PAPER

Individuality-Preserving Silhouette Extraction for Gait Recognition and Its Speedup

Masakazu IWAMURA[†a)], *Senior Member*, Shunsuke MORI[†*], Koichiro NAKAMURA[†**], Takuya TANOUE[††], *Nonmembers*, Yuzuko UTSUMI[†b)], Yasushi MAKIHARA[††c)], Daigo MURAMATSU[††***d)], *Members*, Koichi KISE[†e)], *Fellow*, *and* Yasushi YAGI[††f)], *Member*

SUMMARY Most gait recognition approaches rely on silhouette-based representations due to high recognition accuracy and computational efficiency. A fundamental problem for those approaches is how to extract individuality-preserved silhouettes from real scenes accurately. Foreground colors may be similar to background colors, and the background is cluttered. Therefore, we propose a method of individuality-preserving silhouette extraction for gait recognition using standard gait models (SGMs) composed of clean silhouette sequences of various training subjects as shape priors. The SGMs are smoothly introduced into a well-established graph-cut segmentation framework. Experiments showed that the proposed method achieved better silhouette extraction accuracy by more than 2.3% than representative methods and better identification rate of gait recognition (improved by more than 11.0% at rank 20). Besides, to reduce the computation cost, we introduced approximation in the calculation of dynamic programming. As a result, without reducing the segmentation accuracy, we reduced 85.0% of the computational cost.

key words: *silhouette extraction, gait recognition, shape prior, graph-cut segmentation, nearest neighbor search*

1. Introduction

Person authentication from surveillance cameras plays an increasingly important role in forensics (e.g., person re-identification and verification of a perpetrator and a suspect). Gait biometrics [1] has been considered a promising cue for person authentication. It can be utilized even if the perpetrator/suspect is captured at a distance from the surveillance camera.

Approaches to gait recognition mainly fall into two families [2]: model-based and appearance-based. The appearance-based approaches (such as [3]–[5]) have been dominant in the gait recognition community since they work well even for lower-resolution images with less computational cost than the model-based ones. In particular, a mainstream of the appearance-based approaches exploits silhouette-based representations [6]–[11] because they are unaffected by clothing color and texture. Gait recognition accuracy using silhouette-based representations is, however, subject to silhouette quality.

Silhouette extraction, i.e., foreground/background segmentation, has been studied for a long time in image processing and computer vision fields [12]. While traditional approaches to background subtraction exploit pixel-wise background modeling [13], recent approaches take adjacent connectivity or smoothness into consideration for better segmentation. A seminal work on this topic is graph-cut segmentation [14] and its variants: GrabCut [15] and mutual GrabCut [16]. In addition, soft segmentation, a.k.a. alpha matte process of foreground/background, is also considered by the image segmentation community [17], and its effectiveness is demonstrated in the gait recognition community [18]. These approaches work well as long as wrongly assigned regions (e.g., over-segmentation in background and under-segmentation in foreground) are small enough because they can be corrected by imposing the smoothness. However, they do not work if the wrongly assigned regions are too large to be corrected (e.g., the bulk of the under-segmentation in the leg region in Fig. 2 (e)). Hence, the segmentation problem for gait recognition is still unsolved and challenging.

To solve the challenging task, a shape prior is incorporated in the segmentation framework [19]. For example, Liu and Sarkar [20] train an eigen stance, i.e., an eigen space of silhouettes at each gait stance, from clean silhouettes of multiple training subjects, and reconstruct silhouettes of a test subject through the eigen stance. However, the reconstructed silhouettes may not preserve the individuality of the test subject since their variations are limited to the eigen space, i.e., a weighted linear sum of training subjects' silhouettes. Wang et al. [21] also incorporate the shape before silhouette extraction. They matched a standard gait model (SGM) to initially extracted silhouettes. They improved the silhouettes while considering a balance between the matched SGM and the initial silhouettes containing the test subject's individuality in the graph-cut segmentation framework. However, since they use a single SGM, the improved silhouettes tend to be close to the single SGM and may reduce inter-subject variations.

Therefore, we propose a method of individuality-preserving silhouette extraction which efficiently exploits

Manuscript received November 10, 2020.
Manuscript publicized March 24, 2021.

[†]The authors are with Dept. of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai-shi, 599–8531 Japan.
[††]The authors are with Osaka University, Suita-shi, 565–0871 Japan.
[*]Presently, with KYOCERA Document Solutions Inc.
[**]Presently, with Canon Inc.
[***]Presently, with Seikei University.
a) E-mail: masa@cs.osakafu-u.ac.jp
b) E-mail: yuzuko@cs.osakafu-u.ac.jp
c) E-mail: makihara@am.sanken.osaka-u.ac.jp
d) E-mail: muramatsu@st.seikei.ac.jp
e) E-mail: kise@cs.osakafu-u.ac.jp
f) E-mail: yagi@am.sanken.osaka-u.ac.jp

DOI: 10.1587/transinf.2020ZDP7500

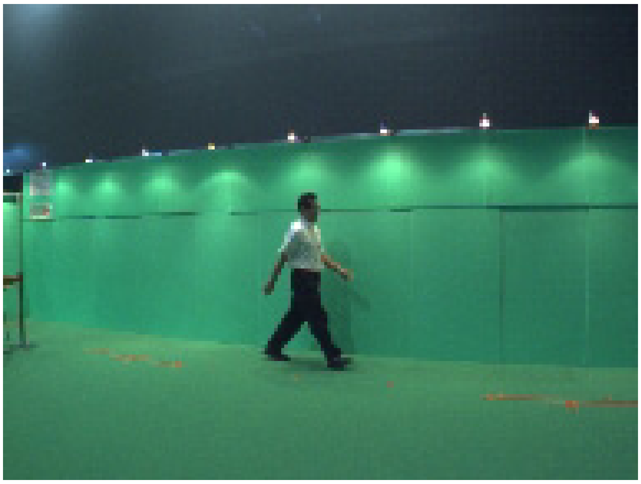

Fig. 1 Sample image from the OU-ISIR Large Population Gait Database.

multiple SGMs in conjunction with the graph-cut segmentation framework. In this context, contributions of this paper are summarized as the following two points.

1. Individuality-preserving silhouette extraction using multiple SGMs: While previous studies [20], [21] may wash out the individuality in the silhouettes, our proposed method keeps the individuality as much as possible. It is realized by selecting the best-fit SGM from multiple SGMs for each test subject and balancing the matched SGM and the initial silhouettes containing the test subject's individuality in the graph-cut segmentation framework.

