Gait-Based Person Recognition Using Arbitrary View Transformation Model

Daigo Muramatsu, *Member, IEEE*, Akira Shiraishi, Yasushi Makihara, Md. Zasim Uddin, and Yasushi Yagi, *Member, IEEE*

***Abstract*— Gait recognition is a useful biometric trait for person authentication because it is usable even with low image resolution. One challenge is robustness to a view change (cross-view matching); view transformation models (VTMs) have been proposed to solve this. The VTMs work well if the target views are the same as their discrete training views. However, the gait traits are observed from an arbitrary view in a real situation. Thus, the target views may not coincide with discrete training views, resulting in recognition accuracy degradation. We propose an arbitrary VTM (AVTM) that accurately matches a pair of gait traits from an arbitrary view. To realize an AVTM, we first construct 3D gait volume sequences of training subjects, disjoint from the test subjects in the target scene. We then generate 2D gait silhouette sequences of the training subjects by projecting the 3D gait volume sequences onto the same views as the target views, and train the AVTM with gait features extracted from the 2D sequences. In addition, we extend our AVTM by incorporating a part-dependent view selection scheme (AVTM_PdVS), which divides the gait feature into several parts, and sets part-dependent destination views for transformation. Because appropriate destination views may differ for different body parts, the part-dependent destination view selection can suppress transformation errors, leading to increased recognition accuracy. Experiments using data sets collected in different settings show that the AVTM improves the accuracy of cross-view matching and that the AVTM_PdVS further improves the accuracy in many cases, in particular, verification scenarios.**

***Index Terms*— Gait recognition.**

I. Introduction

GAIT recognition is a biometric method for recognizing persons using features extracted from their walking style [1]. Different from many biometric modalities [2], gait recognition has some remarkable characteristics: a gait feature, which has traits for discriminating individuals, can be acquired even from the unconscious gait of an uncooperative subject, at a good distance from the camera, and from relatively low image resolution (e.g., a person with 30-pixel height in an image). Gait recognition can therefore be usefully applied to surveillance or crime investigation using CCTV footage.

However, because gait features of uncooperative subjects may contain covariates that influence the gait itself and/or the appearance of a walking person, robustness to such covariates is quite important for accurate gait recognition. Covariates include, but are not limited to views [3]–[5], walking speeds [6], clothing [7], and belongings [8]. Among the covariates, a change in view occurs frequently in real situations and has a large impact on the appearance of the walking person. Matching gait across different views (cross-view matching) is therefore one of the most challenging and important tasks in gait recognition.

Two different families of approaches for gait recognition have been proposed: appearance-based [4], [8]–[10] and model-based [11]–[15] methods. Appearance-based approaches use captured image sequences directly to extract gait features, while model-based methods extract model parameters from the images. In particular, 3D model-based approaches [13], [15], [16] are preferable for cross-view matching owing to their view-invariant nature. It is, however, difficult in general to calculate 3D model parameters with high accuracy from low quality images captured during surveillance or at a crime scene. Therefore, we focus on appearance-based approaches in this paper.

Many methods have been proposed to tackle the view issue in appearance-based approaches [3]–[5], [17]–[26]. These methods fall into three categories: visual hull-based, view-invariant, and view transformation-based approaches. View-invariant approaches are further divided into three types: geometry-based, subspace-based, and metric learning-based approaches. We summarize the characteristics of these approaches in Table I. Among them, view transformation-based approaches and subspace-based/metric learning-based view invariant approaches have advantages for cross-view matching in surveillance and/or crime investigation thanks to their wider range of applicable view differences and the fact that they do not require the visual hull (3D gait volume) of target subjects. Conversely, the applicable views for these methods are limited to several discrete views included in

Manuscript received January 15, 2014; revised June 28, 2014 and October 30, 2014; accepted October 31, 2014. Date of publication November 20, 2014; date of current version December 9, 2014. This work was supported in part by the Grant-in-Aid for Scientific Research (S) through the Japan Society for the Promotion of Science under Grant 21220003, in part by the Research and Development Program for Implementation of AntiCrime and AntiTerrorism Technologies for a Safe and Secure Society, in part by the Funds for Integrated Promotion of Social System Reform and Research and Development through the Ministry of Education, Culture, Sports, Science and Technology, in part by the Japanese Government, and in part by the Japan Science and Technology Agency CREST Project entitled Behavior Understanding Based on Intention-Gait Model. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xilin Chen.

The authors are with the Institute of the Scientific and Industrial Research, Osaka University, Osaka 567-0047, Japan (e-mail: muramatsu@am.sanken.osaka-u.ac.jp; shiraishi@am.sanken.osaka-u.ac.jp; makihara@am.sanken.osaka-u.ac.jp; zasim@am.sanken.osaka-u.ac.jp; yagi@am.sanken.osaka-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2371335

TABLE I
GAIT RECOGNITION APPROACHES THAT ADDRESS THE VIEW ISSUE IN APPEARANCE-BASED METHODS

Approach	Necessary training data	Cross-view matching	Applicable view	Applicable view difference
Visual hull	Visual hull of all the subjects		arbitrary	-
View invariant (Geometry)	-	✓	limited	narrow
View invariant (Subspace)	Image sequences with target views from non-target subjects	✓	discrete	moderate
View invariant (Metric learning)	Matching score vectors between target views from non-target subjects	✓	discrete	moderate
View transformation	Image sequences with target views from non-target subjects	✓	discrete	moderate
Proposed	Visual hull from non-target subjects	✓	arbitrary	moderate

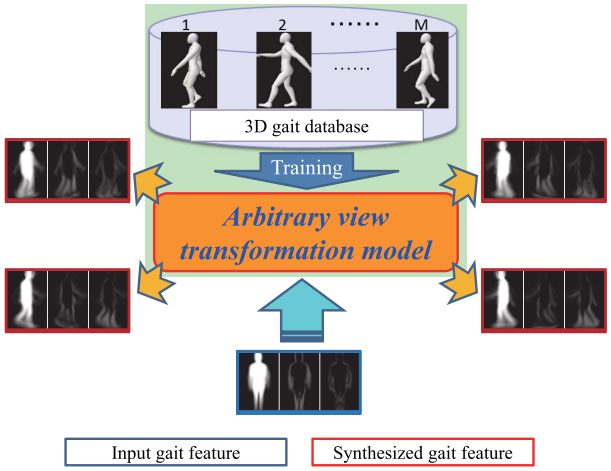

Fig. 1. Framework for arbitrary view transformation model.

the training data (training views) only, and hence, accuracy may drop if the target views are excluded from the training views.

