A Hilbert Warping Algorithm for Recognizing Characters from Moving Camera

Hiroyuki Ishida, Ichiro Ide and Hiroshi Murase
Graduate School of Information Science,
Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi,
464-8601 Japan
{hishi, ide, murase}@murase.m.is.nagoya-u.ac.jp

Tomokazu Takahashi
Department of Economics and Information,
Gifu Shotoku Gakuen University
Nakauzura 1-38, Gifu-shi, Gifu,
501-6194 Japan
ttakahashi@gifu.shotoku.ac.jp

Abstract

We present a method for recognizing characters from image sequences captured by moving camera. In the proposed method, the sequence of the captured images is compared with those of reference character patterns using the concept of analytic signal. Since the captured image sequence can be nonlinearly warped along the time axis due to the movement of a hand-held camera, phase synchronization of two analytic signals is used for the alignment of two image sequences. Hilbert transform is used to convert all the image sequences into analytic signals whose phases are supposed to be increasing. Experimental results showed the usefulness of the proposed phase-based alignment algorithm.

1 Introduction

Character recognition technologies using portable digital cameras have gained attention in recent years in proportion to the diffusion of portable digital imaging devices. To date, many methods have been proposed to solve challenging problems which arise in camera-based character recognition [1]. In this paper, we consider the case where characters cannot be identified from a single frame captured by a still camera. For example, longer texts require multiple frames to be captured with sufficient resolution. In this case, the camera should be moved horizontally along the texts. Another problem arises when the images are captured from a moving camera. Since the camera speed is not constant, the classification task requires sequence alignment between the input and the reference patterns. Such problem occurs also in applications using in-vehicle camera [2].

A sophisticated method called Mozaicing-by-recognition was proposed by Miyazaki et al. in [3] and [4]. It employs DP matching [5], also known as Dynamic Time Warping (DTW), to measure global distances among sequences for classification. Although DTW is a powerful method for sequence alignment, misclassification can still occur because DTW finds the best alignment even for the sequences of incorrect answers. In this paper, we propose a novel method for recognizing image sequences captured by moving camera. Unlike DTW, the proposed method does not ensure proper alignment for incorrect answers. Such incorrect answers are expected to be rejected due to failure of the sequence alignment. This strategy can be implemented through phase synchronization of analytic signals [6], [7]. The analytic signal is a complex signal whose imaginary component is created by applying Hilbert transform [8] to the real component. An important property of the analytic signal is that its instantaneous phase increases at any time. In other words, the trajectory of the analytic signal rotates around the origin in the complex plane. To make use of this property for the sequence alignment, all image sequences are converted into the form of analytic signal. They can then be compared and simultaneously classified by the phase synchronization; two sequences are aligned successfully by correcting the frame-to-frame phase shift, if and only if, they belong to the same category. This property is suitable for classification task because over-fitting to incorrect categories is restricted.

The concept of the analytic signal has been rarely used in image recognition, though it has been applied to artificial one-dimensional signal in [9]. The algorithm in [9] is problematic for applying to actual 2D images. Phase-only correlation [10] is proposed for the matching of fingerprints, though it requires the analysis of the resulting images. Gabor filters are used to obtain the local phase information [11]. In the case where the camera moves toward the same direction, however, the phase shift can be measured more simply from the analytic signal. A Hilbert warping algorithm proposed in this paper is a straightforward approach to recognize image sequences, since it calculates the similarity and the frame-to-frame phase shift at the same time.

The proposed scheme for the sequence alignment is illustrated in Fig. 1. The Hilbert transform is initially ap-

