epsilon} + \sum_{l=1}^{L} \beta_l u^{\mathrm{t}}_{c^*,l,t^*+\epsilon}. \quad (7)$$

This equation determines ϵ, if δ and $\{\beta_l\}$ are given.

Since the prediction length δ will be specified by users, $\{\beta_l\}$ should be determined for fixing ϵ. This can be done by fitting the linear combination model of $\boldsymbol{u}^{\mathrm{t}}_{c^*,l}$ to the actual frame alignment ρ_t (i.e., the actual temporal variation of $\boldsymbol{x}$) for $1 \leqq t \leqq t^*$. The optimal fitting is formulated as the minimization problem of the following criterion:

$$J(\beta_1, \ldots, \beta_L) = \sum_{t=1}^{t^*} \left\| \left\{ t + \left(\mu^{\mathrm{t}}_{c^*,t} + \sum_{l=1}^{L} \beta_l u^{\mathrm{t}}_{c^*,l,t} \right) \right\} - \rho_t \right\|. \quad (8)$$

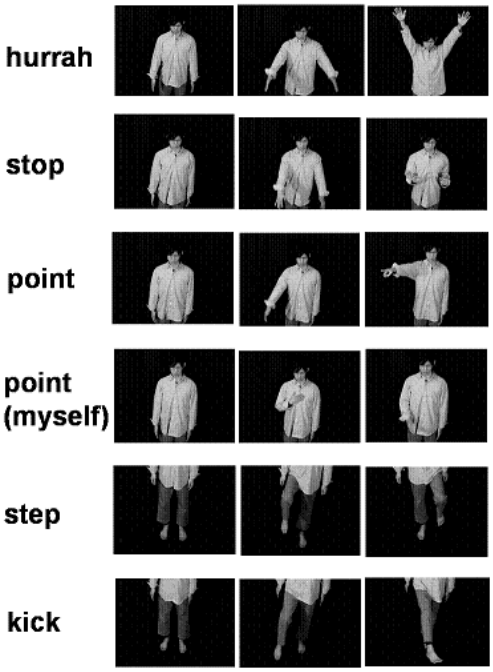

Figure 3 Snapshot of six gestures.

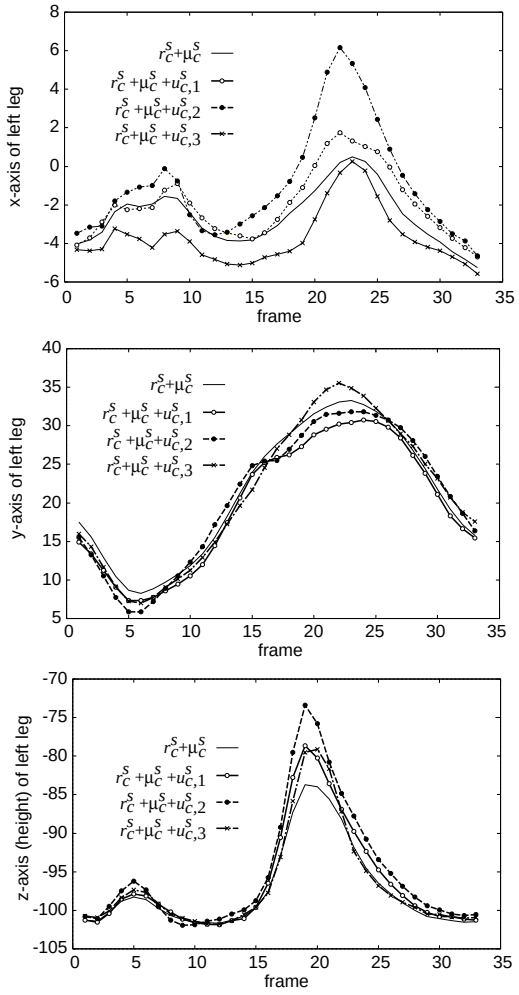

Figure 4 Spatial eigen-gestures for "step". Among 12 features, x-axis of left leg (upper), y-axis of left leg(middle) and z-axis of left leg (lower) are plotted.

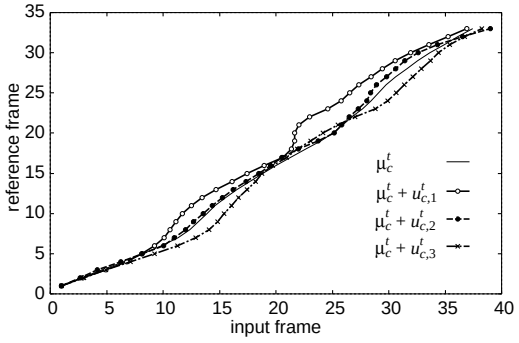

Figure 5 Temporal eigen-gestures for "step".