To solve the problems associated with the discrete nature of the required views, we propose a framework that extends the view transformation-based approach to arbitrary views. More specifically, we focus on a view transformation model (VTM) and propose an arbitrary view transformation model (AVTM) for cross-view matching.[1] The AVTM generates gait features with arbitrary views from a gait feature with a different view as shown in Fig. 1. We use 3D gait volume sequences of multiple auxiliary training subjects, and generate training gait features with target views by projecting the 3D gait volume sequences onto 2D spaces associated with the target views. Then, the generated training gait features are used for training the AVTM. Despite using 3D gait volumes, the proposed method differs from the previously mentioned visual hull-based approaches in that it does not require 3D gait volumes of the *target* subjects, but of multiple *non-target (independent)* subjects.

Moreover, since it is more efficient to transform gait features into an intermediate destination view rather than an original view as reported in [27], we propose incorporating a part-dependent view selection (PdVS) scheme and extend our AVTM to an AVTM_PdVS. The AVTM_PdVS divides the gait feature into several body parts, and sets a different destination view for each body part. Because the view with the lowest transformation error is dependent on the body part, part-dependent destination view selection can suppress transformation errors, which in turn leads to improved recognition accuracy. Note that an arbitrary framework makes this view selection feasible because the AVTM framework can generate a training dataset of destination view candidates at sufficiently fine intervals, unlike the existing framework for discrete VTMs that has difficulty in collecting such a training dataset. We can, therefore, generate multiple VTMs for each view candidate, where the appropriate candidate view is selected based on the part-dependent transformation error criterion for each part.

The contributions of this paper include the following five points[2]:

1) *Applicability to Arbitrary Views in Cross-View Matching:* We propose a framework to overcome the discrete nature of conventional VTMs, and to realize the AVTM. Because 3D gait volumes of target subjects are unnecessary, the proposed method can be applied to cross-view matching where a pair of gait features with different views is provided for authentication.
2) *Analysis of Part-Dependent Transformation Error:* As motivation for the AVTM_PdVS, we show that the transformation error of each body part is dependent on the destination view; that is, the appropriate view that achieves a low transformation error is different for each body part. To the best of our knowledge, this is the first work that quantitatively evaluates the impact of the destination view on the transformation error at sufficiently fine intervals. The framework for the AVTM enables us to perform this evaluation.
3) *Proposed AVTM With Part-Dependent View Selection:* We estimate an appropriate destination view separately for each part using training data, and calculate a matching score for each using a pair of gait features transformed to the estimated view. This PdVS contributes to suppressing transformation errors, and leads to improved recognition accuracy.
4) *Accuracy Improvement of Cross-View Matching:* Because of the arbitrary nature and view transformation of the proposed method, accuracy of cross-view matching is improved. Moreover, owing to the PdVS, the accuracy of cross-view matching is further improved in many settings.
5) *Evaluation Using Large Population Dataset:* We evaluated the proposed method using a large population

[1]The framework proposed in this paper can also be applied to subspace-based/metric learning-based view invariant approaches.

[2]Extensions of our previous conference paper [26] are related to items 2, 3, 4 and 5.

dataset, a subset of the OU-ISIR large population gait dataset [28] (referred to as the course dataset in the experiments), and obtained statistically reliable results. Moreover, the data collection settings for the course dataset are completely different from those for 3D gait volume sequence generation. Therefore, we can show the applicability of the proposed method to realistic scenes.

II. Related Work

View difference is one of the most problematic and challenging issues in appearance-based gait recognition, and several methods have been proposed to address this. Approaches addressing view difference issues are classified into three groups: the visual hull-based approach [17]–[20] the view-invariant approach [3], [21], [22], [29]–[31], and the view transformation-based approach [4], [5], [23]–[26].

The visual hull-based approach assumes that the 3D gait volumes, which can be constructed via a visual intersection method (i.e., a visual hull) from temporally synchronized multi-view gait image sequences, of all the target subjects are available. Images from a virtual view are generated by re-projecting the 3D gait volumes onto the image plane; these generated images are then used for recognition. Although this family of approaches achieves high accuracy, a limitation is that data from all the target subjects captured by multiple temporally synchronized cameras are necessary. Therefore, this type of approach cannot be applied in cases where only two gait image sequences (one for the gallery and the other for the probe) are provided for verification.

Different from the visual hull-based approach, the view-invariant approach does not require such synchronized multi-view gait data from the target subjects, and hence, this method is applicable to cases where only two gait image sequences are provided for verification. This approach can be further divided into geometry-based [3], [21], [22], subspace-based [29], [30], and metric learning-based [31] methods. The geometry-based approach makes use of geometrical properties for feature extraction. For example, Kale et al. proposed a method to synthesize side-view gait images from any other arbitrary view [3] under the assumption that the person (3D object) is represented by a planar object on a sagittal plane. This approach works well in cases where the angle between the sagittal plane of the person and the image plane is small; however, in cases where the angle is large, accuracy deteriorates drastically.

Conversely, the subspace- and metric learning-based approaches do not make such an assumption. Instead, the subspace-based approach learns the joint subspace using training data, and calculates view-invariant features by projecting the original features onto the learned subspace [29], [30]. For example, Lu and Tan proposed uncorrelated discriminant simplex analysis to learn the feature subspace [29], while Liu et al. applied joint principal component analysis to learn the joint subspace of the gait feature pair with different views [30]. The metric learning-based approach learns a weight vector that sets the importance of a matching score associated with each feature, and uses the weight vector to calculate a final score. For example, Martín-Félez and Tao applied the pairwise RankSVM algorithm [32] to improve the accuracy of gait recognition with covariate (view, clothing, and carrying condition) variations [31]. A limitation of the subspace- and metric learning-based methods is that the applicable views are limited to discrete training views.

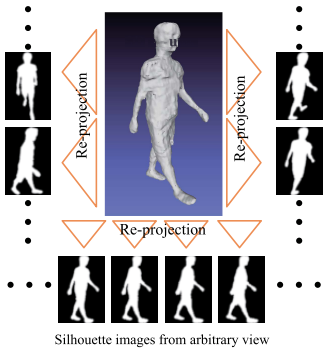

Fig. 2. Re-projection for silhouette image generation.

The view transformation-based approach constructs a VTM using auxiliary training gait image sequences of multiple non-target subjects, and generates a gait feature from a target view using the learned VTM. Note that unlike the visual hull-based approach, the view transformation-based approach does not require synchronized multi-view gait data from the target subjects, and hence, it is applicable to cases where only gallery and probe images from different views are available. For training the VTMs, a matrix factorization process through singular value decomposition (SVD) [4], [23], [26], [33], or regression [5], [24], [25] is exploited. Limitations of this type of approach are that the applicable views are limited to discrete training views and accuracy is degraded if the target views differ from the training views.

In this paper, we propose a framework to overcome the limitations associated with the discrete nature of the above methods, and apply the framework to one of the view transformation-based approaches using the matrix factorization process. Note that the proposed framework is also applicable to the view transformation-based approach with regression and/or the subspace-based/metric learning-based view-invariant approach. Moreover, we propose a PdVS scheme for the AVTM. We show that feature transformation to a part-dependent destination view improves both the transformation and recognition accuracy.