978-0-7695-3337-7/08 $25.00 © 2008 IEEE
DOI 10.1109/DAS.2008.38

IEEE computer society

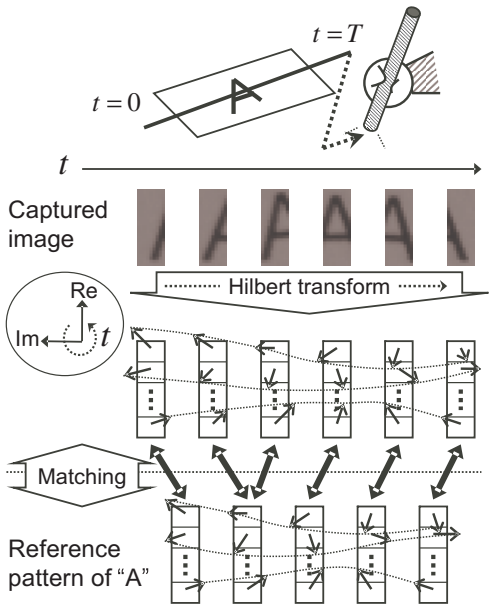


Figure 1. Hilbert warping algorithm for character recognition. Each vertical set of rectangles with rotating phasors represents a vector consisting of analytic signals (ASV).

plied to each captured image along the horizontal axis. This transformation generates a vector which we shall call analytic signal vector (ASV). Likewise, the reference pattern is composed of the sequence of ASVs. The recognition result is obtained by comparing the sequences of the input ASVs and the reference ASVs of all categories.

This paper is organized as follows. Section 2 describes the Hilbert transform to obtain ASVs. Section 3 describes the proposed Hilbert warping algorithm for character recognition. Results are shown in Section 4.

2 Hilbert transform

The Hilbert transform [8] is a process for generating an analytic signal which rotates around the origin in the complex plane as shown in Fig. 2. It is written in terms of the convolution notation as

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

The analytic signal $a(t)$ composed of the original signal $f(t)$ as a real part and its Hilbert transform $\mathcal{H}[f(t)]$ as an imaginary part is denoted by

$$a(t) = f(t) + j\mathcal{H}[f(t)] = |a(t)|e^{j\phi(t)}, \qquad (2)$$

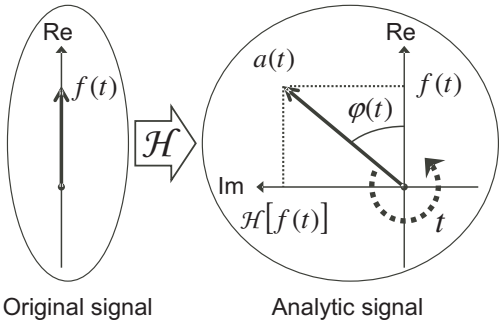


Figure 2. Construction of analytic signal by Hilbert transform. $\mathcal{H}[f(t)]$ is the Hilbert transform of the original signal $f(t)$ along the t axis.

where $\phi(t)$ is defined as instantaneous phase, and $w(t)$ is defined as instantaneous frequency. They are given by

$$\phi(t) = \arctan\frac{\mathcal{H}[f(t)]}{f(t)}, \qquad (3)$$

$$\omega(t) = \frac{d}{dt}\phi(t). \qquad (4)$$

In principle, $a(t)$ rotates counter-clockwise in the complex plane, since the spectrum of the analytic signal is non-negative. Let $F(\omega) = \mathcal{F}[f(t)]$ be the Fourier transform of $f(t)$. The analytic signal $a(t)$ is constructed as follows:

$$\begin{aligned}
\mathcal{F}[a(t)] &= \mathcal{F}[f(t) + j\mathcal{H}[f(t)]] \\
&= \mathcal{F}[f(t)] + j\mathcal{F}[\mathcal{H}[f(t)]] \\
&= \mathcal{F}[f(t)] + j\mathcal{F}[1/\pi t]\,\mathcal{F}[f(t)] \\
&= F(\omega) + j(-j)\mathrm{sgn}(\omega)F(\omega) \\
&= F(\omega)[1 + \mathrm{sgn}(\omega)] \\
&= \begin{cases} 2F(\omega) & (\omega > 0) \\ F(\omega) & (\omega = 0) \\ 0 & (\omega < 0) \end{cases} \qquad (5)
\end{aligned}$$

2.1 Hilbert transform of an image

In the proposed method, the discrete Hilbert transform [1] [12] is applied to images along their horizontal axis, assuming that the camera moves from left to right. Let $f_t(x, y)$ denote the t-th image in the captured sequence;

$$\begin{aligned} x &= 1, \cdots, X, \cdots, 2^n \quad (2^{n-1} < X \le 2^n) \\ y &= 1, \cdots, Y, \end{aligned}$$

[1] We developed a library hht.h for using the Hilbert transform in MIST [13].