2. Accuracy improvement in gait recognition: While previous studies [20], [21] did not report the accuracy improvement in gait recognition, we demonstrate that the proposed silhouette extraction actually improves gait recognition accuracy thanks to the individuality-preserving property described above.

Compared with our previous paper [22], in the current paper, we made the following extensions.

1. **Speeding up of the proposed method**
   We improve the proposed method to reduce its processing time by introducing an approximate nearest neighbor method.
2. **Use of new database**
   The database used in our previous paper contained image sequences of pupils wearing backpacks, and the backpacks were included in their silhouettes. In the current paper, to avoid the backpack problem, we prepared a new database based on OU-ISIR Large Population Gait Database [23]. Since image sequences of the OU-ISIR Large Population Gait Database were recorded under a controlled environment (see Fig. 1), it is easy to segment the body region from the background. Hence, to make the problem realistic, we synthesized the background for the database images, and they are not easy to segment anymore.
3. **New evaluation criterion**
   In evaluation, in addition to gait recognition accuracy, silhouette extraction accuracy is evaluated.

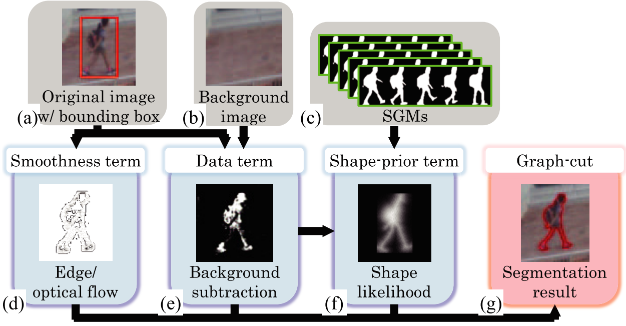

Fig. 2 Framework of the proposed method which leverages a shape prior (f) derived from multiple SGMs (c) as well as the data (e) and the smoothness (d) in graph-cut segmentation for better results (g).

2. Proposed Method

2.1 Problem Setting

In this study, we consider the person authentication of pedestrians captured by two different cameras. Under this problem setting, we assume that the cameras are static and background image sequences without pedestrians for background modeling are available. Moreover, since we focus on silhouette extraction for gait recognition, we assume that well-established pedestrian detectors [24] and trackers [25] give bounding box sequences for individual pedestrians.

2.2 Framework

As with most segmentation approaches, we adopt a graph-cut segmentation framework [14] that assigns a foreground/background label to each pixel through energy minimization. Figure 2 illustrates the framework of the proposed method. Given an original image (Fig. 2(a)) and background image (Fig. 2(b)), a foreground/background likelihood as a data term (Fig. 2(e)) based on background subtraction, and a smoothness term (Fig. 2(d)) to enhance foreground/background label consistency in the spatio-temporal proximity are computed. Besides, the best-matched SGM is found by matching SGMs of multiple persons (Fig. 2(c)) to the data term to compute the shape-prior term (Fig. 2(f)). Then, an energy function $E(X)$ is defined as a weighted linear sum of the data term $E_{\mathrm{dt}}(X_q)$, the smoothness term $E_{\mathrm{sm}}(X_p, X_q)$, and the shape-prior term $E_{\mathrm{sh}}(X_q)$ as

$$\begin{aligned} E(X) = & w_{\mathrm{dt}} \sum_{q \in Q} E_{\mathrm{dt}}(X_q) + w_{\mathrm{sm}} \sum_{(p,q) \in P} E_{\mathrm{sm}}(X_p, X_q) \\ & + w_{\mathrm{sh}} \sum_{q \in Q} E_{\mathrm{sh}}(X_q), \end{aligned} \tag{1}$$

where Q and P are sets of sites (pixels) and edges (pairs of spatio-temporally adjacent pixels), X_q is a foreground/background label at the site q (FG: foreground, BG: background), X is a set of labels for all the sites Q, and w_{dt}, w_{sm}, and w_{sh} are weights to consider the tradeoff among individual terms. Finally, the optimal label assignment is obtained by minimizing the energy function $E(X)$ with the min-cut algorithm. We describe the details of the individual procedures in the following subsections.

2.2.1 Data Term

The data term (corresponding to Fig. 2(e)) is to segment the given input image into the foreground and background regions. In this section, we present how to calculate the background and foreground data terms. The background data term is calculated as follows. First, a pixel-wise background model is trained as a single Gaussian from the given background image sequence. Specifically, as the pixel-wise background model, a mean color vector $\vec{\mu}_{\mathrm{bg},q} \in \mathbb{R}^3$ and a covariance matrix $\Sigma_{\mathrm{bg},q} \in \mathbb{R}^{3\times 3}$ are computed at each site q. Then, the Mahalanobis distance $d_{\mathrm{bg},q}$ between an input color vector $\vec{c}_q$ and the trained background model $\{\vec{\mu}_{\mathrm{bg},q}, \Sigma_{\mathrm{bg},q}\}$ is computed at each site q as

$$d_{\mathrm{bg},q} = \left(\vec{c}_q - \vec{\mu}_{\mathrm{bg},q}\right)^T \Sigma_{\mathrm{bg},q}^{-1} \left(\vec{c}_q - \vec{\mu}_{\mathrm{bg},q}\right). \qquad (2)$$

Finally, the background data term $E_{\mathrm{dt}}(X_q = \mathrm{BG})$ is defined as

$$E_{\mathrm{dt}}(X_q = \mathrm{BG}) = \exp\left(-\kappa_{\mathrm{bg}} d_{\mathrm{bg},q}\right), \qquad (3)$$

where κ_{bg} is a hyper-parameter.

Once we obtain the background data term, then we calculate the foreground data term. The foreground color is represented as a Gaussian mixture model (GMM). Specifically, foreground sample regions are extracted by background subtraction (i.e., thresholding the background data term (Eq. (3)), followed by applying morphological operations containing closing, opening, and area filter. Then, by applying k-mean clustering algorithms to a set of color vectors within the foreground sample regions, a set of means and covariance matrices for the GMM, denoted by $\{\vec{\mu}^k_{\mathrm{fg}}, \Sigma^k_{\mathrm{fg}}\}$ $(k = 1, \ldots, K)$, where K is the number of mixtures, are obtained. Thirdly, Mahalanobis distances $d^k_{\mathrm{fg},q}$ between an input color vector $\vec{c}_q$ are calculated at each site q and the k-th component of the trained foreground GMM as

$$d^k_{\mathrm{fg},q} = \left(\vec{c}_q - \vec{\mu}^k_{\mathrm{fg}}\right)^T \Sigma^{k\ -1}_{\mathrm{fg}} \left(\vec{c}_q - \vec{\mu}^k_{\mathrm{fg}}\right). \qquad (4)$$

Finally, the foreground data term $E_{\mathrm{dt}}(X_q = \mathrm{FG})$ is defined as

$$E_{\mathrm{dt}}(X_q = \mathrm{FG}) = \exp\left(-\kappa_{\mathrm{fg}} \min_k d^k_{\mathrm{fg},q}\right), \qquad (5)$$

where κ_{fg} is a hyper-parameter.