III. Algorithm

A. Framework for Arbitrary View

The discrete nature of the VTM-based approach and the subspace-based/metric learning-based view-invariant approach stems from the discreteness of the training views. Generally, the training dataset is composed of 2D image sequences from several discrete views; however, it is impossible to generate 2D image sequences from arbitrary views from these 2D training datasets. Therefore, we propose using 3D volume sequences for training. Using a 3D volume, 2D images from any arbitrary view can be re-projected using associated projection matrices as shown in Fig. 2. Therefore, we can prepare

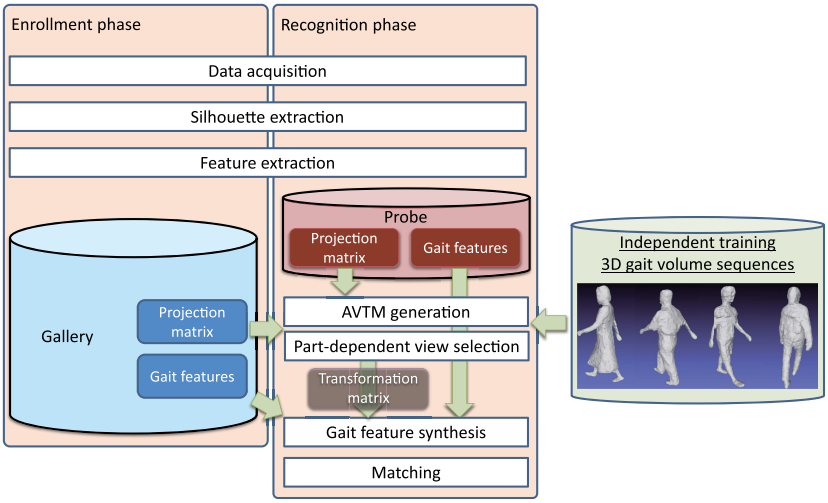

Fig. 3. Process flow of proposed method.

training data associated with any pairs of target views, which enables the construction of an AVTM.

B. Overview of Gait Recognition Algorithm Using AVTM

Fig. 3 shows the process flow of the proposed method using an AVTM. The method includes an enrollment phase and a recognition phase. In both phases, gait image sequences are input together with an ID number. We assume that projection matrices associated with the data are available; these projection matrices are used to generate training data with the target views. Silhouette image sequences are extracted from the gait image sequences using background subtraction-based graphcut segmentation [34], and then gait features are extracted from the silhouette image sequences. Thereafter, the extracted gait features are enrolled with their ID number and associated projection matrices as the gallery and probe in the enrollment and recognition phases, respectively. For AVTM generation, in the recognition phase, training gait features with the target views are generated from independent 3D gait volume sequences associated with multiple non-target subjects using the projection matrices associated with the target views. In this paper, we consider the gallery and probe views as source views, and the intermediate view between these source views as the destination view. All these views are included as target views. The VTMs are trained using a set of generated training gait features with these target views. Note that we consider part-dependent destination views in this paper, and therefore, an appropriate destination view for each part is selected, and gait features transformed to the selected view are used to calculate a matching score for each part. Finally, the part-based matching scores are summed for recognition.

In the remainder of this section, we explain the processing of the AVTM after feature extraction. The process comprises AVTM generation (training), PdVS, gait feature synthesis (transformation), and matching.

C. AVTM Generation

AVTM generation comprises two steps: image re-projection and transformation matrix generation. Let $P(\theta_n)$, $n = 0, 1, \ldots, N+1 \in \mathbb{R}^{3\times 4}$ be the n-th projection matrix associated with view θ_n, and $\mathbf{V}^m_{1:T_m} = (V^m_1, V^m_2, \ldots, V^m_{T_m})$, $m = 1, 2, \ldots, M$ be a cycle of the 3D gait volume sequence associated with the m-th training subject, where M is the number of training subjects and T_m is the number of volumes (period) included in one gait cycle of the m-th training subject.

Step 1) Image Re-Projection: A silhouette image sequence of the m-th subject $\mathbf{S}^m_{1:T_m}(\theta_n) = \left(S^m_1(\theta_n), \ldots, S^m_{T_m}(\theta_n)\right)$ from view θ_n is generated using the 3D gait volume sequence of the m-th subject by

$$S^m_t(\theta_n) = \text{Project}(V^m_t; P(\theta_n)), \quad t = 1, 2, \ldots, T_m, \tag{1}$$

where $\text{Project}(\cdot; P)$ is a function that projects a 3D volume onto a 2D image plane using projection matrix P. Then, we extract gait feature $X^m(\theta_n) \in \mathbb{R}^{K\times 1}$ by

$$X^m(\theta_n) = \text{Extract}(\mathbf{S}^m_{1:T_m}(\theta_n)), \tag{2}$$

where $\text{Extract}(\cdot)$ is a function that extracts gait features from a gait silhouette image sequence. As the gait feature, any type of gait feature with a fixed dimension (e.g., gait energy image (GEI) [9] or frequency domain feature (FDF) [4]) is acceptable.

Step 2) Transformation Matrix Generation: Using the extracted gait features of the training data, we generate a training matrix for VTM generation.

Let $\theta_0 = \theta_G$ and $\theta_{N+1} = \theta_P$ be the gallery and probe views, respectively. Moreover, let $\psi \in \{\theta_0, \ldots, \theta_{N+1}\}$ be the destination view. Using the training gait features of θ_G, θ_P, and destination view ψ, we generate training matrix $\mathbf{D}_\psi$ by

$$\mathbf{D}_\psi = \begin{bmatrix} X^1(\theta_G) & \cdots & X^m(\theta_G) & \cdots & X^M(\theta_G) \\ X^1(\psi) & \cdots & X^m(\psi) & \cdots & X^M(\psi) \\ X^1(\theta_P) & \cdots & X^m(\theta_P) & \cdots & X^M(\theta_P) \end{bmatrix}. \tag{3}$$

Note that the training matrix is essentially dependent on the three views, θ_G, θ_P, and ψ, but for simplicity we describe it by D_ψ. Then, we decompose the matrix by applying SVD:

$$\begin{aligned} \mathbf{D}_\psi &= \mathbf{U}_\psi \mathbf{S}_\psi \mathbf{V}^T_\psi \\ &= \begin{bmatrix} R_\psi(\theta_G) \\ R_\psi(\psi) \\ R_\psi(\theta_P) \end{bmatrix} \begin{bmatrix} v^1_\psi & \cdots & v^M_\psi \end{bmatrix}, \end{aligned} \tag{4}$$

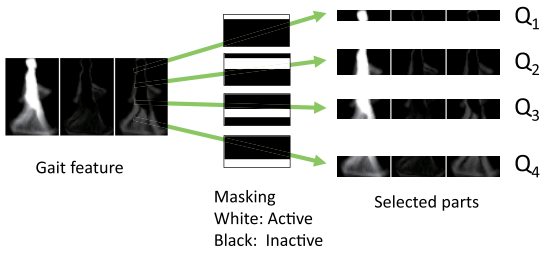

Fig. 4. Part division by masking.

where $\mathbf{U}_\psi \in \mathbb{R}^{3K\times M}$ and $\mathbf{V}_\psi \in \mathbb{R}^{M\times M}$ are orthogonal matrices, and $\mathbf{S}_\psi \in \mathbb{R}^{M\times M}$ is a diagonal matrix where the on-diagonal elements are singular values. In this framework, gait feature $X^m(\theta), \theta \in \{\theta_G, \theta_P, \psi\}$ in the training data can be described as

$$X^m(\theta) = R_\psi(\theta) v^m_\psi. \tag{5}$$

Here, $R_\psi(\theta) \in \mathbb{R}^{K\times M}$ is a subject-independent matrix, and $v^m_\psi \in \mathbb{R}^{M\times 1}$ is a subject-dependent intrinsic vector. In this paper, we call $R_\psi(\theta)$ the transformation matrix to view θ.