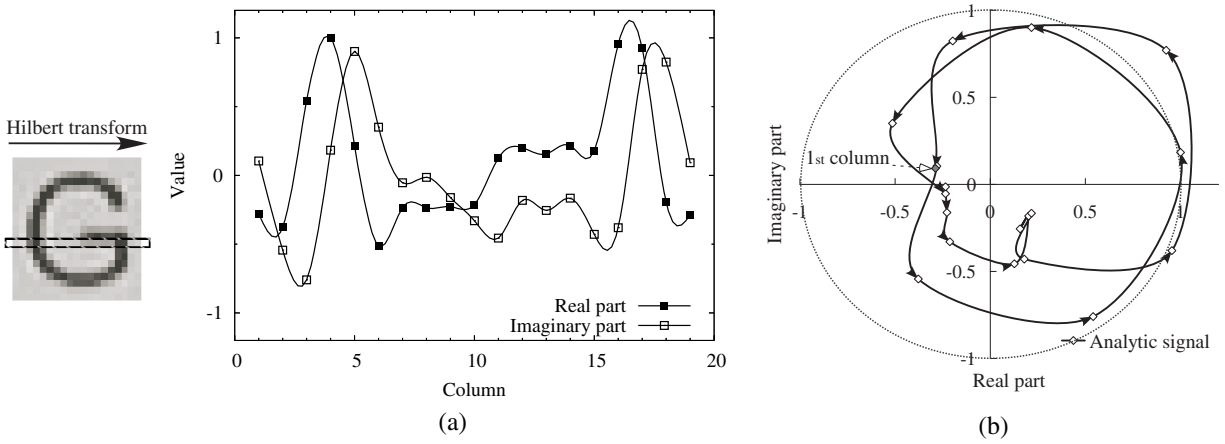


Figure 3. Analytic signals made of an image row indicated by rectangle. (a) The real part of the analytic signal is identical to the intensity. Its Hilbert pair is in the imaginary part. (b) Analytic signal contour in the complex plane.

where X is the image width, and Y is the image height. Let the individual intensities of the image $f_t(x, y)$ be shifted so as to make their mean 0. The analytic expression of the image in the frequency domain is obtained by

$$\mathcal{F}\left[a_t(u,y)\right] = \begin{cases} F_t(u,y) & (u=0) \\ 2F_t(u,y) & (u=1,\cdots,2^{n-1}-1) \\ 0 & (u=2^{n-1},\cdots,2^n-1) \end{cases} \tag{6}$$

with $F_t(u, y)$ given by applying the 1D discrete Fourier transform to each image row ($F_t(u, y) = \mathcal{F}[f_t(x, y)]$). By Eq. (6), the positive spectrum is doubled, and the negative spectrum is restricted to 0. The analytic signal $a_t(x, y)$ can be obtained simply by the inverse discrete Fourier transform. An example of $a_t(x, y)$ is shown in Fig. 3, where we see that the instantaneous phase increases, as x increases. Next, $a_t(x, y)$ is converted to an ASV $\boldsymbol{a}(t)$, whose norm is 1. If the slit width (the number of image columns used for the sequence alignment [3]) is 1, the ASV is represented as

$$\boldsymbol{a}(t) = \frac{1}{\sqrt{\sum_y \overline{a_t(X/2,y)} a_t(X/2,y)}} \begin{bmatrix} a_t(X/2,1) \\ a_t(X/2,2) \\ \vdots \\ a_t(X/2,Y) \end{bmatrix}. \tag{7}$$

2.2 Calculation of similarity and phase shift

The Hermitian inner product of ASVs can be used both for frame-to-frame similarity evaluation and for phase synchronization. Suppose that input ASVs $\boldsymbol{a}^{in}(t)$ are compared to reference ASVs $\boldsymbol{a}^{(c)}(t)$ of category c. The Hermitian inner product $s^{(c)}(t_1, t_2)$ between the t_1-th frame of the reference ASVs and the t_2-th frame of the input ASVs is given by

$$s^{(c)}(t_1,t_2) = \left[\boldsymbol{a}^{(c)}(t_1)\right]^* \boldsymbol{a}^{in}(t_2), \tag{8}$$

where the superscript $*$ denotes the complex conjugate transpose of a vector. Figure 4 shows an example of $s^{(c)}(t_1, t_2)$.

