Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A Hilbert warping method for handwriting gesture recognition

Hiroyuki Ishida [a,*], Tomokazu Takahashi [b], Ichiro Ide [a], Hiroshi Murase [a]

[a] Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan
[b] Department of Economics and Information, Gifu Shotoku Gakuen University, Nakauzura 1-38, Gifu-shi, Gifu 501-6194, Japan

ARTICLE INFO

Article history:
Received 24 February 2009
Received in revised form
24 February 2010
Accepted 25 February 2010

Keywords:
Sequence alignment
Classification method
Analytic signal
Gesture recognition

ABSTRACT

We propose a novel sequence alignment algorithm for recognizing handwriting gestures by a camera. In the proposed method, an input image sequence is aligned to the reference sequences by phase-synchronization of analytic signals which are transformed from original feature values. A cumulative distance is calculated simultaneously with the alignment process, and then used for the classification. A major benefit of this method is that over-fitting to sequences of incorrect categories is restricted. The proposed method exhibited higher recognition accuracy in handwriting gesture recognition, compared with the conventional dynamic time warping method which explores optimal alignment results for all categories.

1. Introduction

Camera-based analysis of human behavior has been studied for decades [1]. Specifically, there has been a great interest in realizing hand gesture recognition by a computer [2]. One of its applications is handwriting gesture recognition system in which characters written by a finger are identified as shown in Fig. 1. It has gained attention as a novel means of man-machine interaction [3] because: (I) users can operate computers just by simple gestures and (II) it does not require extra equipments except for a camera. The handwriting gesture recognition system could be used as an input device for PDA or ATM machines.

There are two major approaches to handwriting gesture recognition: hidden Markov model (HMM) [4] and dynamic time warping (DTW) [5]. In the HMM-based methods [6–8], the motion of the fingertip is used as input to the HMM network. The direction of the fingertip motion is encoded into a two-dimensional chain code and treated as an observation symbol of the HMM network. In this framework, tracking of the fingertip location is essential. Currently, there are several problems with this framework. The first is the failure of tracking. It can be caused when the fingertip is occluded by the palm. The second is the costs of the fingertip detection. It is hard to accomplish a real-time recognition, if great efforts and costs are required to process each frame. An alternative way to extract features is the eigenspace method [9] which does not require tracking nor image analysis. Unfortunately, there is no straightforward way to employ the eigenspace method in the HMM framework because observation symbols such as the motion direction are not given.

On the other hand, the DTW provides more flexible solutions in terms of feature extraction. In the methods [10–12] based on the DTW, feature vectors such as hand location are compared to those of the reference sets based on the distance calculation. Distances between feature vectors are calculated, followed by a cumulative distance between sequences; an input sequence is classified to a reference sequence which gives the minimum cumulative distance. More recently, it has been proven that the combination of the DTW and the eigenspace method makes a powerful tool for the image sequence recognition [13]. It not only solves the problems of the fingertip detection, but also achieves the robustness against image noise by using principal components as the feature values.

However, the DTW has a drawback for the classification task. The DTW searches the optimal alignment for the reference sequences of all categories, which often causes misclassification to incorrect categories due to over-fitting. To cope with this problem, we propose a "Hilbert warping" method which searches the proper alignment only for a correct category. In the proposed method, sequences are converted into the form of analytic signals [14]. An important property of the analytic signal is that its instantaneous phase increases monotonically. Using this property, both of the sequences are aligned by phase-synchronization of their analytic signals. Undesirable over-fitting to incorrect categories is restricted if the sequence alignment is performed by the phase-synchronization.

In this paper, we apply the proposed method to camera-based recognition of handwriting gestures. Fig. 2 shows the flow of the proposed method. Firstly, image sequences are converted to a series of feature vectors by the eigenspace method [9]. As

[*] Corresponding author.
E-mail address: hishi@murase.m.is.nagoya-u.ac.jp (H. Ishida).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.02.021

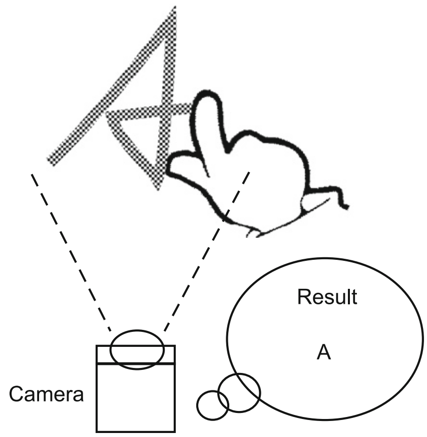


Fig. 1. Example of a handwriting gesture recognition system using a camera.

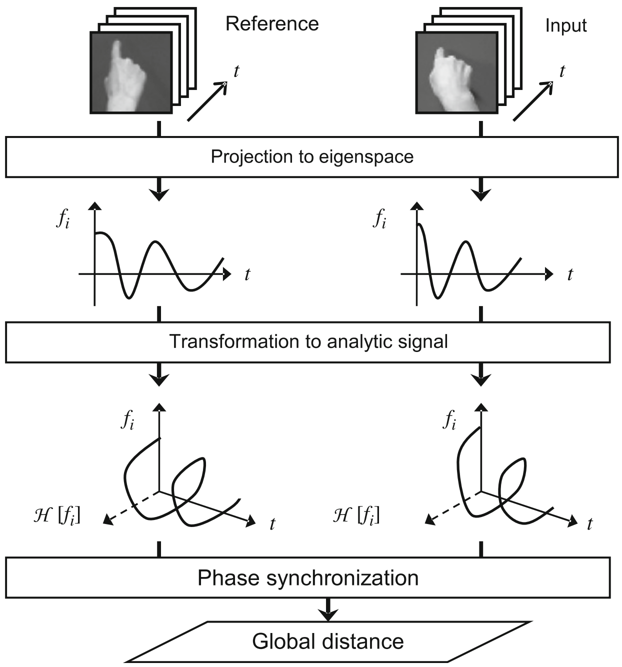


Fig. 2. Proposed Hilbert warping method for handwriting gesture recognition. The alignment is performed by synchronizing the phases of the spiral-shaped trajectories.

proposed in a gesture recognition method [15], we adopt an appearance-based approach in order to avoid the detection error of a fingertip. Secondly, each feature value is transformed to an analytic signal. The empirical mode decomposition (EMD) [16] is introduced in the training step, to ensure that the phase of the analytic signal should become monotonic. Finally, the cumulative distance between two sequences are calculated by synchronizing the phase of the analytic signals.