D. Part-Dependent View Selection

Since recognition accuracy of the VTM-based approach is influenced by the destination view [5], [27], an appropriate destination view for recognition must first be selected. Under the assumption that the appropriate destination view differs for each body part, we propose a PdVS scheme where an appropriate destination view for each body part is selected independently. In this paper, we consider four body parts based on known anatomical properties [35], and separate a gait feature into the four parts using a mask matrix $\mathbf{M}_q, q \in \{Q_1, \ldots, Q_4\}$ as shown in Fig. 4. Each part Q_1, Q_2, Q_3, and Q_4 includes a single region, namely, the head, chest, waist, and legs, respectively.

For each of these parts, PdVS is realized in three steps. First, we divide the training data into two non-overlapping data subsets: model training data ($\mathbf{G}^T$) and validation data ($\mathbf{G}^V$), subject to $\mathbf{G}^T \cap \mathbf{G}^V = \phi$, where ϕ denotes the empty set. We then generate transformation matrices $R'_\psi(\theta_G)$, $R'_\psi(\psi)$, and $R'_\psi(\theta_P)$ using the model training data in Eq. (4).

In the second step, using the generated transformation matrices, we transform gait features in the validation data, and measure the transformation difference of each view separately as

$$E_q(\psi) = \sum_{l\in\mathbf{G}^V} \left|\left|\mathbf{M}_q \left(R'_\psi(\psi)\left(R'^{+}_\psi(\theta_G)X^l(\theta_G) - R'^{+}_\psi(\theta_P)X^l(\theta_P)\right)\right)\right|\right|. \tag{6}$$

Here, R^+ denotes the pseudo inverse matrix of matrix R. We then select the appropriate destination view ψ^*_q of the part $q \in \{Q_1, Q_2, Q_3, Q_4\}$ by

$$\psi^*_q = \underset{\psi\in\{\theta_0,\ldots,\theta_{N+1}\}}{argmin} E_q(\psi), \tag{7}$$

and use $\{R_{\psi^*_q}(\theta_G), R_{\psi^*_q}(\theta_P)$, and $R_{\psi^*_q}(\psi^*_q)\}$ for transformation related to part q.

E. Gait Feature Synthesis

Let $x^G_{\theta_G} \in \mathbb{R}^{K\times 1}$ be a gallery gait feature of a target subject from view θ_G, and $x^P_{\theta_P} \in \mathbb{R}^{K\times 1}$ be a probe gait feature of a target subject with view θ_P for recognition, where $\theta_G \neq \theta_P$ in cross-view settings. We denote the test gait features of the target subject by symbol x and not X to distinguish test gait features from independent training gait features, although the type of both gait features is the same.

Using a subset of the transformation matrices and the given gait feature set, we first estimate the intrinsic vector $\hat{v}^{\xi\in\{G,P\}}$ associated with the target subject by

$$\hat{v}^\xi = \underset{v}{argmin} \left|\left|x^\xi_{\theta_\xi} - R_\psi(\theta_\xi)v\right|\right|^2. \tag{8}$$

while the intrinsic vector of the target subject $\hat{v}^\xi$ is estimated using the least squares method by

$$\hat{v}^\xi = \left(R_\psi(\theta_\xi)^T R_\psi(\theta_\xi)\right)^{-1}(R_\psi(\theta_\xi))^T x^\xi_{\theta_\xi}. \tag{9}$$

Using the estimated $\hat{v}^\xi$ and transformation matrix associated with a destination view ψ, we synthesize a gait feature with view ψ by

$$\hat{x}^\xi_\psi = R_\psi(\psi)\hat{v}^\xi. \tag{10}$$

F. Matching

When applying the conventional AVTM, we calculate the dissimilarity score as

$$D_{G2P}(x^G_{\theta_G}, x^P_{\theta_P}) = ||\hat{x}^G_{\theta_P} - x^P_{\theta_P}||. \tag{11}$$

Moreover, when applying AVTM with PdVS, we define a dissimilarity score between the two cross-view gait features as

$$D_{B2PdV}(x^G_{\theta_G}, x^P_{\theta_P}) = \sum_{q\in\{Q_1,..,Q_4\}} \left|\left|\mathbf{M}_q(\hat{x}^G_{\psi^*_q} - \hat{x}^P_{\psi^*_q})\right|\right|. \tag{12}$$

Note that in part-based gait recognition, the fusion method for combining the part-based scores plays an important role [7]; in this study, however, we simply sum the part-based scores, because we want to exclude the positive impact of the fusion method.

IV. Implementation

A. 3D Gait Volume Sequence

The following steps were used to generate 3D gait volume sequences for training.

First, we collected gait image sequences from multiple subjects using the multi-view temporally synchronized camera system shown in Fig. 5. We constructed a circular studio with a treadmill in the center of the studio, and set 12 poles at 30 deg intervals around the studio. Two cameras were placed on each pole at a height of 130 and 200 cm, respectively, and one camera was placed above the treadmill. In total, 25 temporally synchronized cameras were positioned around the studio. The spatial resolution of the captured images from the cameras was 640 by 480 pixels, and 60 images (frames) were captured per second by each camera. Projection matrices for the individual cameras were computed according to the

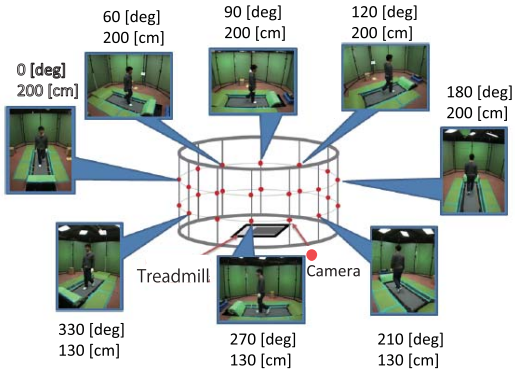

Fig. 5. Data acquisition setup for treadmill dataset.