The magnitude of the complex number $s^{(c)}(t_1, t_2)$ is defined as a similarity measure among the frames. It is given by

$$\left|s^{(c)}(t_1,t_2)\right| = \sqrt{\mathrm{Re}\left[s^{(c)}(t_1,t_2)\right]^2 + \mathrm{Im}\left[s^{(c)}(t_1,t_2)\right]^2}. \tag{9}$$

Likewise, the angle of $s^{(c)}(t_1, t_2)$ is given by

$$\angle s^{(c)}(t_1,t_2) = \frac{\mathrm{Im}\left[s^{(c)}(t_1,t_2)\right]}{\mathrm{Re}\left[s^{(c)}(t_1,t_2)\right]}. \tag{10}$$

Note that the phases of the ASVs increase, as t increases [2], since the directions of the x-axis and the t-axis in Eq. (7) are equivalent if the camera moves along the x-axis. Therefore,

[2] This is not strictly correct. Although the spectrum of the analytic signal is 0 for negative frequencies, the instantaneous frequency can be negative [14]. For example, the analytic signal contour in Fig. 3 has a small loop due to the negative instantaneous frequency. Fortunately, such loops tend to be small if the original signal is made from black and white images such as characters. The problem of the loops is almost negligible here because the proposed method evaluates the Hermitian inner product which averages phase shifts over elements of ASVs.

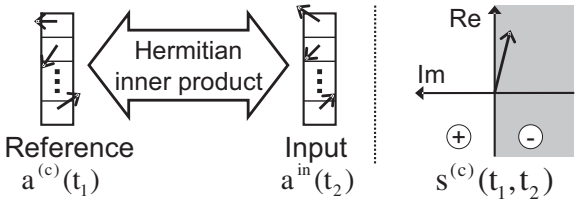


Figure 4. Hermitian inner product of ASVs.

if the input sequence belongs to category c, the following relation is available to predict the frame-to-frame alignment:

$$\left|s^{(c)}\left(t_1 + \mathrm{sgn}\angle s^{(c)}(t_1,t_2), t_2\right)\right| > \left|s^{(c)}(t_1,t_2)\right| \quad (11)$$

This property is used to find the time-warping path.

3 Hilbert warping

The proposed Hilbert warping algorithm explores the time-warping path by tracing the node where $\angle s^{(c)}(t_1,t_2) = 0$. While the conventional DTW calculates the time-warping path also from a cross-similarity matrix (Fig. 5 (a)), the proposed method uses the magnitude $\left|s^{(c)}(t_1,t_2)\right|$ (Fig. 5 (b)) and the angle $\angle s^{(c)}(t_1,t_2)$ (Fig. 5 (c)) for classification and path searching, respectively. Note that the calculation of all nodes is not required. The similarity between sequences can be obtained one by one by following the sign of the phase shift as illustrated in Fig. 6. The algorithm shown in Table 1 calculates the similarity $S^{(c)}$ to category c. The input is classified to the category with the largest similarity.

4 Experimental results

Experiments were conducted to examine the effectiveness of the Hilbert warping algorithm. It was compared to the conventional DTW used in [3]. The cross-similarity matrix used for the conventional DTW was calculated by normalized correlation of real-valued vectors instead of ASVs. The main difference is that the conventional DTW used only the real components in Eqs. (7) and (8). The slope constraints [5] of the DTW were set as

$$(t_1 - k, t_2 - 1) \rightarrow (t_1, t_2), \qquad (0 \le k < K), \quad (12)$$

which means that the maximum allowable camera speed is $K - 1$ [pixel/frame] [4]. This type of constraints is used also in [3]. Using Eq. (12), we can easily compare the similarities of reference sequences (with various lengths) to the given input sequence ($1 \le t_2 \le T_2$).