Up to now we have presented recognition methods using the Hilbert warping in [17,18]. However, in [17], we did not present the application to the handwriting gesture classification task. Meanwhile, in [18] we did not present a training process using the EMD. This paper presents the application of the Hilbert warping algorithm to the handwriting gesture recognition task and the EMD-based training algorithm.

This paper is organized as follows: Section 2 introduces the property of analytic signals. In Section 3, the proposed Hilbert warping method is described. Results are presented in Section 4.

2. Analytic signal

Analytic signal is often used in the domain of signal processing, mainly for the analysis of temporal properties of signals. This section describes its properties which are useful for the sequence alignment, together with the transformation process.

An image sequence is transformed to analytic signals [14] for the sequence alignment. Let $f(t)$ be a feature value obtained from the t-th image in the sequence. An analytic signal $a(t)$ is composed of the original signal $f(t)$ as the real part and its Hilbert transform $\mathcal{H}[f(t)]$ as the imaginary part [19]. It is denoted as

$$a(t)=f(t)+j\mathcal{H}[f(t)]=|a(t)|\exp[j\varphi(t)], \tag{1}$$

where its argument $\varphi(t)$ is defined as the instantaneous phase given by

$$\varphi(t)=\arctan\frac{\mathcal{H}[f(t)]}{f(t)}. \tag{2}$$

In principle, $\varphi(t)$ increases monotonically, which means that $a(t)$ rotates counter-clockwise in the complex plane as illustrated in Fig. 3. This is an important property for the alignment of different sequences because frames with the same value of $\varphi(t)$ can be aligned simply by phase-synchronization. The instantaneous phase $\varphi(t)$ generally satisfies the monotonicity, since the spectrum of the analytic signal contains only non-negative frequency components. The reason why negative frequency components are removed is shown by the following formulae: Let $F(\omega)=\mathcal{F}[f(t)]$ be the Fourier transform of $f(t)$, and the Hilbert transform be written in terms of the convolution notation as

$$\mathcal{H}[f(t)]=\frac{1}{\pi t}*f(t)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f(\tau)}{t-\tau}\,d\tau, \tag{3}$$

then the analytic signal $a(t)$ in frequency domain is represented as

$$\begin{aligned}\mathcal{F}[a(t)]&=\mathcal{F}[f(t)+j\mathcal{H}[f(t)]]=\mathcal{F}[f(t)]+j\mathcal{F}[\mathcal{H}[f(t)]]\\&=\mathcal{F}[f(t)]+j\mathcal{F}[1/\pi t]\mathcal{F}[f(t)]=F(\omega)+j(-j)\mathrm{sgn}(\omega)F(\omega)\\&=F(\omega)[1+\mathrm{sgn}(\omega)]=\begin{cases}2F(\omega) & (\omega>0)\\ F(\omega) & (\omega=0)\\ 0 & (\omega<0).\end{cases}\end{aligned} \tag{4}$$

Now we can see that $\mathcal{F}[a(t)]$ does not contain negative frequency components.

This equation indicates also that analytic signals are calculated simply via the Fourier transforms [20]. Since Eq. (3) involves integration over an infinite range of τ, it is more practical to

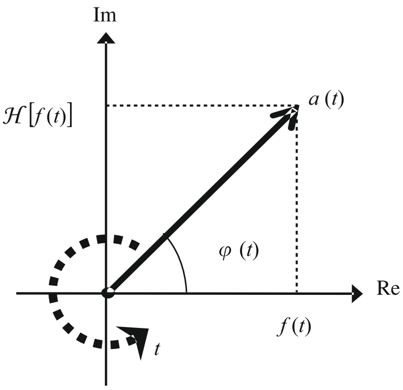


Fig. 3. Construction of analytic signal. $\mathcal{H}[f(t)]$ is the Hilbert transform of $f(t)$. In principle, instantaneous phase $\varphi(t)$ increases monotonically.

obtain the analytic signal $a(t)$ from Eq. (4). In practice, analytic signals are obtained by the following way:

1. Calculate the Fourier transform of $f(x)$.
2. Double the positive frequency components $F(\omega > 0)$, and remove the negative frequency components $F(\omega < 0)$.
3. Calculate the inverse Fourier transform.

3. Hilbert warping method

The method for the sequence classification is described in this section. The usefulness of the analytic signal for matching two warped signals is indicated in [21]. However, practical alignment algorithms for classification tasks have not been presented. Since the algorithm in [21] used a single-dimensional signal without removing noise components, its performance was not satisfactory for complicated signals. The problem of over-fitting to different categories was not discussed either. We focused on this problem and presented a video-sequence classification method using the analytic signals in [17]. However, it was applicable only to the camera-based printed character recognition. In order to apply it to various types of video sequences, we introduced the EMD which has an ability to analyze complicated signals. Although this idea is presented in [18], the method for separating undesirable components from the analytic signals has not been presented. This paper proposes a new Hilbert warping method using a multi-dimensional feature space. Also, a technique to use the EMD in the Hilbert warping is presented. The proposed method ensures proper alignment for a correct category, but avoids over-fitting to incorrect categories.

This section is organized as follows: Feature vectors used for the recognition is introduced in Section 3.1. The definition of the phase-difference used in the alignment process is detailed in Section 3.2. The proposed Hilbert warping algorithm is presented in Section 3.3. An alternative definition of the phase-difference using the EMD is introduced in Section 3.4.

3.1. Feature vector

The proposed method uses feature vectors for the calculation of a distance between sequences. Low-dimensional feature vectors are obtained from images using the eigenspace method [9]. Initially, a mean vector $\boldsymbol{\mu}$ and an R-dimensional eigenspace $\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_R\}$ are constructed from all reference images. Let the t-th image in a sequence be represented by a normalized vector $\boldsymbol{x}(t)$. It is projected on the eigenspace as a point $\boldsymbol{g}(t)$ by

$$\boldsymbol{g}(t) = [\boldsymbol{e}_1 \quad \cdots \quad \boldsymbol{e}_R]^\top(\boldsymbol{x}(t) - \boldsymbol{\mu}) \tag{5}$$

$$= [f_1(t) \quad \cdots \quad f_R(t)]^\top. \tag{6}$$

Then the image sequence is transformed to a trajectory of points in the eigenspace as shown in Fig. 4. These $f_i(t)$ $(1 \le i \le R)$ are used as the feature values for the sequence alignment.