Fig. 6. Example silhouette images for 3D model generation.

camera calibration. We asked 103 subjects to walk naturally on the treadmill operating at a speed of 4 km/hour. During walking, gait image sequences were captured by the 25 synchronized cameras. We call the views of these cameras *training views*.

Second, we corrected lens distortion using a non-parametric calibration method [36], and extracted silhouette images from the undistorted images by applying background subtraction and graph-cut based segmentation [34]. To obtain high-quality silhouette images, we modified the silhouette image sequences manually. The extracted silhouette images are shown in Fig. 6.

Finally, we constructed a 3D volume from the silhouette images for multiple views using a visual cone intersection technique [37]. Samples of the constructed 3D gait volume sequences are shown in Fig. 3.

B. Gait Feature

From a gait image sequence we extracted the FDF [4] as the gait feature, because this achieves comparable recognition accuracy to the GEI [9] for a large-population dataset [28], and the dimension of the FDF is higher than that of the GEI. A higher dimensional feature is preferable for the VTM. Since the FDF is an image-based feature, we simply converted the 2D feature into a feature vector by means of a raster scan.

C. Projection Matrix of Destination View

To select an appropriate destination view, image sequences of multiple destination views were needed. Projection matrices of the destination views were generated from the projection matrices of the gallery/probe view. We decomposed the projection matrix into an intrinsic and extrinsic parameter matrix comprising a rotation matrix and a translation vector. By modifying the rotation matrix and translation vector, we generated the projection matrix of an intermediate destination view that differs from the gallery/probe view.

V. Experiments

A. Overview

We evaluated the accuracy of the proposed AVTM using two types of data from the OU-ISIR database [28][3]: subsets of the treadmill dataset and the large population dataset. We call the former set the *treadmill dataset*, and the latter the *course dataset*. For the evaluations, we carried out two different experiments.

The aim of the first experiment was to evaluate the applicability of the AVTM in different view settings rather than showing superiority of the AVTM over the other benchmarks such as the discrete VTM (DVTM), using the treadmill dataset. Although the views of the treadmill dataset are almost the same as those of the dataset used for 3D construction, we nevertheless, generated training data for the VTM by projecting 3D volumes with projection matrices associated with the target views because any effect of the projection/re-projection should be included in the accuracy of AVTM.

The second experiment compared the recognition accuracy of the proposed AVTM and AVTM_PdVS with that of the benchmarks in settings where the target views differed from the training views, using mainly the course dataset.

Each experiment consisted of a preliminary experiment and a main experiment. In the preliminary experiment of the first evaluation, we confirmed that the proposed method could synthesize a single gait feature from the gait features obtained from the different views. In the preliminary experiment of the second evaluation, we first showed the synthesized gait features generated by the AVTM. Then, we showed that transformation errors are affected by the destination views, and that appropriate destination views with the lowest transformation errors differ for different parts.[4] In both the main experiments, we evaluated the accuracy of the proposed method for verification and identification tasks.

B. Databases

1) Treadmill Dataset: This dataset is composed of gait silhouette image sequences from 100 subjects. The data in this dataset were acquired in the same setting as the images for 3D gait volume generation, but the subjects were completely different and the camera's pose was slightly different from those for the 3D gait volume generation; therefore, the views of the treadmill dataset were not exactly the same as the training views. Each subject in the treadmill dataset walked on the treadmill twice on the same day, while gait image sequences were captured by 25 cameras. The spatial resolution of each captured image was 640 by 480 pixels, and 60 images (frames) were captured per second by each camera. From the captured image sequences, silhouette image sequences were

[3]http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitTM.html

[4]This preliminary experiment was performed using 3D gait volume sequences since ground truth data for each destination view were needed.

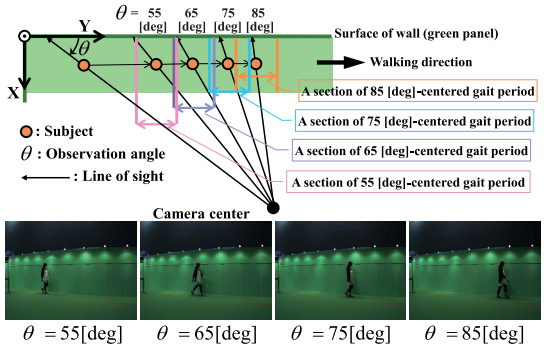

Fig. 7. Walking images of different views.

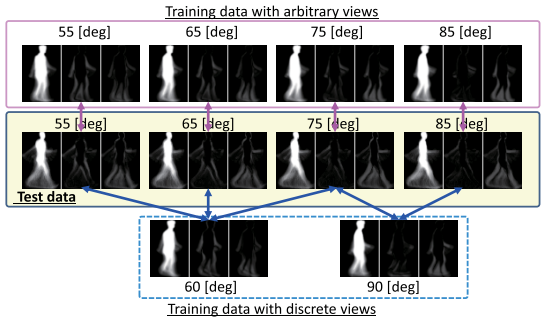

Fig. 8. Gait features of course (test) data, training data with discrete views, and training data with arbitrary views.

generated and FDFs were extracted. We used gait image sequences from the cameras at a height of 200 cm and observation azimuth angles of 0, 30, 60, 90, 120, 150, and 180 deg, where 0, 90, and 180 deg denote the frontal, right side, and rear views, respectively.

2) Course Dataset: We used a subset of the OU-ISIR large population dataset, and chose gait image sequences from 1,912 subjects whose data were captured for evaluation by calibrated cameras. Each subject was asked to walk along a course twice in a natural manner, while gait data were captured using a single camera placed approximately 5 m from the course at a height of 100 cm as shown in Fig. 7. The spatial resolution of each captured image was 640 by 480 pixels, and 30 images (frames) were captured by the camera. Four subsets with different observation views, that is, observation azimuth angles of 55, 65, 75, and 85 deg as shown in Fig. 7, are available in this dataset; we used all four types of data and extracted FDFs for evaluation. This dataset is suitable for evaluating the accuracy of the AVTM, because the two image sequences per subject were captured under similar conditions (e.g., on the same day and with the same attire), excluding covariates other than view. Moreover, the views differ from those in the training data used for 3D gait volume generation described in Section IV and shown in Fig. 8. This dataset was used to evaluate the usefulness of the arbitrary nature of the proposed method.