In the experiments, recognition accuracy for individual characters was evaluated. The number of categories used

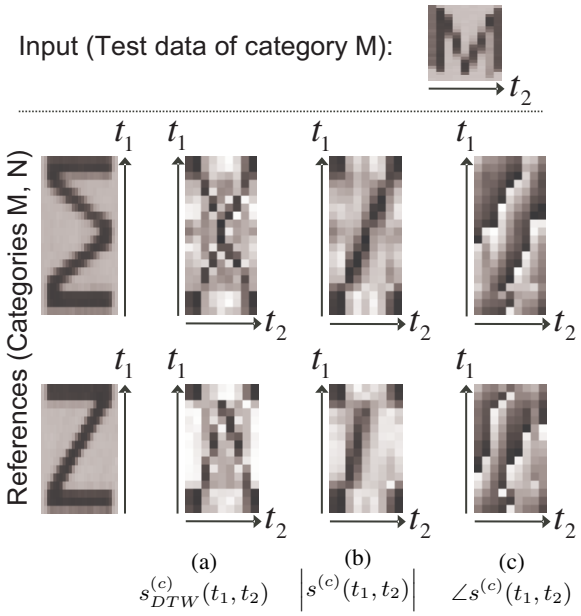


Figure 5. Cross-similarity and phase shift matrices for recognizing an input sequence (category M). (a) Conventional cross-similarity matrix based on normalized correlation (white: 0 → black: 1). (b) Similarities between ASVs (white: 0 → black: 1). (c) Phase shifts between ASVs (white: $-\pi$ → black: π).

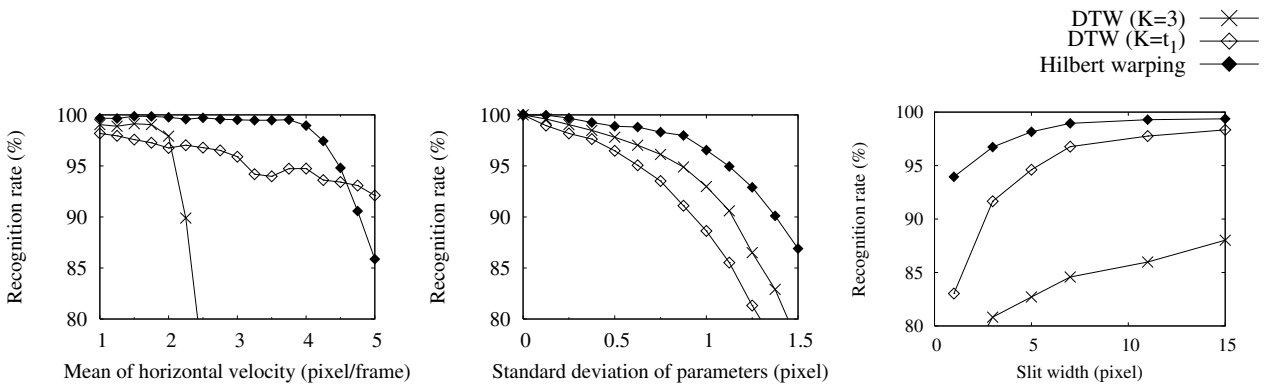
for the experiments was 62 (A–Z, a–z, 1–9: Ariel font). Reference image sequences were generated as illustrated in Fig. 7. Here we introduced parameters μ_{vx}, σ_{vx}, μ_y and σ_y to simulate the horizontal velocity $v_x \approx \mathcal{N}(\mu_{vx}, \sigma_{vx}^2)$ [pixel/frame] and the vertical shift $y \approx \mathcal{N}(\mu_y, \sigma_y^2)$ [pixel] of camera movement. To obtain the images (25×25 pixel) of the reference sequences, these parameters were set as $\mu_{vx} = 1$ and $\mu_y = \sigma_{vx} = \sigma_y = 0$.