The minimization of J with respect to $\beta_1, \ldots, \beta_L$ is also a simple least square error problem.

5. Experimental results

Motion prediction experiments were conducted using about 600 samples of six gesture categories (i.e., about 100 samples for each category). Figure 3 shows the six categories, "stop", "hurrah", "point", "point myself", "kick", and "step". Among the 100 samples, 50 samples were used as training samples for the estimation of the eigen-gestures. The remaining 50 samples were used as test samples for performance evaluation.

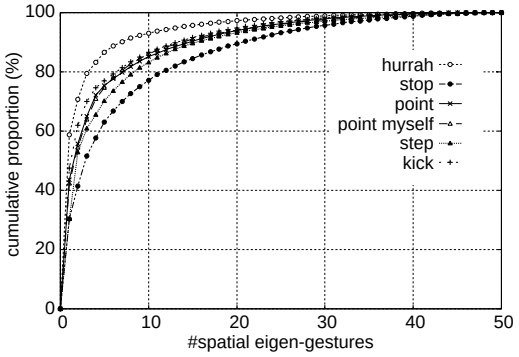

Figure 6 Cumulative proportion of spatial eigen-gestures.

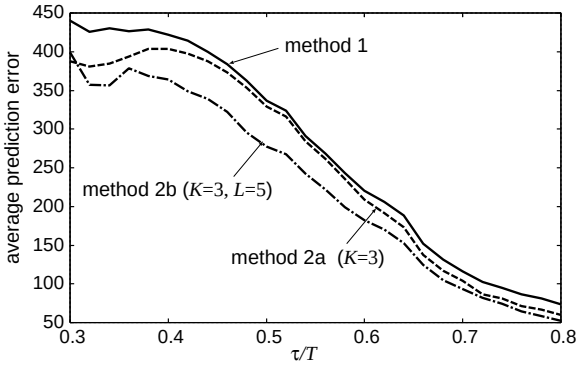

Figure 7 Average prediction error.

Each gesture sample was represented by a sequence of 12-dimensional feature vectors (i.e., $D = 12$). The elements of the feature vector were the relative 3-D position of both hands and feet to the head. The position was acquired by real-time stereo measurement (camera: Sony DFW-X700, 15 frames/s).

5.1 Estimated eigen-gestures

The spatial and the temporal eigen-gestures of each category were estimated from the spatial and the temporal difference vectors, $\{\boldsymbol{v}_{c,1}^s, \ldots, \boldsymbol{v}_{c,50}^s\}$ and $\{\boldsymbol{v}_{c,1}^t, \ldots, \boldsymbol{v}_{c,50}^s\}$, respectively. Those difference vectors were obtained via DP matching between one reference pattern $\boldsymbol{r}_c$ and each of the 50 training samples (i.e., $N = 50$).

Figure 4 shows top three spatial eigen-gestures for "step", which is comprised of one cycle of a step action (that is, each of two legs is lifted once). Among the 12 features, three features, the x-axis of left leg position (i.e., forward direction), the y-axis of left leg position (i.e., right-to-left direction) and the z-axis of left leg position (i.e., height of the left leg) were plotted in this figure. The spatial eigen-gestures for the z-axis of left leg position commonly represent the heigh variation at the peak of leg lifting. Those for the x-axis and y-axis show rather different motions.

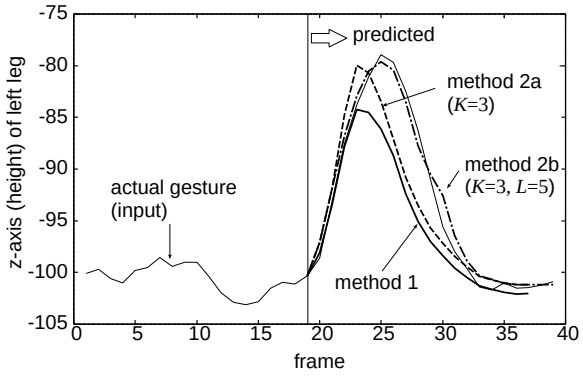

Figure 8 Predicted left leg motions for "step" by three prediction methods.