2.2.2 Smoothness Term

The smoothness term (corresponding to Fig. 2(d)) enhances foreground/background label consistency in the spatio-temporal proximity. First, a set of edges P is defined as pairs of spatio-temporally adjacent pixels. While we simply use four connected neighbors for the spatial domain, we use optical flow correspondences [26] for the temporal domain. Then, the smoothness term is defined as

$$E_{sm}(X_p, X_q) = \begin{cases} 0 & \text{if } X_p = X_q, \\ \exp\left(-\kappa_{sm} \frac{\|\vec{c}_q - \vec{c}_p\|^2}{\|\vec{c}_q + \vec{c}_p\|^2 + \varepsilon}\right) & \text{otherwise.} \end{cases} \qquad (6)$$

where $\vec{c}_q$ is an RGB color vector at the site q, $\|\cdot\|$ stands for the L_2 norm, and κ_{sm} and ε are hyper-parameters.

2.2.3 Shape-Prior Term

We introduce the shape-prior term to preserve the individuality. The shape-prior term is calculated by selecting the best-match SGM from SGMs of multiple persons (Fig. 2(c)) in the following manner; a sequence of background subtraction images (i.e., the foreground images of the data term calculated in Sect. 2.2.1) is matched to the SGMs for each frame, and the most similar one is found as the best-match SGM. More specifically, we prepare a sequence of images whose pixel values are set to the background data term at the corresponding site. Then, as the foreground images, we extract a sequence of cropped images based on given bonding boxes, $\{\vec{f}(n)\}$ $(n = 1, \ldots, N)$, where N is the number of frames.

We then introduce a set of SGMs from M training subjects. The SGM for the m-th training subject is composed of an entire period (let it be N^P_m) of clean silhouette sequences, as shown in Fig. 3 (left). The SGM is regarded as a period image with regard to frames, and an index for the frame (or phase, gait stance) is denoted as ϕ $(\phi = 1, \ldots, N^P_m)$.

Since the given bounding box sequences for a target person may contain small deviations from the ground truth, we consider variations to scaling s, horizontal translating t_x, and vertical translating t_y, in addition to the gait stance ϕ. More specifically, we quantize each variation with empirically determined quantization steps as $s = 1 + 0.01 s_s$, $t_x = 0.01 h s_x$, $t_y = 0.01 h s_y$, where h is the image height of the SGM, and s_s, s_x, and s_y are all integers (let a set of the variations of scaling, translation, and phase be a vector $\vec{s} = [\phi, s_s, s_x, s_y]^T$). We further consider the variation range empirically as $S = \{\vec{s} \mid \phi = 1, \ldots, N^P_m, |s_s| \leq 5, |s_x| \leq 5, |s_y| \leq 25\}$, where N^P_m is a complete period (i.e., the number of frames) for the m-th SGM. We can now define an unfolded image vector for the m-th subject's SGM with the variation $\vec{s}$ as $\vec{g}_m(\vec{s})$ $(m = 1, \ldots, M, \vec{s} \in S)$.

Since the variation is represented by a 4-dimensional vector $\vec{s} \in \mathbb{R}^4$, selection of the best-match SGM is cast as a search problem in the 4-dimensional state space by minimizing a certain cost function composed of two terms: a matching cost and a transition cost.

The matching cost for a specific pair of the SGM $\vec{g}_m(\vec{s})$ and $\vec{f}(n)$ is defined by Tanimoto distance [27], which is a standard dissimilarity measure for shape matching, as

$$D_{\text{match}}(\vec{f}(n), \vec{g}_m(\vec{s})) = 1 - \frac{\sum_{(x,y)} \min\{f(x,y;n), g_m(x,y;\vec{s})\}}{\sum_{(x,y)} \max\{f(x,y;n), g_m(x,y;\vec{s})\}}, \tag{7}$$

where $f(x, y; n)$ and $g(x, y; \vec{s})$ are pixel values at the position (x, y) for the background subtraction image $\vec{f}(n)$ and the SGM $\vec{g}(\vec{s})$, respectively. Note that the Tanimoto distance is zero when the two images $\vec{f}(n)$ and $\vec{g}(\vec{s})$ are identical, and that it is one when there is no overlapping between them.

If we search the best-match SGM just by minimizing the matching cost frame-by-frame, it may contain abrupt changes of the variation (e.g., abrupt change of the phase, the scale, or the translation between adjacent frames). We therefore consider the transition cost from state $\vec{s}_m(n-1)$ at the $(n-1)$-th frame to $\vec{s}_m(n)$ at the n-th frame for the m-th SGM to ensure the smoothness as

$$\begin{aligned} D_{\text{trans}}(\vec{s}_m(n-1), \vec{s}_m(n)) =& \Delta_m(\phi_m(n-1), \phi_m(n)) \\ &+ |s_{m,s}(n) - s_{m,s}(n-1)| \\ &+ |s_{m,x}(n) - s_{m,x}(n-1)| \\ &+ |s_{m,y}(n) - s_{m,y}(n-1)| \end{aligned} \tag{8}$$

$$\begin{aligned} \Delta_m(\phi_m(n-1), \phi_m(n)) = \min\{&|\phi(n) - (\phi(n-1)+1)|, \\ & N_m^P - |\phi(n) - (\phi(n-1)+1)|\}. \end{aligned} \tag{9}$$

Note that the phase smoothness is defined by considering periodicity as well as the phase evolution (i.e., the gait stance is evolved as the frame is incremented).

Consequently, the search problem in the 4-dimensional state space is defined as a minimization problem of the weighted sum of the matching cost and the transition cost as

$$\begin{aligned} \{\vec{s}^*_m(n)\} = \mathop{\arg\min}_{\{\vec{s}_m(n)\}} \Bigg\{ & \sum_{i=1}^{N} D_{\text{match}}(\vec{f}(n), \vec{g}_m(\vec{s}_m(n))) \\ & + \alpha \sum_{i=2}^{N} D_{\text{trans}}(\vec{s}_m(n-1), \vec{s}_m(n)) \Bigg\}, \end{aligned} \tag{10}$$

where α is a weight for the transition cost and set to be 0.05 empirically.