4.1 Experiment using artificial test data

The performance of the methods was initially tested using artificial test data generated under various conditions by changing the defined parameters. One hundred image sequences were generated for each category, which resulted in obtaining 6,200 sets of test data.

4.1.1 Performance under various camera velocities

Experimental results obtained by changing the mean of the velocity μ_{vx} are presented in Fig. 8 (a). The other parameters were $\mu_y = 0$ and $\sigma_{vx} = \sigma_y = 0.25$.



(a) μ_{vx} is changed. ($\mu_y = 0,\ \sigma_{vx} = \sigma_y = 0.25$)

(b) σ_{vx} and σ_y are changed. ($\mu_{vx} = 1,\ \mu_y = 0$)

(c) Slit width is changed. ($\mu_{vx} = 2,\ \mu_y = 0,\ \sigma_{vx} = \sigma_y = 1.0$)

Figure 8. Recognition rates for test data generated with horizontal velocity $\mathcal{N}(\mu_{vx}, \sigma_{vx}^2)$ and vertical shift $\mathcal{N}(\mu_y, \sigma_y^2)$. Slit width is 1 except for (c).

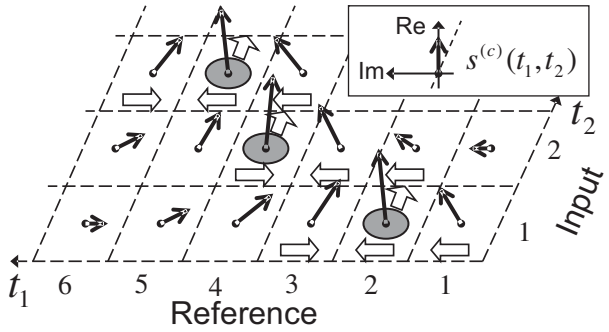


Figure 6. Graph representing the process to search a time-warping path. For each node in the matrix, complex-valued Hermitian inner product $s^{(c)}(t_1, t_2)$ is shown with a black arrow. In this example, it is calculated in the order indicated by a white arrow. Gray circles on the matrix show the resulting time-warping path.

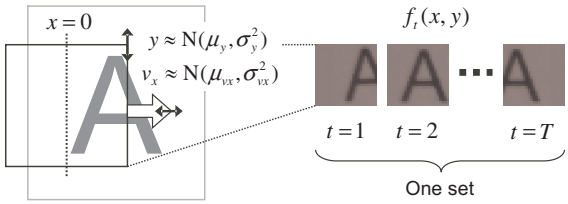


Figure 7. Generation of an image sequence. For each frame, the horizontal velocity parameter v_x and the vertical shift parameter y are generated from normal distributions.

The Hilbert warping algorithm outperformed the conventional DTW while $\mu_{vx} < 4$ even if K in Eq. (12) was set identical to t_1 (without velocity constraints). However, its performance suddenly dropped once $\mu_{vx} > 4$. One reason is that the time-warping path was constructed wrongly once the phase shift went out of the range $(-\pi, \pi)$.

4.1.2 Performance against camera vibration

For simplicity, we simulated the effect of camera vibration by controlling σ_{vx} and σ_y together. The other parameters were $\mu_{vx} = 1$ and $\mu_y = 0$. Results are presented in Fig. 8 (b).

The results indicate that the Hilbert warping algorithm can cope with a certain amount of vibration. To obtain robustness against more serious vibration, learning of various reference sequences will be required.

4.1.3 Expansion of slit width

As proposed in [4], we also investigated the case where a wider slit was used. Provided that the slit width is denoted as w, the vector defined in Eq. (7) consists of elements in the range of $(X/2 - \lfloor w/2 \rfloor \leq x \leq X/2 + \lfloor w/2 \rfloor)$. Results are presented in Fig. 8 (c), where the parameters

Table 1. Hilbert warping algorithm for calculating the similarity $S^{(c)}$ to category c. The Hermitian inner product between the $t_1[i]$-th frame of reference ASVs and the t_2-th frame of input ASVs is denoted as $s^{(c)}(t_1[i], t_2)$, where i is an index used for the iteration.