C. Experiment on the Treadmill Dataset

1) Preliminary Experiment: Fig. 9(a) shows the synthesized gait features with different views that were synthesized from a gait feature with different views. In this figure, we show the

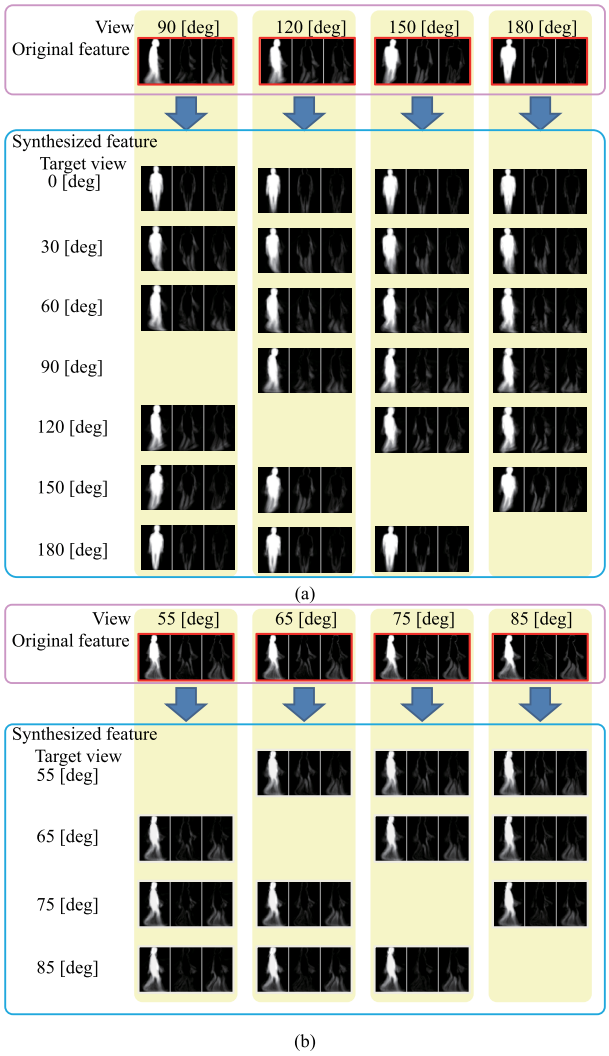

Fig. 9. Synthesized gait features: (a) treadmill dataset, (b) course dataset.

synthesized gait feature with views of the observation azimuth angles of 0, 30, 60, 90, 120, 150, and 180 deg from gait features with views of the observation azimuth angles of 90, 120, 150, and 180 deg associated with the treadmill dataset.

From the figure, it is clear that gait features with different views can be synthesized using the AVTM, and that the synthesized gait images are similar to those of the target views. These results show that the AVTM has the potential to improve the accuracy of cross-view gait recognition.

However, it should be noted that the information included in the gait features depends on the views, and information that is not included cannot be recovered using the view transformation approach. For example, although body width information is included in gait features with a frontal or rear view (e.g., 0 deg or 180 deg in the treadmill dataset), this information is not included in gait features with a side view (e.g., 90 deg in the treadmill dataset). This is one reason why the width of the synthesized gait features with a view of 180 deg from a gait feature with a view of 90 deg (in the first column and the eighth row of Fig. 9(a)) is narrower than that of a gait feature with a frontal view (in the fourth column

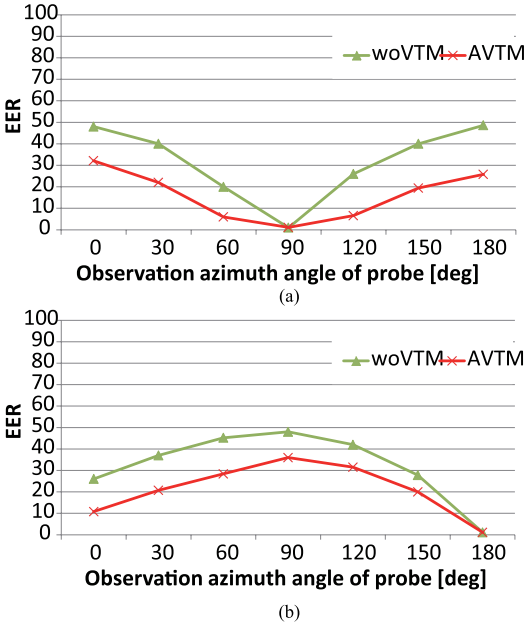

Fig. 10. Equal error rates of different cross-view matching with score normalization: (a) gallery: 90 deg, (b) gallery: 180 deg.

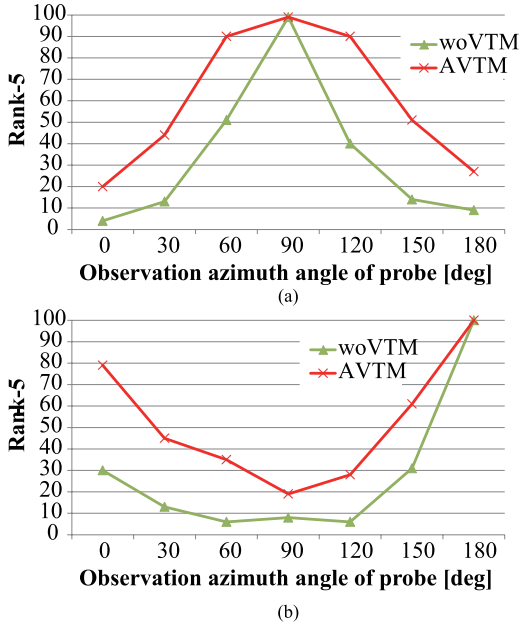

Fig. 11. Rank-5 identification rates of different cross-view matching: (a) Gallery: 90 deg, (b) Gallery: 180 deg.

and the first row of Fig. 9(a)). This observation implies that the recognition accuracy improvement by the AVTM of gait features with a frontal/rear view and those with a side view is small.

2) Main Experiment: We evaluated the accuracy of the proposed AVTM for two different tasks: verification and identification. For the verification task, we used the target-dependent score normalization technique [38] before matching.

For comparison purposes, we also evaluated the accuracy of a method without a view transformation model (woVTM). The woVTM matches gait features from different views directly without any transformation. Because the views of the treadmill dataset are almost the same as the discrete training views, the accuracy of the AVTM is almost the same as that of the DVTM. Therefore, in this experiment we compared the accuracy of the woVTM with that of the AVTM.

For this dataset, we chose gait features with observation azimuth views of 90 and 180 deg for the gallery, and evaluated the recognition accuracy against probe gait features with observation azimuth angles of 0, 30, 60, 90, 120, 150, and 180 deg. We computed equal error rates (EERs) and rank-5 identification rates as the accuracy criteria for the verification task and identification task, respectively. Fig. 10 shows the EERs of different cross-view matching associated with the side and rear views, while Fig. 11 shows rank-5 identification rates (Rank-5) of the different cross-view matching associated with the side and rear views. In these figures, the matching results for the woVTM are plotted together with those of the proposed AVTM. In Figs. 10(a) and 11(a), we plot the EERs and Rank-5 values where gait features with a side view (observation azimuth angle of 90 deg) are used as the gallery. In these figures, we can see that the accuracy of both the woVTM and AVTM deteriorates with an increasing view difference between the gallery and the probe; however, the accuracy of the AVTM is always better than that of the woVTM.