For the efficient optimization of the above cost function, we employ a DP framework and introduce the optimal cumulative cost up to the n-th frame for the m-th SGM at the state $\vec{s}_m(n)$ as $C_m(n, \vec{s}_m(n))$. We assume that the optimal path from the first frame to the state $\vec{s}_m(n)$ at the n-th frame is selected.

First, the cumulative cost for the first frame is initialized as follows.

$$C_m(1, \vec{s}_m(1)) = 0 \qquad \forall \vec{s}_m(1) \tag{11}$$

Next, we define a set of previous states, $S(n-1; \vec{s}_m(n))$, which can be transited to a current state $\vec{s}_m(n)$ as

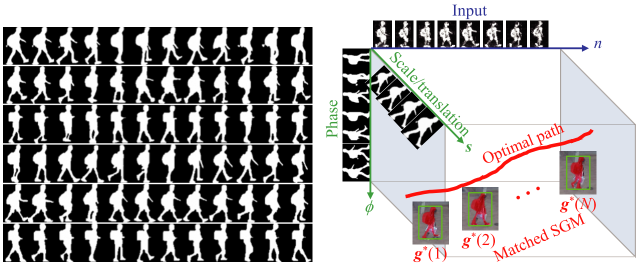

Fig. 3 SGMs (left) and its DP matching (right)

$$\begin{aligned} S(n-1; \vec{s}_m(n)) = \{\vec{s}_m(n-1) | & \Delta_m(\phi_m(n-1), \phi_m(n)) \le \phi^{\text{tol}}, \\ & |s_{m,s}(n) - s_{m,s}(n-1)| \le s_s^{\text{tol}}, \\ & |s_{m,x}(n) - s_{m,x}(n-1)| \le s_x^{\text{tol}}, \\ & |s_{m,y}(n) - s_{m,y}(n-1)| \le s_y^{\text{tol}}\}, \end{aligned} \tag{12}$$

where ϕ^{tol}, s_s^{tol}, s_x^{tol}, and s_y^{tol} are transition tolerance parameters. We set them to 1, respectively, so as to limit the transition to the adjacent states.

We then select the optimal previous state, which transits to the current state $\vec{s}_m(n)$ as

$$\begin{aligned} \vec{s}_m^{\,\text{opt}}(n-1; \vec{s}_m(n)) = & \\ \mathop{\arg\min}_{\vec{s}_m(n-1) \in S(n-1; \vec{s}_m(n))} & \{C_m(n-1, \vec{s}_m(n-1)) \\ & + \alpha D_{\text{trans}}(\vec{s}_m(n-1), \vec{s}_m(n))\}. \end{aligned} \tag{13}$$

Now, we can calculate the cumulative cost at the n-th frame as a sum of the cumulative cost at the previous frame, the transition cost, and the Tanimoto distance at the current frame in a recursive way as

$$\begin{aligned} C_m(n, \vec{s}_m(n)) =& C_m(n-1, \vec{s}_m^{\,\text{opt}}(n-1; \vec{s}_m(n))) \\ &+ \alpha D_{\text{trans}}(\vec{s}_m^{\,\text{opt}}(n-1; \vec{s}_m(n)), \vec{s}_m(n)) \\ &+ D_{\text{match}}(\vec{f}(n), \vec{g}_m(\vec{s}_m(n))). \end{aligned} \tag{14}$$

Once the cumulative costs are calculated, the optimal path is found by backtracking from the optimal state at the last frame as follows.

$$\vec{s}^*_m(N) = \mathop{\arg\min}_{\vec{s}_m(N)} C_m(N, \vec{s}_m(N)) \tag{15}$$

$$\vec{s}^*_m(n-1) = \vec{s}_m^{\,\text{opt}}(n-1; \vec{s}^*_m(n)) \tag{16}$$

We do this process for all the training subjects and select the best training subject with minimal cost as

$$m^* = \mathop{\arg\min}_{m} C_m(N, \vec{s}^*_m(N)). \tag{17}$$

Subsequently, we formulate the shape-prior term based on the matched SGM $\{\vec{g}_{m^*}(\vec{s}^*_{m^*}(n))\}$ (we denote it $\{\vec{g}^*(n)\}$ for simplicity). After we compute the signed distance $d_{\text{sh},q}$ of the matched SGM $\{\vec{g}^*(n)\}$ for the site q (i.e., positive and negative values for inside and outside of the silhouette, respectively), we compute the background/foreground shape-prior terms using a sigmoid function as

$$E_{\mathrm{sh}}(X_q = \mathrm{BG}) = \frac{1}{1 + \exp(-\kappa_{\mathrm{sh}} d_{\mathrm{sh}})} \tag{18}$$

$$E_{\mathrm{sh}}(X_q = \mathrm{FG}) = 1 - E_{\mathrm{sh}}(X_q = \mathrm{BG}), \tag{19}$$

where κ_{sh} is the gain for this sigmoid function.

A property of this representation is that the shape-prior is mitigated near the silhouette contour while it becomes stronger as the site is further from the silhouette contour (i.e., probable inside or outside, see Fig. 2(f)). Thanks to this property, we avoid making the segmentation results too close to the matched SGM, which is beneficial when the matched SGM deviates from the ground truth of a test subject. Moreover, thanks to the multiple SGMs, we can suppress this deviation by selecting the best matched SGM. Therefore, we successfully handle the tradeoff between the data and the shape prior.

2.3 Speedup of Selecting the Matched SGM

Though the proposed framework described in Sect. 2.2 is effective, it requires a long processing time. It is because all the dissimilarities between the input sequence and SGMs are calculated, as shown in Fig. 4. However, only a part of them is actually used. Hence, we propose to reduce the computational cost by limiting the calculation of dissimilarities, as shown in Fig. 5.

Before explaining the proposed method, let us see the detailed process of finding the matched SGM. The process consists of two steps. One is to calculate all the dissimilarities between the input sequence and SGMs. The other is global optimization based on DP matching. Most of the computational cost is spent on the former. However, not all dissimilarities are used in DP matching. Hence, it is preferable to predict dissimilarities required for DP matching and to calculate only the required dissimilarities. To reduce the computational cost, the predicting process should be achieved with less computational cost than calculating the *not-required* dissimilarities.

We propose to use a k-approximate nearest neighbor search method to meet the requirement above. The k-nearest neighbor search is finding the closest k points (nearest neighbors) in the dataset to the given query. The naive way to find them is to calculate the distances (dissimilarities) between the query and all points in the dataset, and then output the k points which have the minimum distances. While the k-nearest neighbor search problem can always be solved, the necessary computational time increases as the number of points in the dataset increases. Hence, an approximation is often introduced to the k-nearest neighbor search problem to drastically reduce computational time while it finds a wrong nearest neighbor with some probability. Usually, k-(approximate) nearest neighbor search is realized by following two steps. In the first step, relatively close points are selected from the points in the dataset in a concise way. In the second step, similarly to the naive way, the distances between the query and points in the dataset are calculated, and the k points whose distances are the minimum are selected. Of the two steps, the first step is suitable for selecting SGMs close to the input.