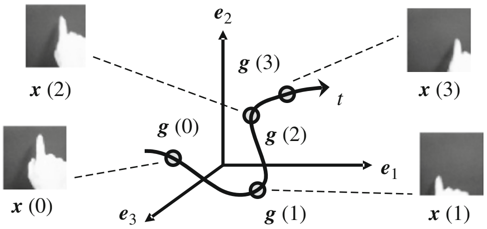


Fig. 4. Trajectory of feature vectors in eigenspace.

3.2. Calculation of phase-difference

The proposed method requires a definition of the phase-difference between an input frame and a reference frame. In order to evaluate instantaneous phases, feature values need to be transformed to analytic signals.

A feature vector $\boldsymbol{g}(t)$ is converted to an analytic signal vector (ASV) by transforming each element $f_i(t)$ to an analytic signal using Eq. (1). Let $\boldsymbol{\alpha}^{(c)}(t)$ be a reference ASV of category c, and $\boldsymbol{\alpha}^{in}(t)$ be an input ASV. These ASVs are denoted using reference analytic signals $a_i^{(c)}(t)$ and input analytic signals $a_i^{in}(t)$ by

$$\boldsymbol{\alpha}^{(c)}(t) = [a_1^{(c)}(t) \quad \cdots \quad a_R^{(c)}(t)]^\top, \tag{7}$$

$$\boldsymbol{\alpha}^{in}(t) = [a_1^{in}(t) \quad \cdots \quad a_R^{in}(t)]^\top. \tag{8}$$

The proposed method evaluates phase-difference from the argument ($\angle$) of the Hermitian inner product $p^{(c)}(t_1,t_2)$ given by

$$p^{(c)}(t_1,t_2) = [\boldsymbol{\alpha}^{(c)}(t_1)]^* \boldsymbol{\alpha}^{in}(t_2), \tag{9}$$

where the superscript * denotes the complex conjugate transpose of a vector, and the variables t_1 and t_2 represent the frame indices of the reference sequence and the input sequence, respectively.

An attractive property of the Hermitian inner product is that it measures which instantaneous phases of the ASVs are more advanced. We shall show this using Eqs. (1) and (9). First, let us consider the case $R=1$, where the signal is one-dimensional. $\angle p^{(c)}(t_1,t_2)$ is identical to the phase-difference between two analytic signals because

$$p^{(c)}(t_1,t_2)_{R=1} = \overline{a_1^{(c)}(t_1)} a_1^{in}(t_2) = |a_1^{(c)}(t_1)||a_1^{in}(t_2)|\exp[j\{\varphi_1^{in}(t_2) - \varphi_1^{(c)}(t_1)\}], \tag{10}$$

$$\angle p^{(c)}(t_1,t_2)_{R=1} = \varphi_1^{in}(t_2) - \varphi_1^{(c)}(t_1). \tag{11}$$

The sign of the phase-difference indicates which instantaneous phase is advanced; thus we increment t_1 if the phase-difference is positive, or decrement t_1 if negative. This procedure drives $\angle p^{(c)}(t_1,t_2)$ closer to zero, at which two analytic signals have the same instantaneous phase values. This is why the analytic signal is useful for matching warped signals. In the cases $R>1$, $p^{(c)}(t_1,t_2)$ indicates the weighted sum of complex-valued phase-differences of each principal component. It is given by

$$p^{(c)}(t_1,t_2) = \sum_{r=1}^{R} \overline{a_r^{(c)}(t_1)} a_r^{in}(t_2)$$
$$= \sum_{r=1}^{R} |a_r^{(c)}(t_1)||a_r^{in}(t_2)|\exp[j\{\varphi_r^{in}(t_2) - \varphi_r^{(c)}(t_1)\}]. \tag{12}$$

By averaging $\overline{a_r^{(c)}(t_1)} a_r^{in}(t_2)$ over the principal components $(1 \le r \le R)$, it is possible to estimate which phase as a whole is advanced or delayed. Based on this property, the frame index t_1 is updated by

$$t_1 \leftarrow t_1 + \operatorname{sgn} \angle p^{(c)}(t_1,t_2), \tag{13}$$

$$\operatorname{sgn} \angle p^{(c)}(t_1,t_2) = \begin{cases} +1 & (\angle p^{(c)}(t_1,t_2) > 0), \\ -1 & (\angle p^{(c)}(t_1,t_2) < 0), \end{cases} \tag{14}$$

because it is expected that

$$|\angle p^{(c)}(t_1 + \operatorname{sgn} \angle p^{(c)}(t_1,t_2), t_2)| < |\angle p^{(c)}(t_1,t_2)|. \tag{15}$$

In the alignment stage, frame t_1 which corresponds to frame t_2 is sequentially searched according to the sign of $\angle p^{(c)}(t_1,t_2)$.

Table 1
Hilbert warping algorithm for calculating the cumulative distance $D^{(c)}$ to category c.

Hilbert warping algorithm

```
/* Initialization */
$D^{(c)} \leftarrow 0,\ t_1[1] \leftarrow 1,\ t_2 \leftarrow 1,\ i \leftarrow 1$
do
  do
    calculate $p^{(c)}(t_1[i],t_2)$
    /* Search by the sign of the phase-difference */
    $t_1[i+1] \leftarrow t_1[i] + \mathrm{sgn}\angle p^{(c)}(t_1[i],t_2)$
    $i \leftarrow i+1$
  until sign of $\angle p^{(c)}(t_1[i],t_2)$ changes
  /* Distance $d^{(c)}(t_1,t_2)$ is calculated */
  $D^{(c)} \leftarrow D^{(c)} + \min_i d^{(c)}(t_1[i],t_2)$
  $t_1[1] \leftarrow \mathrm{argmin}_{t_1[i]} d^{(c)}(t_1[i],t_2)$
  $t_2 \leftarrow t_2+1,\ i \leftarrow 1$
until $t_2$ reaches the last frame
return $D^{(c)}$
```