Hilbert warping algorithm for character recognition	
	/* Initialization */
1	$S^{(c)} \leftarrow 0, \quad t_1[1] \leftarrow 1, \quad t_2 \leftarrow 1, \quad i \leftarrow 1$
2	**do**
3	**do**
	/* Search using the sign of the phase shift */
4	$t_1[i+1] \leftarrow t_1[i] + \mathrm{sgn}\angle s^{(c)}(t_1[i], t_2)$
5	**until** $\mathrm{sgn}\angle s^{(c)}(t_1[i], t_2)$ changes
	/* t_2 is aligned to $t_1^{\angle 0}$
	which gives the minimum phase shift */
6	$t_1^{\angle 0} \leftarrow \arg\min_{t_1[i]} \left\lvert \angle s^{(c)}(t_1[i], t_2) \right\rvert$
7	$S^{(c)} \leftarrow S^{(c)} + \left\lvert s^{(c)}\left(t_1^{\angle 0}, t_2\right) \right\rvert$
8	$t_1[1] \leftarrow t_1^{\angle 0}, \quad t_2 \leftarrow t_2 + 1, \quad i \leftarrow 1$
9	**until** t_2 reaches the last frame T_2
10	**return** $S^{(c)}$

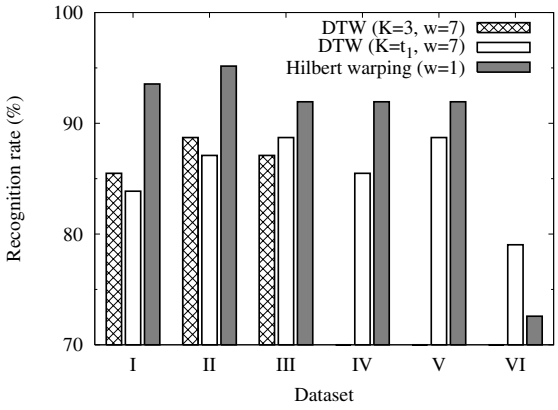


Figure 9. Recognition rates for datasets I–VI captured by six persons. Recognition rates by DTW ($K = 3$) for datasets IV, V and VI were 61.3%, 59.7% and 29.0%, respectively.

Table 2. Estimated horizontal velocity $\hat{\mu}_{vx}$ of datasets.

Dataset	I	II	III	IV	V	VI
$\hat{\mu}_{vx}$	0.73	0.75	1.36	1.96	2.42	3.93

were $\mu_{vx} = 2$, $\mu_y = 0$ and $\sigma_{vx} = \sigma_y = 1$.

The expansion of the slit width improved the performance of both methods. It is remarkable that the Hilbert warping algorithm achieved higher accuracy with narrower slits.

4.2 Experiment using hand-held camera

The second experiment was conducted using test data captured by a moving camera. The test data were composed of six datasets (I–VI), which were captured by six different inexperienced users. To obtain each dataset, 62 image sequences were captured by moving the camera along 62 different characters printed on a paper. The camera speed was determined arbitrarily by each user. All the images were trimmed and normalized such that the height of the character became approximately 25 pixels.

Results are presented in Fig. 9. In this experiment, the Hilbert warping algorithm was compared with the conventional DTW which used slits of wider width ($w = 7$ [pixel]). For each dataset, horizontal velocity $\hat{\mu}_{vx}$ was estimated from the ratio of the length of the reference sequence to that of the input sequence. Table 2 shows the estimated horizontal velocities.

4.3 Discussion

The proposed Hilbert warping algorithm exhibited the highest rates for the datasets I–V in spite of the narrow slit width. Figure 10 shows an example in which the proposed algorithm provided a correct answer n, while the others recognized n as m. By evaluating the value obtained from Eq. (8), the proposed algorithm found a correct alignment for the correct category and an incorrect alignment for the incorrect category, which is a desirable property for classification.

If the camera speed was too high such as the case of the dataset VI, however, the proposed algorithm did not work properly due to the failure of the phase synchronization. Future work will consider extending the algorithm to overcome the velocity limitation.