In Figs. 10(b) and 11(b), we plot EERs and Rank-5 values where gait features with a rear view (observation azimuth angle of 180 deg) are used as the gallery. In these figures, the accuracy of both the woVTM and the AVTM deteriorates with an increasing view difference between the gallery and probe; the accuracy is at a minimum for a probe view of 90 deg. Thereafter, accuracy improves with an increasing view difference.

From the viewpoint of view difference from the 180 deg observation angle, a probe gait feature with a 30 deg observation azimuth angle is much larger than that with a 90 deg observation azimuth angle. However, from the viewpoint of the included information in a gait feature, a probe gait feature with a 30 deg observation azimuth angle has more common information with a gait feature with a 180 deg observation azimuth angle than that of a gait feature with a 90 deg observation azimuth angle.

D. Experiment on the Course Dataset

1) Preliminary Experiment: Fig. 9(b) shows the synthesized gait features with different destination views that were synthesized from a gait feature with different views. From this figure, we can confirm that the AVTM can synthesize gait features in the setting where the target views differ from the training views.

We evaluated to what extent the destination view impacts the transformation error using training sets of 3D gait volume sequences. In this experiment, we set the gallery and probe views of the course dataset to $\theta_G = 55$ deg and $\theta_P = 85$ deg, respectively, and considered multiple destination views $\psi = 55, 57, \ldots, 85$ deg at two degree intervals. We generated 2D gait image sequences for these views by

projecting the 3D volume sequences onto a 2D image plane using associated projection matrices, and then generated gait features for these views. The generated gait features were divided into two subsets: a subset for model training (G^T), and a subset for validation (G^V). The subset for model training was used for VTM generation, while that for validation was used to calculate the transformation error. In this experiment, we performed 10-fold cross validation, and evaluated three transformation errors for each part q: the transformation error from gallery (E_q^G), that from probe (E_q^P), and the total transformation error (E_q^T) defined as

$$E_q^G(\psi) = \sum_{\upsilon=1}^{10} \sum_{m \in G_\upsilon^V} \left|\left| M_q \left(R'_{\psi,\upsilon}(\psi) R^{\prime +}_{\psi,\upsilon}(\theta_G) X^m(\theta_G) - X^m(\psi) \right) \right|\right|, \quad (13)$$

$$E_q^P(\psi) = \sum_{\upsilon=1}^{10} \sum_{m \in G_\upsilon^V} \left|\left| M_q \left(R'_{\psi,\upsilon}(\psi) R^{\prime +}_{\psi,\upsilon}(\theta_P) X^m(\theta_P) - X^m(\psi) \right) \right|\right|, \quad (14)$$

$$E_q^T(\psi) = E_q^G(\psi) + E_q^P(\psi). \quad (15)$$

Here, $R'_{\psi,\upsilon}(\theta)$ is the transformation matrix of view θ trained using gait features of view θ_G, θ_P, and a destination view ψ in the υ-th subset for model training G_υ^T. These errors are illustrated in Fig. 12.

From these figures, we observe that the transformation error increases monotonically as the view difference between the original and destination views increases (Fig. 12(a) and (b)). On the other hand, the total transformation error of each part forms the minimum between the probe and gallery views, while the destination view with the lowest error differs for each body part. The experimental results show that the appropriate destination view for each part is different, and part-dependent destination view selection has the potential to improve recognition accuracy.

2) Main Experiment: We evaluated the accuracy of the proposed AVTMs for two different tasks: verification and identification. For the verification task, we evaluated the accuracy in two scenarios: with and without score normalization [38]. We plotted receiver operation characteristic (ROC) curves to compute EERs to evaluate verification accuracy, and cumulative matching characteristic (CMC) curves to compute rank-1 identification rates to evaluate identification accuracy.

In this experiment, we evaluated the accuracy of both the AVTM and AVTM_PdVS. For comparison, we evaluated the state-of-the-art algorithm, RankSVM [31], as well as the DVTM and the woVTM.

Different from the AVTM, applicable views for the DVTM and RankSVM are limited to discrete training views only; in this experiment, the views associated with gait image sequences from 25 cameras used for 3D gait volume generation were used as training views for DVTM and RankSVM. Of the 25 discrete views, we selected two views taken from a height of 130 cm with observation azimuth angles of 60 and 90 deg, respectively, as candidates for the nearest

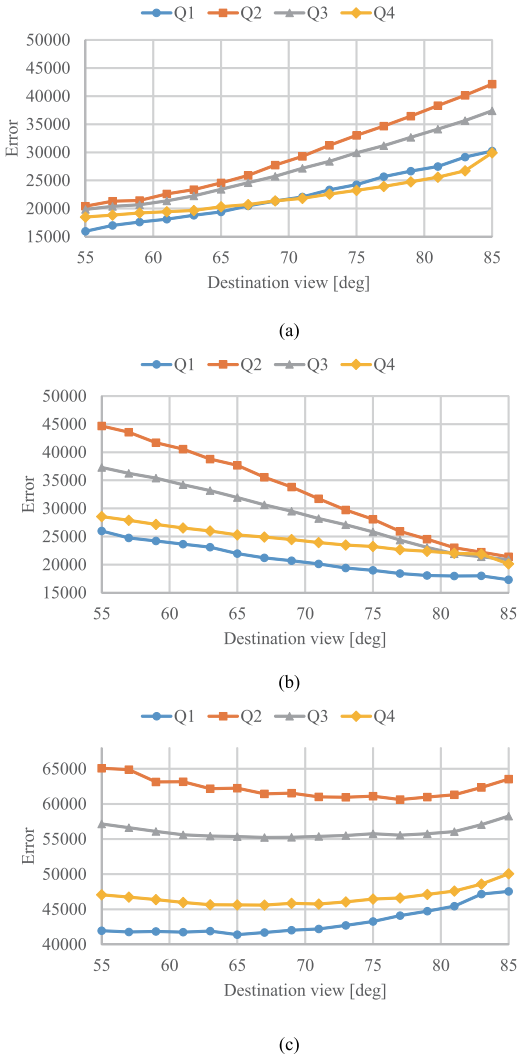

Fig. 12. Impact of the destination view against transformation error. (a) Transformation error of gait features with view $\theta_G = 55$ [deg]. (b) Transformation error of gait features with view $\theta_P = 85$ [deg]. (c) Total transformation error $E_q(\psi)$.