3.3. Hilbert warping algorithm

The proposed method associates a reference sequence $(1 \le t_1 \le T_1)$ with an input sequence $(1 \le t_2 \le T_2)$ by the algorithm in Table 1. This algorithm explores the correspondence of frames as illustrated in Fig. 6. The pair of aligned frames are obtained by tracing the node (t_1,t_2) where $\angle p^{(c)}(t_1,t_2) \approx 0$, and simultaneously the cumulative distance $D^{(c)}$ is computed. In this algorithm, the frame-to-frame distance $d^{(c)}(t_1,t_2)$ is defined as a distance between ASVs by

$$d^{(c)}(t_1,t_2) = \sqrt{\|\boldsymbol{\alpha}^{(c)}(t_1) - \boldsymbol{\alpha}^{in}(t_2)\|^2}. \tag{16}$$

Finally, the input sequence is classified to

$$\hat{c} = \arg\min_c \left(D^{(c)} + \sum_{t_1=1}^{t_1'-1} d^{(c)}(t_1,1) + \sum_{t_1=t_1''+1}^{T_1} d^{(c)}(t_1,T_2) \right), \tag{17}$$

where t_1' and t_1'' are the frame indices of $\boldsymbol{\alpha}^{(c)}(t_1)$ which are aligned to $\boldsymbol{\alpha}^{in}(t_2=1)$ and $\boldsymbol{\alpha}^{in}(t_2=T_2)$ by the algorithm, respectively. The last two terms in Eq. (17) are introduced for penalizing a path in which the input sequence is aligned to only a short part of the reference sequence. The searched path ($\angle p^{(c)}(t_1,t_2) \approx 0$) does not coincide with the path giving the minimal $D^{(c)}$ if the two sequences cannot be aligned consistently. Consequently, this method avoids over-fitting to incorrect categories.

3.4. Calculation of phase-difference using EMD

The phase-synchronization by Eq. (9) is effective for the alignment only if the phase increases monotonically. Unfortunately, such requirement is not satisfied unless the original $f_i(t)$ has zero-crossing points between the local maxima and the local minima [22]. For example, an analytic signal generated from $f_i(t)$ in Fig. 5(a) has local loops which do not enclose the origin as seen in Fig. 5(b). The loops often occur at the beginning and the end of strokes where the feature values tend to be a local maxima or a local minima. If such loops exist, phases cannot be synchronized because of the lack of monotonicity. In order to eliminate these loops, we apply the EMD[1] to decompose $f_i\ (t)$ of the reference sequences to oscillation functions which are called "intrinsic mode functions (IMFs)" (Figs. 5(c) and (d)). The EMD is applied only to the reference sequence in the training step. Some of the IMFs should be excluded during a period where they are considered to make loops. Suppose that $b_i^{(c)}(t)$ is a component which makes loops in the reference ASVs, the following vector should be subtracted.

$$\boldsymbol{\beta}^{(c)}(t) = [b_1^{(c)}(t) \ \ \cdots \ \ b_R^{(c)}(t)]^\top. \tag{18}$$

This vector is calculated from the IMFs previously in the training step. In the alignment step, $\boldsymbol{\beta}^{(c)}(t)$ is excluded both from reference ASVs $\boldsymbol{\alpha}^{(c)}(t)$ and input ASVs $\boldsymbol{\alpha}^{in}(t)$. Accordingly, the right side of Eq. (9) is modified as

$$[\boldsymbol{\alpha}^{(c)}(t_1) - \boldsymbol{\beta}^{(c)}(t_1)]^*[\boldsymbol{\alpha}^{in}(t_2) - \boldsymbol{\beta}^{(c)}(t_1)]. \tag{19}$$

This modified phase-difference is useful for synchronizing the phases, since $\boldsymbol{\beta}^{(c)}(t_1)$ is determined such that the remaining component $\boldsymbol{\alpha}^{(c)}(t_1) - \boldsymbol{\beta}^{(c)}(t_1)$ satisfies the monotonicity. Fig. 7 shows an example of the analytic signal trajectory after subtracting $b_i^{(c)}(t)$, where we can see that the loops in the original trajectory have been eliminated successfully. Algorithms for the EMD and for the calculation of $b_i^{(c)}(t)$ are described below. Both of them are employed in the training step.

The algorithm of the EMD proposed by Huang et al. [24] is outlined in Table 2. It decomposes a signal $f_i(t)$ to IMFs $s_{(i,m)}(t)$ $(1 \le m \le M)$ and a residual $r_i(t)$ as

$$f_i(t) \rightarrow \sum_{m=1}^{M} s_{(i,m)}(t) + r_i(t). \tag{20}$$

The value M represents the number of IMFs. It depends on the shapes of the original signal $f_i(t)$, especially on the number of local maxima and minima. In general, IMFs $s_{(i,m)}(t)$ with large m consist of low frequency components.

The component $b_i^{(c)}(t)$ used for the Hilbert warping in Eq. (19) is determined as shown in Table 3. This algorithm extracts the component $b_i^{(c)}(t)$ from IMFs if the phase-difference $|\angle z|$ between the IMF and the remaining component is large.

4. Experimental results

Experiments were conducted using hand-writing gesture datasets[2] (Fig. 8) which consists of ten datasets written by ten persons individually. Each dataset contains 26 image sequences (48 × 48 pixels, 30 frames per second) of finger-writing characters (uppercase A–Z). Recognition rates were evaluated by leave-one-out cross-validation; all the sequences except for an input dataset were used as references. Even though some characters in the datasets contained several patterns of writing order, it was possible to obtain a correct result as long as the reference datasets contained characters written in the same order. The classification was based on the nearest neighbor rule (1-NN).

The performance of the proposed method (HW+EMD) was compared with the DTW. The cumulative distance $D^{(c)}\ (T_1,T_2)$ of the DTW was calculated by

$$D^{(c)}(0,0) = 0, \tag{21}$$

$$D^{(c)}(t_1,t_2) = \min_k\{D^{(c)}(t_1-k,t_2-1)\} + d^{(c)}(t_1,t_2)(0 < t_1 \le T_1, 0 < t_2 \le T_2), \tag{22}$$

[1] We developed a library hht.h for using the EMD and Hilbert transform which is distributed as part of MIST libraries [23].

[2] Since a widely used dataset for hand-writing gesture recognition was not available, we constructed the datasets and used them in this experiment. The datasets can be downloaded freely for evaluation from http://www.murase.m.is.nagoya-u.ac.jp/~hishi/finger-writing.html.