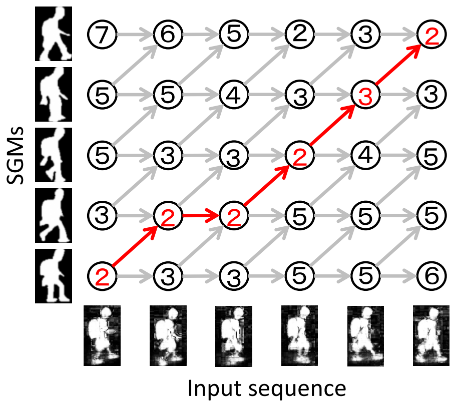

Fig. 4 In the conventional way, all the dissimilarities between the input sequence and SGMs are calculated before DP matching.

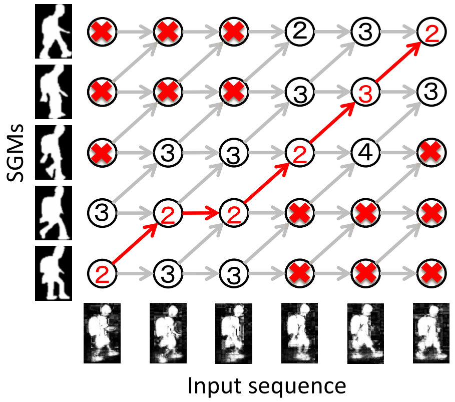

Fig. 5 In the proposed way, only selected dissimilarities are calculated. Red crosses represent the dissimilarities that are not calculated.

The procedure of the proposed speedup method is as follows. First, the dimensionality of the feature of an input and SGMs is reduced by principal component analysis (PCA). Next, some SGMs close to the input are selected by a k-approximate nearest neighbor search method. As the k-approximate nearest neighbor search method, in this paper, we use Bucket Distance Hashing (BDH) [28]. Then, the selected dissimilarities are calculated, as shown in Fig. 5. As for the dissimilarities that are not calculated (red crosses in Fig. 5), we assign the maximum dissimilarity. Finally, the global optimization based on DP matching is performed.

3. Experiments

3.1 Setup

We used image sequences of 109 subjects from OU-ISIR Large Population Gait Database [23]. In the database, the image sequences were captured in 30 frames per second, and the resolution of the images was 800 × 600. Of 4,007 subjects, we used 109 in the experiments; nine for SGMs and 100 for training and test in the task of person authentication.

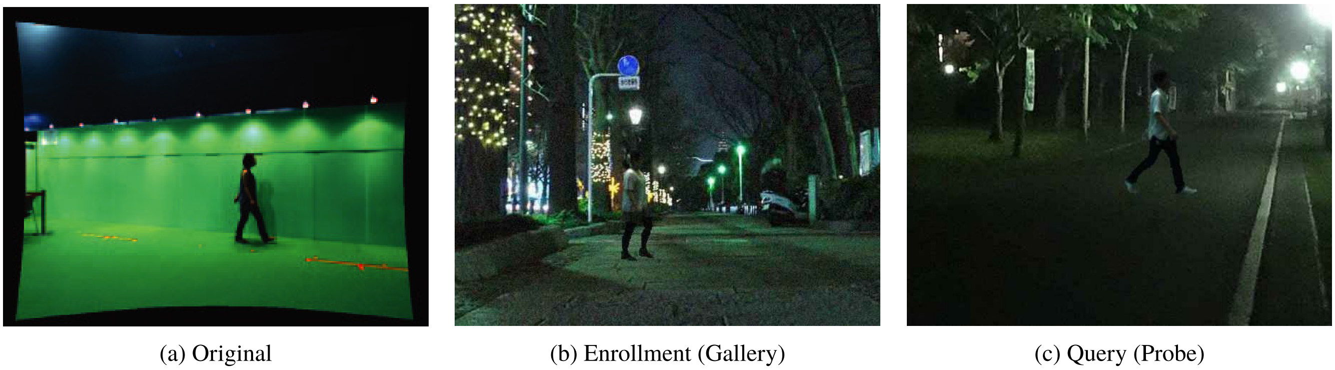

Fig. 6 (a) Original image in OU-ISIR Large Population Gait Database and (b) (c) images used in experiment.

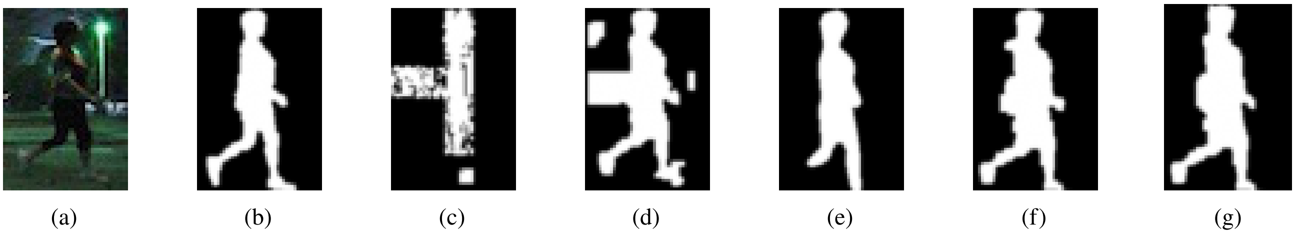

Fig. 7 Results of silhouette extraction. (a) Original image, (b) Ground truth, (c) GrabCut [15], (d) Graph-cut [29], (e) RefineNet [30], (f) Graph-cut with a single SGM [21], (g) Proposed method.

For each subject, two image sequences are available. Hence, one was used as an enrollment (gallery), and the other a query (probe). Each image sequence consisted of from 26 to 39 frames. For SGMs, we selected subjects whose image sequences have 30 frames or more and cropped them to 30 frames per image sequence by extracting 30 consecutive frames. The database was captured under a green-screen background (see Fig. 6(a)). Hence, as shown in Figs. 6(b) and 6(c), we synthesized images using Adobe Photoshop CS6 so that the person parts of the database were put on a realistic background image. Moreover, for increasing the reality by degrading the image quality, we added Gaussian noise which changed for each sequence and converted to JPEG format using lossy compression. Since person detection itself is not a scope of this study, we substituted it by manually annotating rough bounding boxes to pedestrians in a test sequence.