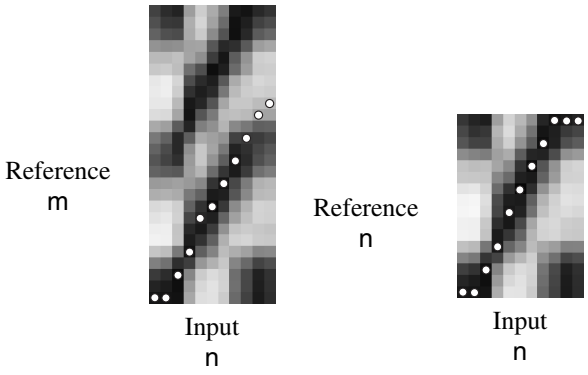


Figure 10. Cross-similarity matrix constructed by Hilbert warping algorithm to recognize n. Circles in the matrix represent the time-warping path through which similarity is calculated.

5 Conclusion

In this paper, we proposed a character recognition method from image sequences captured by moving camera. For this purpose, a sequence alignment algorithm by phase synchronization of analytic signals is introduced. The experimental results showed the usefulness of the proposed method for sequence classification. In future work, the Hilbert warping algorithm will be extended to cope with arbitrary camera movement. A 2D extension of the Hilbert warping algorithm will be interesting for future consideration.

Acknowledgement

Parts of this research were supported by the Grants-In-Aid for JSPS Fellows (19-6540) and Scientific Research (16300054). This work is implemented based on the MIST library [13].

References

[1] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: a survey," Int. Journal of Document Analysis and Recognition, vol.7, no.2–3, pp.84–104, July 2005.

[2] J. Sato, T. Takahashi, I. Ide, and H. Murase, "Change detection in streetscapes from GPS coordinated omnidirectional image sequences," Proc. 18th Int. Conf. on Pattern Recognition, vol.4, pp.935–938, Hong Kong, China, August 2006.

[3] H. Miyazaki, S. Uchida, and H. Sakoe, "Mosaicing by recognition: a technique for video-based text recognition," Proc. 8th Int. Conf. on Document Analysis and Recognition, pp.904–908, Seoul, Korea, August 2005.

[4] S. Uchida, H. Miyazaki, and H. Sakoe, "Mosaicing-by-recognition for video-based text recognition," Pattern Recognition, Elsevier Science Inc., vol.41, no.4, pp.1230–1240, New York, USA, April 2008.

[5] H. Sakoe and S. Chiba, "A dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoustics, Speech and Signal Processing, vol.26, no.1, pp.43–49, February 1978.

[6] L. Cohen, "Time-frequency analysis," Prentice Hall Signal Processing Series, Prentice Hall, Upper Saddle River, NJ, 1995.

[7] J. Horel, "Complex principal component analysis: theory and examples," Journal of Climate and Applied Meteorology, vol.23, no.12, pp.1660–1673, December 1984.

[8] S. Hahn, "Hilbert transforms in signal processing," Artech House, Norwood, Maryland, 1996.

[9] A. Maheswaran and B. Davis, "Analytical signal processing for pattern recognition," IEEE Trans. Acoustics, Speech and Signal Processing, vol.38, no.9, pp.1645–1649, September 1990.

[10] K. Ito, H. Nakajima, K. Kobayashi, T. Aoki, and T. Higuchi, "A fingerprint matching algorithm using phase-only correlation," IEICE Trans., vol.E-87-A, no.3, pp.682–691, March 2004.

[11] J. Daugman, "Complete discrete 2D Gabor transforms by neural networks for image analysis and compression," IEEE Trans. Acoustics, Speech and Signal Processing, vol.36, no.7, pp.1169–1179, July 1988.

[12] A. Oppenheim and R. Schafer, "Discrete-time signal processing," Prentice Hall Signal Processing Series, Prentice Hall, Upper Saddle River, NJ, 1999.

[13] MIST project, *http://mist.suenaga.m.is.nagoya-u.ac.jp/trac-en/*.

[14] T. Zagajewski, "Criticism of the definition of instantaneous frequency," Bull. of the Polish Academy of Sciences, vol.37, no.7–12, pp.571–580, 1989.