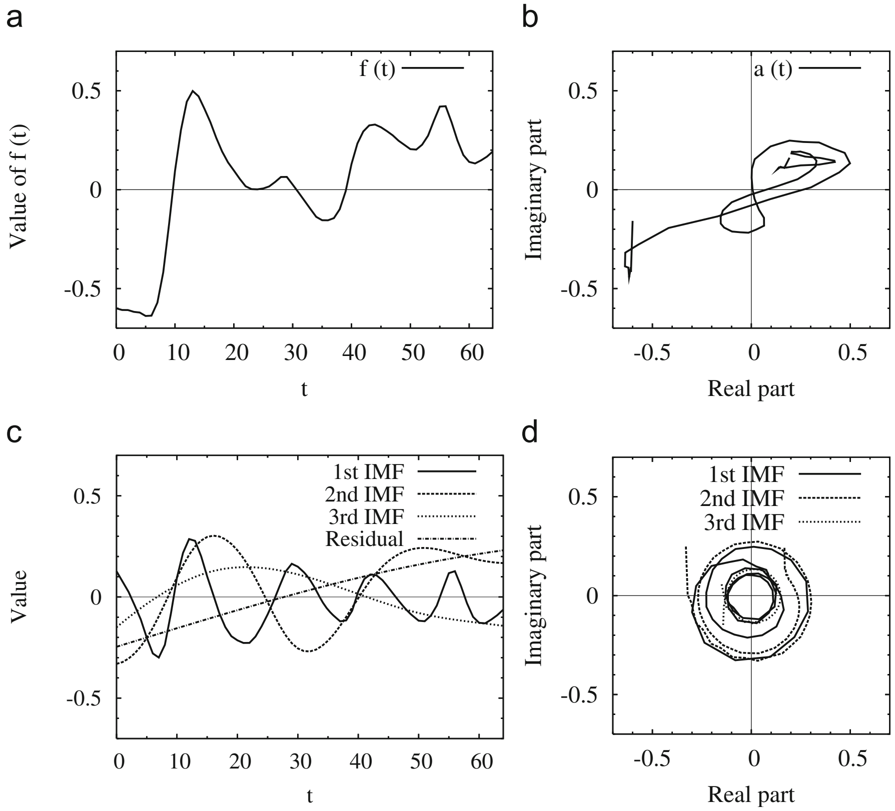


Fig. 5. Examples of analytic signals: (a) $f(t)$, (b) analytic signal of (a), (c) IMFs of (a), and (d) analytic signals of (c).

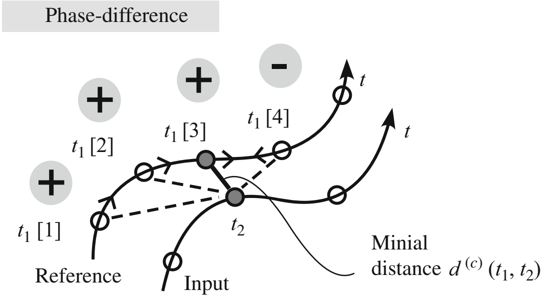


Fig. 6. Phase-synchronization process for sequence alignment.

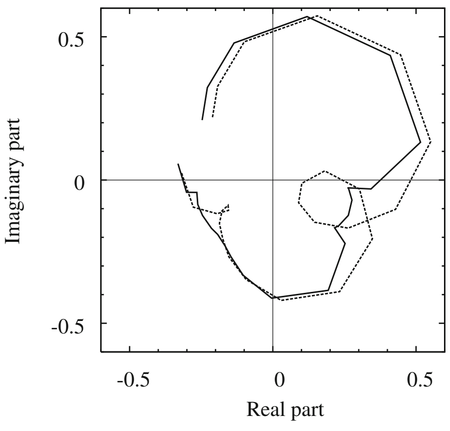


Fig. 7. Elimination of loops in the analytic signal trajectory. The dotted curve represents the original trajectory $a_i^{(c)}(t)$. The solid curve represents the trajectory $a_i^{(c)}(t)-b_i^{(c)}(t)$ obtained by subtracting the loop component $b_i^{(c)}(t)$.

Table 2
The algorithm of the EMD.

Empirical mode decomposition.

```
    /* Initialization */
1   $r_i(t) \leftarrow f_i(t)$, $m \leftarrow 1$
2   do
3     $h_1(t) \leftarrow r_i(t)$, $n \leftarrow 1$
      /* Calculation of the m-th IMF $s_{(i,m)}(t)$ */
4     do
5       Extract all the local maxima and minima of $h_i(t)$
6       Calculate an upper envelope $u_n(t)$ and a lower envelope $v_n(t)$ by interpolating the local maxima and minima, respectively, using a cubic spline
7       $h_{n+1}(t) \leftarrow h_n(t)-[u_n(t)+v_n(t)]/2$    // subtract the mean
8       $n \leftarrow n+1$
9     until the following stopping criterion is satisfied
        $\sum_t[|h_{n-1}(t)-h_n(t)|^2/h_{n-1}(t)^2] \leq \varepsilon \approx 0.3$
10    $s_{(i,m)}(t) \leftarrow h_n(t)$    // m-th IMF
11    $r_i(t) \leftarrow r_i(t)-s_{(i,m)}(t)$    // update the residual
12    $m \leftarrow m+1$
13  until the number of extrema of $r_i(t)$ becomes zero
14  $M \leftarrow m-1$    // the number of IMFs
15  return $s_{(i,m)}(t)$, $r_i(t)$, $M$
```

According to Huang [16], it is desirable to set the threshold ε of the stopping criterion (in Line 7) to around 0.3.

where $d^{(c)}(t_1,t_2)$ used for the DTW was the Euclidean distance in the eigenspace. The following two types of slope constraints $\{k\}$ were tested:

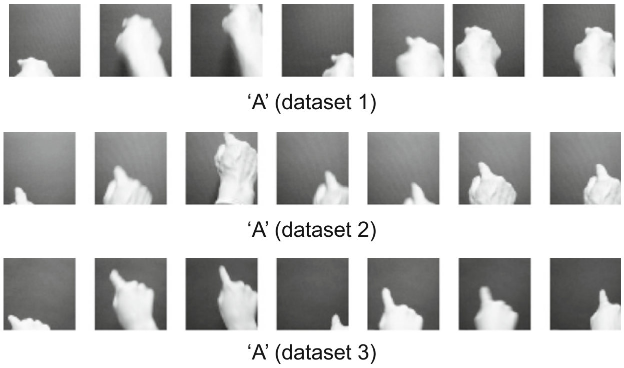
- Type 1: k=0,1,2.This constraint is used generally for recognition tasks using the DTW.
- Type 2: $k=-1$,0,1,2. This less-constrained version of the DTW is effective for gesture recognition [11].