We set the hyper-parameters as follows. The weights in the objective function are $w_{\mathrm{dt}} = 0.7$, $w_{\mathrm{sm}} = 1.0$, and $w_{\mathrm{sh}} = 0.3$. Those of the data term, the smoothness term, the shape-prior term are $\kappa_{\mathrm{fg}} = 0.3$ and $\kappa_{\mathrm{bg}} = 0.02$; $\kappa_{\mathrm{sm}} = 0.01$ and $\varepsilon = 63$; and $\kappa_{\mathrm{sh}} = 0.2$. As for Graph-cut [29], we experimentally set the hyper-parameters of term weights as $w_{\mathrm{dt}} = 1.0$ and $w_{\mathrm{sm}} = 1.0$. In the proposed speedup method, the hyper-parameters of BDH are experimentally set as $k = 100$, and $c = 10$.

We employed the computer where the CPU (Intel Core i7-5820K CPU@3.30GHz), 16GB memory, and OS (Windows 10 Education) were installed.

3.2 Comparison with Benchmarks

We compared the proposed method with four benchmarks: GrabCut [15], Graph-cut [29], RefineNet [30], and *Graph-cut with a single SGM* [21]. As the proposed method uses multiple SGMs, the comparison with *Graph-cut with a single SGM* [21] shows the effectiveness of using multiple SGMs. We selected RefineNet as a representative of deep neural networks because it is one of the most standard method for semantic segmentation at the time of the experiment. As of the time of the paper submission, it is still ranked in a high position in the leaderboard of Cityscapes dataset [31] among those whose source codes are publicly released.

First, we evaluate silhouette extraction performance on two image sequences of 100 subjects. Figured 7 shows typical silhouette extraction results for a qualitative evaluation. Table 1 shows the silhouette extraction accuracy evaluated by Intersection over Union (IoU). They show that the benchmarks without the shape priors (Fig. 7(c), (d), and (e)) suffered from the noise and over-segmentation. In particular, the GrabCut (Fig. 7(c)) was of far less accurate than the others because it was affected by the small color difference between the foreground and background. On the other hand, the benchmarks with the shape prior (Fig. 7(f), and (g)) succeeded in extracting detailed body parts such as as legs, and achieved higher accuracies. Compared with *Graph-cut with a single SGM* (Fig. 7(f)), the proposed method could extract the silhouette more accurately guided by selecting

Table 1 Quantitative evaluation on silhouette extraction.

Method	IoU
GrabCut [15]	19.3
Graph-cut [29]	68.2
RefineNet [30]	69.2
Graph-cut with a single SGM [21]	74.8
Proposed method	77.1

more suitable SGM, and improved the accuracy by 2.3%. RefineNet was no better than Graph-cut. In the experiment, we used the pre-trained network on Person_Parts dataset, which was provided by the authors, without fine-tuning as the dataset was small.

Second, we evaluated the effectiveness of the proposed silhouette extraction in gait recognition, i.e., gait-based person authentication. For this purpose, we adopted GEI [6] as the most widely used silhouette-based gait feature and matched them by Euclidean distance for simplicity. The gait recognition accuracy was evaluated by the cumulative matching characteristics (CMC) curve for an identification scenario (i.e., one-to-many matching). One image sequence of 100 subjects was used for gallery, and the other was used for probe. Figure 8 shows the recognition accuracy of four methods, which exclude GrabCut, where the silhouette extraction was so less accurate that we could not extract GEI. The graph shows that the proposed method improved the accuracy of all other methods while the advantage of the proposed method against *Graph-cut with a single SGM* disappeared in ranks 68 and above. The second best method was *Graph-cut with a single SGM*, which achieved better accuracy than Graph-cut in ranks 10 and above. The third best method was Graph-cut, which achieved better accuracy than Graph-cut in ranks 1 to 49. The order of the methods was almost the same as the quantitative evaluation on the silhouette extraction shown in Table 1, except for RefineNet and Graph-cut. The order of the proposed method (using nine SGMs), *Graph-cut with a single SGM* (using 1 SGM), and Graph-cut (without using SGM) tells that use of more SGMs improves the accuracy.

3.3 Speedup

We evaluated the proposed speedup method with regard to the reduction of computational cost and recognition accuracy by comparing the presence and absence of the approximate nearest neighbor search method. Before applying the approximate nearest neighbor search method, we reduced the dimensionality to 100. Table 2 shows the average computational cost, which is defined as the average time to find the matched SGM over all frames of all input sequences. The table shows that the proposed speedup method (“approximated” in the table) reduced 85% of the computational time of the conventional way (“brute force search”). Table 3 shows the silhouette extraction performance evaluated by Intersection over Union (IoU). Figure 9 shows that the gait recognition accuracy evaluated by cumulative matching characteristics (CMC) curve. Table 3 and Fig. 9 show that the proposed speedup method did not lose accuracy by introducing the approximation dispite 85% of computational cost is reduced.

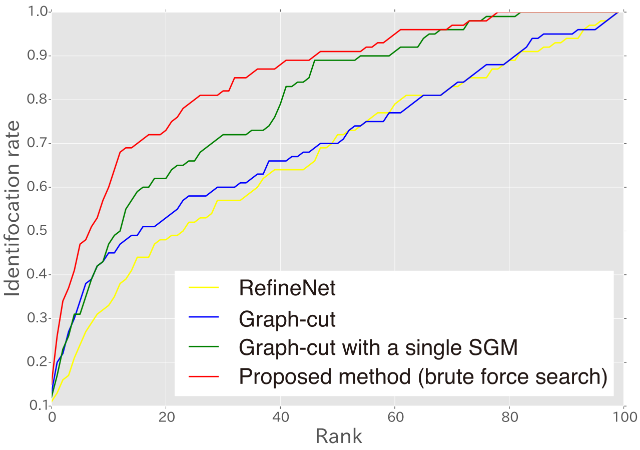

Fig. 8 Quantitative evaluation on gait recognition.

Table 2 Average computational time to find the matched SGM.

Method	Time [ms]
Proposed method (brute force search)	103,982
Proposed method (approximated)	15,567

Table 3 Quantitative evaluation on silhouette extraction.