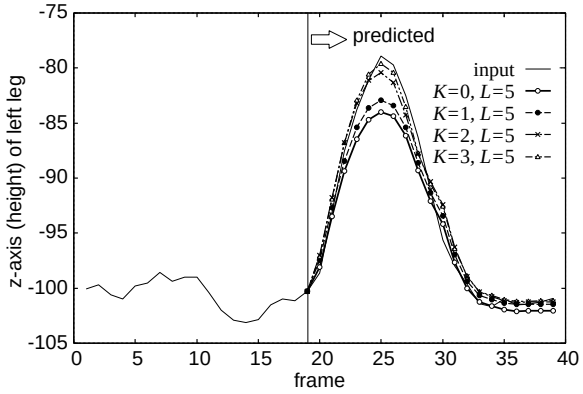

Figure 9 Predicted left leg motions for "step" by Method 2b with different numbers of spatial eigen-gestures.

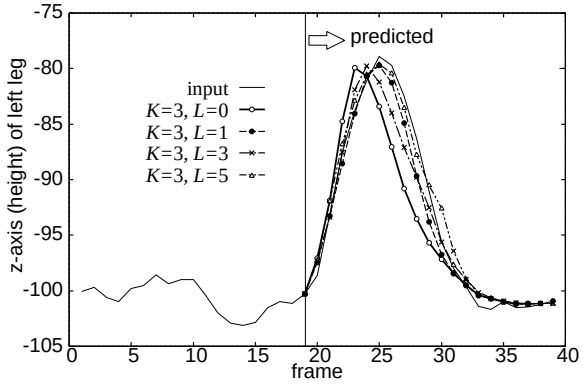

Figure 10 Predicted left leg motions for "step" by Method 2b with different numbers of temporal eigen-gestures.

Figure 5 shows temporal eigen-gestures for "step". The first temporal eigen-gesture explicitly shows the speed change at lifting a leg.

Figure 6 shows the cumulative proportion of the spatial eigen-gestures of each category. Every cumulative proportion shows an early saturation and thus indicates that the spatial variations of gestures have category-specific tendencies.

5.2 Prediction accuracy

The following three prediction methods were compared in terms of prediction accuracy:

Method 1. The prediction based on the simple extrapolation of (1). A mean gesture of the category c was employed as a single reference gesture $\boldsymbol{r}_c$.

Method 2a. The prediction based on (4). That is, only spatial eigen-gestures were used.

Method 2b. The prediction based on (6). That is, both the spatial and the temporal eigen-gestures were used.

Figure 7 shows prediction error averaged in all categories as a function of τ/T, which represents the proportion of the beginning part used for prediction. The prediction error was defined as $\sum_{\delta=0}^{T-\tau} \|\widehat{x}_{\tau+\delta} - x_{\tau+\delta}\|$. Three spatial eigen-gestures were used in Method 2a, whereas five temporal eigen-gestures were also used in Method 2b. As shown in Fig. 7, Method 2b attained the highest accuracy among the three methods at most of τ/T.

5.3 Predicted motion

Figure 8 shows the left leg motions for "step" predicted by Methods 1, 2a and 2b. The proportion τ/T was fixed at 0.5, that is, the latter half of a gesture was predicted from the former half observed. Method 1 has provided a poor prediction result; for example, the predicted peak (i.e., the highest point of the left leg) is different from the actual peak in its height and time (frame). The result by Method 2a shows a good accuracy in its height because the spatial eigen-gestures could represent spatial variation from the reference pattern. The result by Method 2b shows near-perfect accuracy in its time as well as height; the temporal variations were successfully compensated by the temporal eigen-gestures.

Figure 9 shows the prediction results by changing K, i.e., the number of spatial eigen-gestures. As K increases, the spatial difference difference between the predicted motion and the actual input motion is minimized. Figure 10 shows the prediction results by changing L, i.e., the number of temporal eigen-gestures. As L increases, the temporal difference is minimized. Although these figures show only the height of the left leg, positional differences at other parts (hands and right leg) were also minimized successfully by the eigen-gestures.