Table 3
Algorithm for extracting a component of loops.

```
Calculation of $b_i^{(c)}(t)$

/* Initialization */
$b_i^{(c)}(t) \leftarrow 0$
for $m = M-1, M-2, \ldots, 1$    // from low frequency component to high frequency component
    $y \leftarrow \left[\sum_{l=m+1}^{M} a_{(i,l)}(t) - b_i^{(c)}(t)\right]$    // sum of IMFs used for the alignment
    $z \leftarrow \overline{a_{(i,m)}(t)}\, y$    // calculation of phase difference
    if $|a_{(i,m)}(t)| < |y|$    // the $m$-th component makes a loop
        $b_i^{(c)}(t) \leftarrow b_i^{(c)}(t) + \frac{1-\cos \angle z}{2} a_{(i,m)}(t)$
next
$b_i^{(c)}(t) \leftarrow b_i^{(c)}(t) + r_i(t)$
return $b_i^{(c)}(t)$
```

The analytic signal of the IMFs $s_{(i,m)}(t)$ is denoted by $a_{(i,m)}(t)$. The number of IMFs determined by the EMD algorithm is denoted by M.


'A' (dataset 1)

'A' (dataset 2)

'A' (dataset 3)

Fig. 8. Example of images in the datasets.

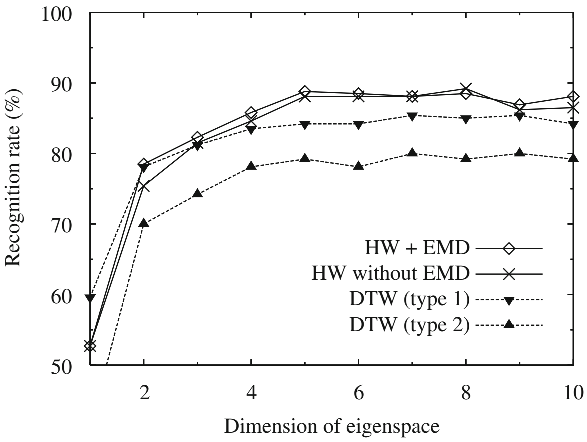
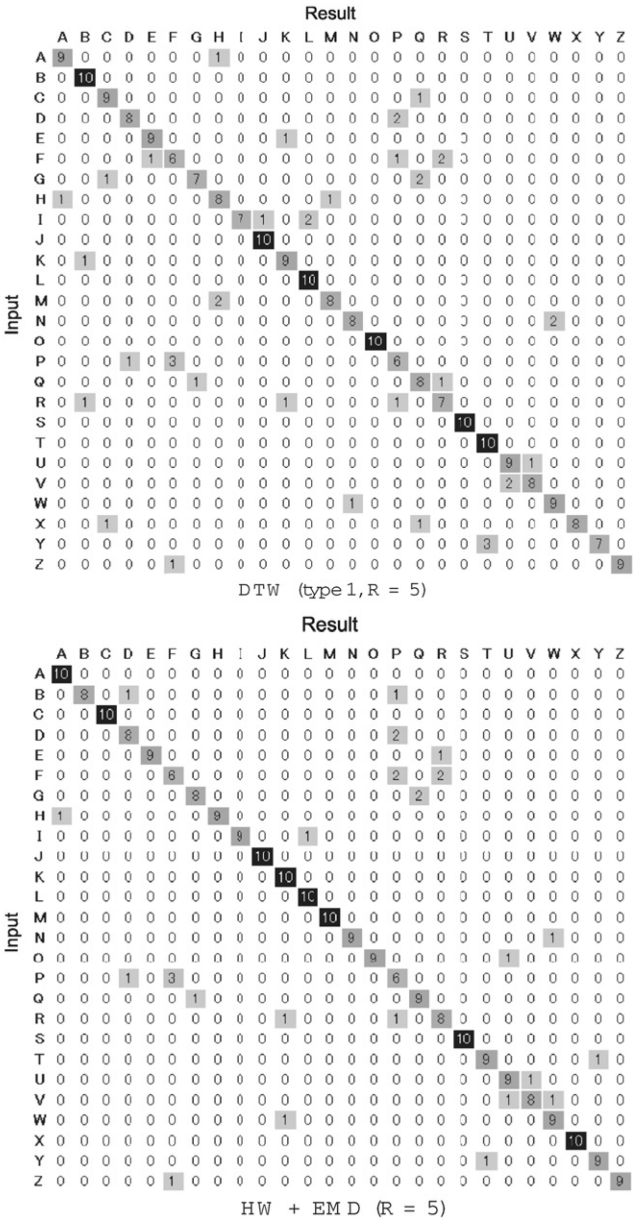


Fig. 9. Recognition rates of finger-writing characters.

The proposed method was compared also to the simple Hilbert warping method without EMD (HW without EMD), which substitutes Eq. (9) for Eq. (19). The HW without EMD is identical to the method presented in [17] with respect to the alignment algorithm.

4.1. Main results

Fig. 9 shows the recognition rates for various dimensions. The horizontal axis of the graph represents the dimension R of the eigenspace. According to the results, the proposed method outperformed the DTW regardless to the type of the slope constraint when $R > 2$, despite that the performance of the sequence alignment was not improved. This is because the over-fitting to incorrect categories was restricted.


Result

Input	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	9	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
D	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	9	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	1	6	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0
G	0	0	1	0	0	0	7	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
H	1	0	0	0	0	0	0	8	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	7	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	1	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	2	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	2	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	8	1	0	0	0	0	0	0	0	0
R	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	7	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	1	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	8	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	9	0	0	0
X	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	8	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	7	0
Z	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9

DTW (type 1, R = 5)

Result

Input	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	8	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
C	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
F	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
H	1	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	1	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	1	0	0	0	0	0
P	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	8	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	1	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	1	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8	1	0	0	0
W	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	9	0
Z	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9

HW + EMD (R = 5)

Fig. 10. Confusion matrices.