Method	IoU
Proposed method (brute force search)	77.1
Proposed method (approximated)	77.2

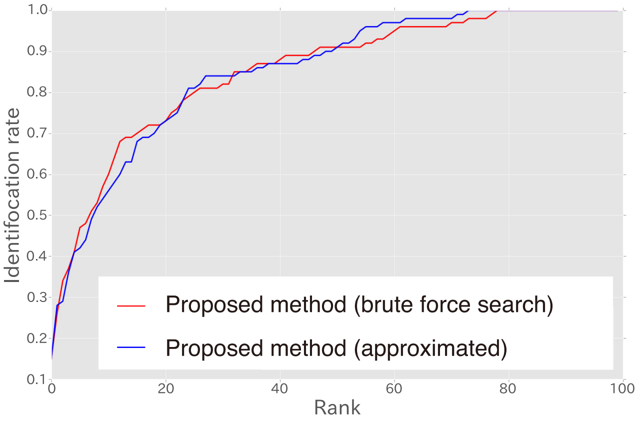

Fig. 9 Quantitative evaluation on gait recognition.

4. Conclusion

We proposed a method of individuality-preserving silhouette extraction for gait recognition. In the problem of silhouette extraction, how to cope with low quality images suffering from noises is an important problem. A feasible solution for this problem is to use a *prior* for the person’s silhouette. The prior can be a silhouette of someone else (i.e., standard gait model (SGM)). However, if the SGM is used as a prior, the individuality of the person of interest can be spoiled because a similar silhouette to that of SGM can be extracted. In this paper, to mitigate the problem, we introduced the use of multiple SGMs. As exploring multiple SGMs takes time, we also proposed a speedup method by using an approximate nearest neighbor search method. The experimental results on silhouette extraction and gait recognition show that the proposed method, which uses multiple SGMs, improved accuracy on both tasks compared to representative methods. Besides, the proposed speedup method reduced 85% of the computational cost without loss of accuracy.

Acknowledgments

This work was partly supported by a JSPS Grant-in-Aid for Scientific Research (A) 15H01693, “R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society”.

References

[1] M.S. Nixon, T. Tan, and R. Chellappa, Human identification based on gait, Springer Science & Business Media, 2006.

[2] Y. Makihara, D.S. Matovski, M.S. Nixon, J.N. Carter, and Y. Yagi, Gait Recognition: Databases, Representations, and Applications, pp.1–15, John Wiley & Sons, Inc., 1999.

[3] J. Little and J. Boyd, “Recognizing people by their gait: The shape of motion,” Videre: J. Comput. Vis. Research, vol.1, no.2, pp.1–13, 1996.

[4] C. BenAbdelkader, R. Culter, H. Nanda, and L. Davis, “Eigengait: Motion-based recognition of people using image self-similarity,” Proc. Int. Conf. on Audio and Video-based Person Authentication, pp.284–294, 2001.

[5] T. Kobayashi and N. Otsu, “Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation,” Proc. the 17th Int. Conf. on Pattern Recognit., pp.741–744, 2004.

[6] J. Han and B. Bhanu, “Individual recognition using gait energy image,” IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.2, pp.316–322, Feb. 2006.

[7] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, “Gait recognition using a view transformation model in the frequency domain,” Proc. ECCV, Lect. Notes Comput. Sci., vol.3953, pp.151–163, Springer, 2006.

[8] K. Bashir, T. Xiang, and S. Gong, “Gait recognition using gait entropy image,” Proc. the 3rd Int. Conf. on Imaging for Crime Detection and Prevention, pp.1–6, Dec. 2009.

[9] K. Bashir, T. Xiang, and S. Gong, “Gait recognition without subject cooperation,” Pattern Recognit. Lett., vol.31, no.13, pp.2052–2060, Oct. 2010.

[10] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan, “Human identification using temporal information preserving gait template,” IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.11, pp.2164 –2176, Nov. 2012.

[11] T.H.W. Lam, K.H. Cheung, and J.N.K. Liu, “Gait flow image: A silhouette-based gait representation for human identification,” Pattern Recognit., vol.44, no.4, pp.973–987, April 2011.

[12] T. Bouwmans, F. Porikli, B. Höferlin, and A. Vacavant, Background modeling and foreground detection for video surveillance, CRC press, 2014.

[13] D.S. Lee, “Effective gaussian mixture learning for video background subtraction,” IEEE Trans. Pattern Analysis & Machine Intelligence, vol.27, no.5, pp.827–832, May 2005.

[14] Y. Boykov and G. Funka-Lea, “Graph cuts and efficient N-D image segmentation,” Int. J. Comput. Vis., vol.70, no.2, pp.109–131, 2006.

[15] C. Rother, V. Kolmogorov, and A. Blake, “GrabCut: Interactive foreground extraction using iterated graph cuts,” ACM Trans. Graph., vol.23, no.3, pp.309–314, Aug. 2004.

[16] Z. Gao, P. Shi, H.R. Karimi, and Z. Pei, “A mutual grabcut method to solve co-segmentation,” EURASIP J. Image and Video Processing, vol.2013, no.1, p.20, 2013.

[17] A. Levin, D. Lischinski, and Y. Weiss, “A closed-form solution to natural image matting,” Proc. IEEE Computer Society Conf. Comput. Vis. Pattern Recognit. (CVPR), pp.228–242, Feb. 2008.

[18] M. Hofmann, S.M. Schmidt, A. Rajagopalan, and G. Rigoll, “Combined face and gait recognition using alpha matte preprocessing,” Proc. the 5th IAPR Int. Conf. on Biometrics, New Delhi, India, pp.1–8, March 2012.

[19] N. Vu and B.S. Manjunath, “Shape prior segmentation of multiple objects with graph cuts,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2008.

[20] Z. Liu and S. Sarkar, “Effect of silhouette quality on hard problems in gait recognition,” IEEE Trans. Systems, Man, and Cybernetics Part B: Cybernetics, vol.35, no.2, pp.170–183, April 2005.

[21] J. Wang, Y. Makihara, and Y. Yagi, “Human tracking and segmentation supported by silhouette-based gait recognition,” Proc. IEEE Int. Conf. Robotics and Automation, pp.1698–1703, 2008.

[22] Y. Makihara, T. Tanoue, D. Muramatsu, Y. Yagi, S. Mori, Y. Utsumi, M. Iwamura, and K. Kise, “Individuality-preserving silhouette extraction for gait recognition,” IPSJ Trans. Comput. Vis. and Applications, vol.7, pp.74–78, 2015.

[23] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, “The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition,” IEEE Trans. Inf. Forensics Security, vol.7, no.5, pp.1511–1521, Oct. 2012.

[24] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” Proc. IEEE Computer Society Conf. Comput. Vis. Pattern Recognit. (CVPR), 2005.

[25] J. Wang and Y. Yagi, “Integrating color and shape-texture features for adaptive real-time object tracking,” IEEE Trans. Image Process., vol.17, no.2, pp.235–240, Feb. 2008.