6. Conclusion

In this paper, two motion prediction methods have been investigated for the realization of "proactive" gesture-based man-machine interaction systems. While the both methods utilize early recognition results of gestures, those methods are different in their prediction models. One method is based on a simple extrapolation model and outputs a part of a reference gesture directly as its prediction result. The other method is based on a more sophisticated model where a reference gesture and eigen-gestures, which

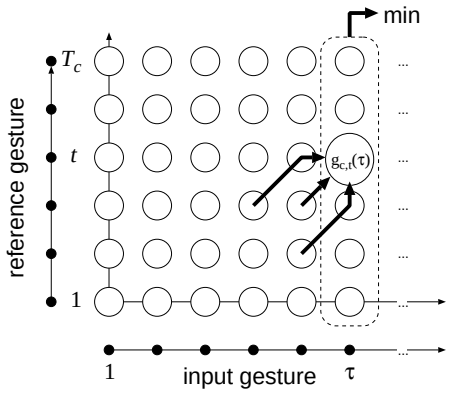

Figure 11 DP matching algorithm.

represent typical variations from the reference gesture, are combined linearly. Experimental results have revealed that the latter method could attain better prediction results because of its ability to deal with various changes in input gestures.

Future work includes the improvement of gesture prediction models. The model based on the linear combination of the eigen-gestures can be extended by some kernel PCA for a more compact representation of the variations of gestures. User-independent/dependent eigen-gestures also should be investigated.

References

[1] Y. Yacoob and M. Black, "Parameterized modeling and recognition of activities," Comput. Vis. Image Und., vol. 73, no. 2, pp. 232–247, 1999.

[2] A. Fod, M. Mataric,and O. C. Jenkins, "Automated derivation of primitives for movement classification," Autonomous Robots, vol. 12, no. 1, pp. 39–54, 2002.

[3] T. Amano, "Image interpolation by high dimensional projection based on subspace method," Proc. ICPR, vol. 4 of 4, pp. 665-668, 2004.

[4] C. Naster, B. Moghaddam, and A. Pentland, "Flexible images: matching and recognition using learned deformations," Comput. Vis. Image Und., vol. 65, no. 2, pp. 179–191, 1997.

[5] T. Darrell and A. Pentland, "Space-time gestures," Proc. CVPR, pp. 335-340, 1993.

[6] S. Seki, K. Takahashi and R. Oka, "Gesture recognition from motion image by spotting algorithm," ACCV1993, vol. 2, pp. 759-762, 1993.

Appendix

1. Early recognition algorithm

The early recognition algorithm is based on DP matching shown in Fig. 11. Specifically, the algorithm provides c^* and t^* by matching the *beginning part* of the reference $r_{c,1}, \ldots, r_{c,T_c}$ to the *beginning part* of the input $x_1, \ldots, x_\tau$. The following pseudo-code illustrates the algorithm.

Step 1: For each input frame $\tau = 1, 2, \ldots 1$, repeat Step 2-4.

Step 2: For each category $c = 1, \ldots, C$, repeat Step 3.

Step 3: For each frame $t = 1, \ldots, T_c$, calculate the following DP-recurrence equation:

$$g_{c,t}(\tau) = \min \begin{cases} g_{c,t-1}(\tau-1) + 3d_{c,t}(\tau) & \text{(a)} \\ g_{c,t-1}(\tau-2) + 2d_{c,t}(\tau-1) + d_{c,t}(\tau) & \text{(b)} \\ g_{c,t-2}(\tau-1) + 3d_{c,t-1}(\tau) + 3d_{c,t}(\tau) & \text{(c)} \end{cases}, \quad \text{(A·1)}$$

where $d_{c,t}(\tau)$ represents local distance between x_τ and $r_{c,t}$ i.e.,

$$d_{c,t}(\tau) = \|x_\tau - r_{c,t}\|. \quad \text{(A·2)}$$

The value $g_{c,t}(\tau)$ becomes the minimum accumulated distance (i.e., matching cost) up to the node at c, t, and τ. In (A·1), (a), (b), and (c) correspond to the path (a), (b), and (c) in Fig. 11, respectively.

Step 4: The early recognition result at the frame τ is provided as

$$(c^*, t^*) = \underset{c,t}{\operatorname{argmin}} \left(g_{c,t}(\tau)/t \right). \quad \text{(A·3)}$$

Only the discrimination rule of Step 4 distinguishes the above early recognition algorithm from conventional DP-based recognition algorithms. Comparing to the conventional rule

$$c^* = \underset{c}{\operatorname{argmin}}\, g_{c,T_c}(\tau),$$

the early recognition rule (A·3) allows the beginning part of reference pattern $\boldsymbol{r}_{c,1}, \boldsymbol{r}_{c,2}, \ldots, \boldsymbol{r}_{c,t}$ $(t < T_c)$ to be a matching candidate. In consequence, the discrimination rule (A·3) provides recognition results before the end of an entire gesture. Note that the path in Fig. 11 shows the nonlinear frame alignment ρ_t, which is used to provide the spatial and the temporal eigen-gestures.