Fig. 10 shows confusion matrices of DTW (type 1) and HW+EMD. Although mis-classifications due to untrained writing patterns could not be corrected by any of the methods, classification performance among some categories was improved by using the HW, For example, categories 'H' and 'M' were distinguished more properly by the proposed method. Unlike the DTW, the proposed method avoided the mis-classification to category 'M'. Fig. 11 presents some of the distance matrices $d^{(c)}$ (t_1,t_2) for recognizing category 'H' in dataset 1. In this case, the DTW incorrectly recognized the input 'H' as 'M', since the optimal alignment path to a reference 'M' exhibited the smallest cumulative distance. On the other hand, the proposed method successfully rejected the incorrect category 'M'. From the lower-right distance matrix of Fig. 11, we can see that the actually searched path was different from the optimal

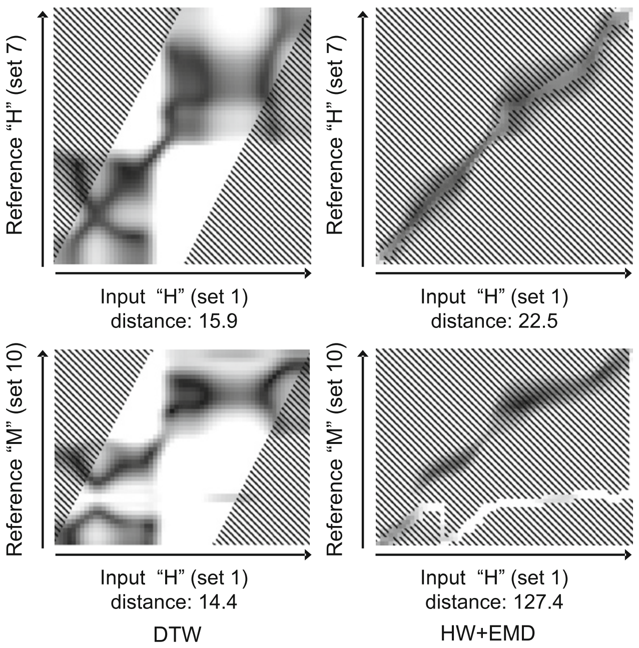


Fig. 11. Example of distance matrices. Values of $d^{(c)}(t_1,t_2)$ are shown by the intensity (0: black). Nodes in the meshed area were not searched.

alignment path indicated by the black-colored nodes. These paths were separated when the finger started writing the horizontal stroke of 'H'. Over-fitting to category 'M' was avoided as a result of the phase-synchronization. Accordingly, it can be stated that the phase-synchronization gave a proper alignment path for the classification.

The results showed that the EMD was effective when $2 \le R \le 6$. This is because the monotonicity of the phase contributes to the proper alignment of sequences especially if the number of available feature values is small. Unlike the case where R was large, it was not effective to counteract the loops of the analytic signals using Eq. (10) in which all the elements of ASV are integrated. Therefore, the use of the EMD is recommended for robust synchronization of phases.

4.2. Comparison with regard to the number of frames

In order to study the performance under various levels of writing speed, the number of frames was changed by three steps. The frame rate of the input sequences was decreased to 20, 15, and 10 fps (frames per second),[3] while the reference frame rate was 30 fps. The resulting recognition rates are shown in Fig. 12, where the DTW (type 3) allowed slopes of $k=3$ in the transition model of Eq. (22). The slope $k=3$ enables to recognize input sequences which are three times as fast as the reference sequences.

The performance of the DTW (type 1) dropped significantly as the number of frames decreased. This is simply because the slope constraint $k \le 2$ was severe. The DTW (type 3) was relatively robust to the lower frame-rate cases, though it was less effective at 30 fps. It is worth noting that the methods using HW yielded higher recognition rates at all frame rates without any given slope constraints. On the other hand, the use of the EMD did not improve the recognition performance at lower frame rates. If the frame rate is low, the step size of the instantaneous phase angle tends to be large. In such case, sparseness of the phase measurement is a greater problem than that caused by the loops of the analytic signals.

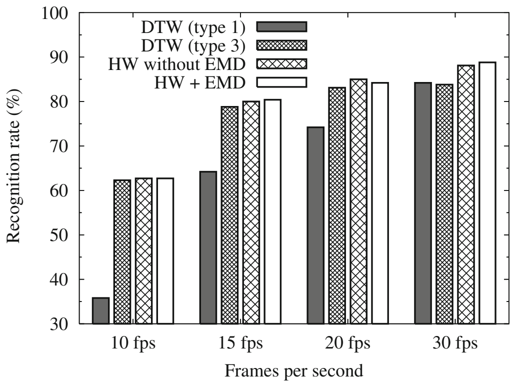


Fig. 12. Recognition rates for various frame rates. The number of dimensions R was 5.

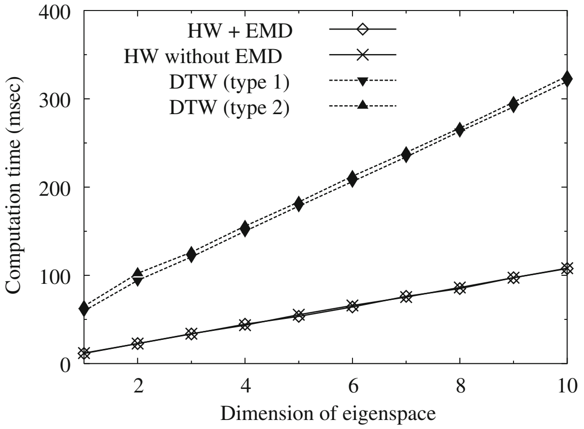


Fig. 13. Average computation time for recognizing one sequence. The experiment was performed on a Pentium IV 3 GHz PC.

4.3. Computational cost

The computation time for recognizing one sequence is shown in Fig. 13, where the results of the Hilbert warping methods include the time required for the Hilbert transformation. The use of the EMD did not affect the computational cost drastically, since it is applied to the reference sequences in the training step. Components $\boldsymbol{\beta}^{(c)}(t)$, which were subtracted from the input sequences in Eq. (18), were calculated from the reference sequences.

The proposed method was approximately three times faster than the conventional DTW, since the calculation of the distance matrices $d^{(c)}(t_1,t_2)$ was drastically reduced. As shown by the meshed area in the example in Fig. 11, the phase-synchronization reduced calculation cycles of the distances $d^{(c)}(t_1,t_2)$. Altogether, the proposed method (HW+EMD) achieved high recognition accuracy with low computational cost.