[26] B.D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” Proc. 7th International Joint Conference on Artificial Intelligence (IJCAI), vol.2, pp.674–679, Aug. 1981.

[27] K.R. Sloan and S.L. Tanimoto, “Progressive refinement of raster images,” IEEE Trans. Comput., vol.28, no.11, pp.871–874, Nov. 1979.

[28] M. Iwamura, T. Sato, and K. Kise, “What is the most efficient way to select nearest neighbor candidates for fast approximate nearest neighbor search?,” Proc. 14th Int. Conf. Comput. Vis. (ICCV 2013), pp.3535–3542, Dec. 2013.

[29] Y. Boykov and G. Funka-Lea, “Graph cuts and efficient N-D image segmentation,” Int. J. Comput. Vision, vol.70, no.2, pp.109–131, Nov. 2006.

[30] G. Lin, A. Milan, C. Shen, and I.D. Reid, “RefineNet: Multi-path refinement networks for high-resolution semantic segmentation,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp.5168–5177, 2017.

[31] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.

Masakazu Iwamura is an associate professor of the Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University. He received the B.E., M.E., and Ph.D. degrees in engineering from Tohoku University, Japan, in 1998, 2000 and 2003, respectively. His research interests include text and object recognition, and visually impaired assistance. He received awards including IAPR/ICDAR Young Investigator Award in 2011, best paper award of IEICE in 2008, IAPR/ICDAR best paper awards in 2007, IAPR Nakano award in 2010, the ICFHR best paper award in 2010, and MVA best paper award in 2017. He worked as the vice-chair of the IAPR technical committee 11 (reading systems) in 2016–2018.

Shunsuke Mori received the B.E., and M.E. degrees in engineering from Osaka Prefecture University, Japan, in 2014 and 2016, respectively. Presently, he works for KYOCERA Document Solutions Inc.

Koichiro Nakamura received the B.E., and M.E. degrees in engineering from Osaka Prefecture University, Japan, in 2017 and 2019, respectively. Presently, he works for Canon Inc.

Takuya Tanoue received the B.S. and and M.S. degrees from Osaka University, Japan, in 2013 and 2015, respectively. His research interest is computer vision including gait recognition.

Yuzuko Utsumi received the MEng degree in 2007 and the Ph.D. degrees in 2010, both from Osaka University. She was an academic visitor in the Active Vision Group at the University of Oxford from 2010 to 2011. After a period as an assistant professor in engineering at Osaka Prefecture University, she was appointed to a lecturer in 2020. She won the MVA best paper award in 2017. She has been involved in research on face recognition and human detection and tracking and, more recently, has focused on image-based plant analysis. She was a research fellowship for young scientists of JSPS from 2009 to 2011. She is a member of IEICE.

Yasushi Makihara received the B.S., M.S., and Ph.D. degrees in Engineering from Osaka University in 2001, 2002, and 2005, respectively. He was appointed as a specially appointed assistant professor (full-time), an assistant professor, and an associate professor at The Institute of Scientific and Industrial Research, Osaka University, in 2005, 2006, and 2014, respectively. He is currently a professor of the Institute for Advanced Co-Creation Studies, Osaka University. His research interests are computer vision, pattern recognition, and image processing. He is a member of IPSJ, IEICE, RSJ, and JSME. He has obtained several honors and awards, including the 2nd Int. Workshop on Biometrics and Forensics (IWBF 2014), IAPR Best Paper Award, the 9th IAPR Int. Conf. on Biometrics (ICB 2016), Honorable Mention Paper Award, and the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, Prizes for Science and Technology, Research Category in 2014. He has served as an associate editor in chief of IEICE Trans. on Information and Systems, an associate editor of IPSJ Transactions on Computer Vision and Applications (CVA), a program co-chair of the 4th Asian Conf. on Pattern Recognition (ACPR 2017), area chairs of ICCV 2019, CVPR 2020, ECCV 2020.

Daigo Muramatsu received the B.S., M.E., and Ph.D. degrees in electrical, electronics, and computer engineering from Waseda University, Tokyo, Japan, in 1997, 1999, and 2006, respectively. He is currently a Professor of The Faculty of Science and Technology, Seikei University. His current research interests include gait recognition, signature verification, and biometric fusion. He is a member of the IEEE, IEICE, and the IPSJ..

Koichi Kise received the B.E., M.E. and Ph.D. degrees in communication engineering from Osaka University, Osaka, Japan in 1986, 1988 and 1991, respectively. From 2000 to 2001, he was a visiting professor at German Research Center for Artificial Intelligence (DFKI), Germany. He is now a Professor of the Department of Computer Science and Intelligent Systems, Osaka Prefecture University, Japan. He received awards including the best paper award of IEICE in 2008, the IAPR/ICDAR best paper awards in 2007 and 2013, the IAPR Nakano award in 2010, the ICFHR best paper award in 2010 and the ACPR best paper award in 2011. He worked as the chair of the IAPR technical committee 11 (reading systems), a member of the IAPR conferences and meetings committee. He is an editor-in-chief of the international journal of document analysis and recognition. His major research activities are in analysis, recognition and retrieval of documents, images and human activities. He is a member of IEEE, ACM, IPSJ, IEEJ, ANLP and HIS.

Yasushi Yagi is a professor of the Institute of Scientific and Industrial Research, Osaka university. He received his Ph.D. degrees from Osaka University in 1991. In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, where he worked on robotics and inspections. He became a Research Associate in 1990, a Lecturer in 1993, an Associate Professor in 1996, and a Professor in 2003 at Osaka University. He was also Director of the Institute of Scientific and Industrial Research, Osaka university from 2012 to 2015, and the Executive Vice President of Osaka university from 2015 to 2019. International conferences for which he has served as Chair include: FG1998 (Financial Chair), OMINVIS2003 (Organizing chair), ROBIO2006 (Program co-chair), ACCV2007 (Program chair), PSVIT2009 (Financial chair), ICRA2009 (Technical Visit Chair), ACCV2009 (General chair), ACPR2011 (Program co-chair) and ACPR2013 (General chair). He has also served as the Editor of IEEE ICRA Conference Editorial Board (2007–2011). He is the Editorial member of IJCV and the Editor-in-Chief of IPSJ Transactions on Computer Vision & Applications. He was awarded ACM VRST2003 Honorable Mention Award, IEEE ROBIO2006 Finalist of T.J. Tan Best Paper in Robotics, IEEE ICRA2008 Finalist for Best Vision Paper, MIRU2008 Nagao Award, and PSIVT2010 Best Paper Award. His research interests are computer vision, medical engineering and robotics. He is a fellow of IPSJ and a member of IEICE, RSJ, and IEEE.