5. Conclusion

In this paper, a Hilbert warping algorithm for the sequence classification was proposed. The sequence alignment process is based on the phase-synchronization of analytic signals. It is

[3] We used 2/3, 1/2, and 1/3 frames from the original sequences, respectively. In the case of 20 fps, we employed linear interpolation in the eigenspace to obtain the feature vectors.

suitable for classification because incorrect categories are rejected in the alignment process. The experimental results showed a high classification performance of the proposed method for the handwriting gesture recognition task. Future works will involve the automatic detection of human behaviors. For this purpose, the algorithm of the Hilbert warping will be extended for the extraction of video sequences which contain specific gestures.

Acknowledgments

Parts of this research were supported by the Grants-In-Aid for JSPS Fellows (19-6540) and Scientific Research (16300054).

References

[1] D. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding 73 (1) (1999) 82–98.

[2] Y. Wu, T. Huang, Vision-based gesture recognition: a review, in: Lecture Notes in Computer Science, vol. 1739, Springer, Berlin, Heidelberg, 1999, pp. 103–115.

[3] V. Pavlovic, R. Sharma, T. Huang, Visual interpretation of hand gestures for human–computer interaction: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 677–695.

[4] R. Duda, P. Hart, D. Stork, Pattern Classification, second ed., Wiley-Interscience, New York, 2001.

[5] H. Sakoe, S. Chiba, A dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech and Signal Processing 26 (1) (1978) 43–49.

[6] T. Sonoda, Y. Muraoka, A letter input system based on handwriting gestures, Electronics and Communications in Japan, Part III 86 (5) (2006) 53–64.

[7] L. Jin, D. Yang, L. Zhen, J. Huang, A novel vision based finger-writing character recognition system, in: Proceedings of the 18th International Conference on Pattern Recognition, 2006, pp. 1104–1107.

[8] Y. Nam, K. Wohn, Recognition of space–time hand-gestures using hidden Markov model, in: Proceedings of the ACM Symposium on Virtual Reality Software and Technology, 1996, pp. 51–58.

[9] H. Murase, S. Nayar, Three-dimensional object recognition from appearance—parametric eigenspace method, Systems and Computers in Japan 26 (8) (1995) 45–54.

[10] T. Nishimura, R. Oka, Spotting recognition of human gestures from time-varying images, in: Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1998, pp. 318–322.

[11] T. Nishimura, T. Mukai, R. Oka, Non-monotonic continuous dynamic programming for spotting recognition of hesitated gestures from time-varying images, in: Proceedings of the Asian Conference on Computer Vision, 1998, pp. 734–741.

[12] J. Blackburn, E. Ribeiro, Human motion recognition using isomap and dynamic time warping, in: Proceedings of the ICCV Workshop on Human Motion, Lecture Notes in Computer Science, vol. 4814, Springer, Berlin, Heidelberg, 2007, pp. 285–298.

[13] J. Sato, T. Takahashi, I. Ide, H. Murase, Change detection in streetscapes from GPS coordinated omni-directional image sequences, in: Proceedings of the 18th International Conference on Pattern Recognition, 2006, pp. 935–938.

[14] L. Cohen, Time–Frequency Analysis, Prentice Hall Signal Processing Series, Prentice-Hall, Upper Saddle River, NJ, 1995.

[15] T. Watanabe, M. Yachida, Real-time gesture recognition using eigenspace from multi-input image sequences, in: Proceedings of the Third International Conference on Automatic Face and Gesture Recognition, 1998, pp. 428–433.

[16] N. Huang, S. Shen, Hilbert–Huang Transform and its Applications, in: Interdisciplinary Mathematical Sciences, vol. 5, World Scientific, Farrer Road, Singapore, 2005.

[17] H. Ishida, T. Takahashi, I. Ide, H. Murase, A Hilbert warping algorithm for recognizing characters from moving camera, in: Proceedings of the Eighth IAPR Workshop on Document Analysis Systems, 2008, pp. 21–27.

[18] H. Ishida, T. Takahashi, I. Ide, H. Murase, A Hilbert warping method for camera-based finger-writing recognition, in: Proceedings of the 19th International Conference on Pattern Recognition, ThCT5.2, 2008.

[19] S. Hahn, Hilbert Transforms in Signal Processing, Artech House, Norwood, MD, 1996.

[20] A. Oppenheim, R. Schafer, Discrete-time signal processing, in: Prentice Hall Signal Processing Series, Prentice Hall, Upper Saddle River, NJ, 1999.

[21] A. Maheswaran, B. Davis, Analytical signal processing for pattern recognition, IEEE Transactions on Acoustics, Speech and Signal Processing 38 (9) (1990) 1645–1649.

[22] T. Zagajewski, Criticism of the definition of instantaneous frequency, Bulletin of the Polish Academy of Sciences 37 (7–12) (1989) 571–580.

[23] Nagoya University, MIST Project 〈http://mist.murase.m.is.nagoya-u.ac.jp/trac-en/〉.

[24] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-steady time series analysis, Proceedings of the Royal Society of London Series A 454 (1998) 903–995.

About the Author—HIROYUKI ISHIDA received his B.S. degree from the Department of Information Engineering at Nagoya University and M.S. and Ph.D. degrees from the Graduate School of Information Science at Nagoya University.

About the Author—TOMOKAZU TAKAHASHI received his B.S. degree from the Department of Information Engineering at Ibaraki University, and his M.S. and Ph.D. from the Graduate School of Science and Engineering at Ibaraki University. His research interests include computer graphics and image recognition.

About the Author—ICHIRO IDE received his B.S. degree from the Department of Electronic Engineering, his M.S. degree from the Department of Information Engineering, and his Ph.D. from the Department of Electrical Engineering at The University of Tokyo.

He is currently an Associate Professor in the Graduate School of Information Science at Nagoya University.

About the Author—HIROSHI MURASE received his B.S., M.S., and Ph.D. degrees from the Graduate School of Electrical Engineering at Nagoya University.

He is currently a Professor in the Graduate School of Information Science at Nagoya University.

He received the Ministry Award from the Ministry of Education, Culture, Sports, Science and Technology in Japan in 2003.

He is a Fellow of the IEEE.