TABLE II

SELECTED DISCRETE TRAINING VIEW (AZIMUTH ANGLE [DEG] AT A HEIGHT OF 130 [CM]) PAIR OF A GALLERY AND A PROBE FOR THE DVTM AND RANKSVM IN EACH SETTING FOR COURSE DATA. NOTE THAT DVTM AND RANKSVM CANNOT BE APPLIED IN SETTINGS MARKED WITH $*$, WHILE HYPHEN (-) DENOTES SAME-VIEW SETTINGS THAT ARE OUT OF SCOPE

Gallery [deg]	55	65	75	85
Probe [deg]	(discrete gallery view, discrete probe view)			
55	-	(60, 60)$*$	(90, 60)	(60, 90)
65	(60, 60)$*$	-	(90, 60)	(60, 90)
75	(60, 90)	(60, 90)	-	(90, 60)
85	(60, 90)	(60, 90)	(60, 90)	-

training views. Gait features with training views and test views are illustrated in Fig. 8. The training views used for the DVTM and RankSVM evaluations are summarized in Table II. In the table, only a single pair of observation azimuth angles are

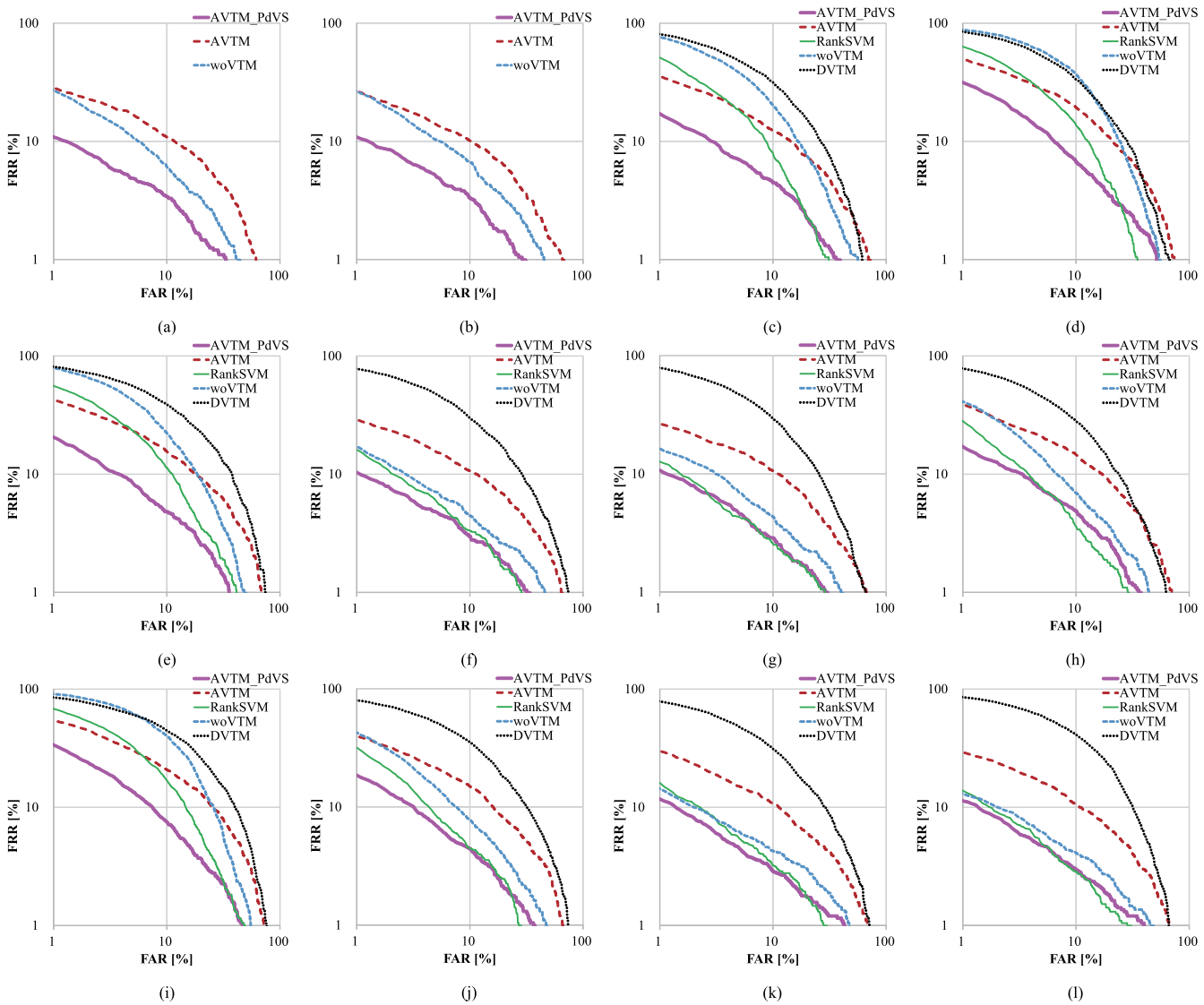

Fig. 13. ROC curves in different cross-view settings; (G, P) = (θ_G, θ_P) means that the observation angles of the gallery and probe are θ_G and θ_P, respectively. (a) (G, P) = (55, 65). (b) (G, P) = (65, 55). (c) (G, P) = (75, 55). (d) (G, P) = (85, 55). (e) (G, P) = (55, 75). (f) (G, P) = (65, 75). (g) (G, P) = (75, 65). (h) (G, P) = (85, 65). (i) (G, P) = (55, 85). (j) (G, P) = (65, 85). (k) (G, P) = (75, 85). (l) (G, P) = (85, 75).

described.[5] Note that for settings (θ_G, θ_P) = (55,65) and (θ_G, θ_P) = (65,55), the corresponding discrete training views of the gallery and the probe are the same as those for 60 deg, and hence, applying DVTM or RankSVM has no effect, resulting in the same accuracy as woVTM. Thus, we did not evaluate the accuracy of these settings.

In all the experiments, we consistently used FDFs [4] as gait features.

Figs. 13, 14, and 15 show, respectively, the ROC curves without score normalization, ROC curves with score normalization, and CMC curves for the AVTM_PdVS, AVTM, RankSVM, DVTM, and woVTM. In these figures, the caption (G,P) = (θ_G,θ_P) means that the observation angles of the gallery and probe are θ_G and θ_P, respectively. EERs with and without score normalization, and Rank-1 values of the considered methods are summarized in Tables III, IV, and V, respectively.[6]

From Figs. 13, 14, and 15, and Tables III, IV, and V, we can see that the AVTM outperforms the DVTM in all cross-view settings. This is because the proposed framework of the AVTM can solve the accuracy degradation caused by view difference between the target and training views.

In the verification tasks with score normalization and identification tasks, the accuracy of the AVTM is also better than that of the woVTM for many considered view pairs; however, when the view difference is small, the AVTM accuracy is sometimes slightly worse than that of the woVTM. We think that the two factors are related to the results: generalization error of view transformation and accuracy degradation caused

[5]Since 60 and 90 deg training azimuth angles are the same distance from the 75 deg azimuth angle of the course dataset, we switched the corresponding discrete view to either 60 or 90 deg flexibly so that the corresponding discrete views of the gallery and probe views were different.

[6]Diagonal parts of these tables correspond to same-view settings. Since accuracy evaluation in these settings is beyond the scope of this paper, we do not compare the accuracy in these settings. Instead, for reference, we show the original accuracy, enclosed in parentheses, of a gait recognition approach using FDFs in the associated view settings.

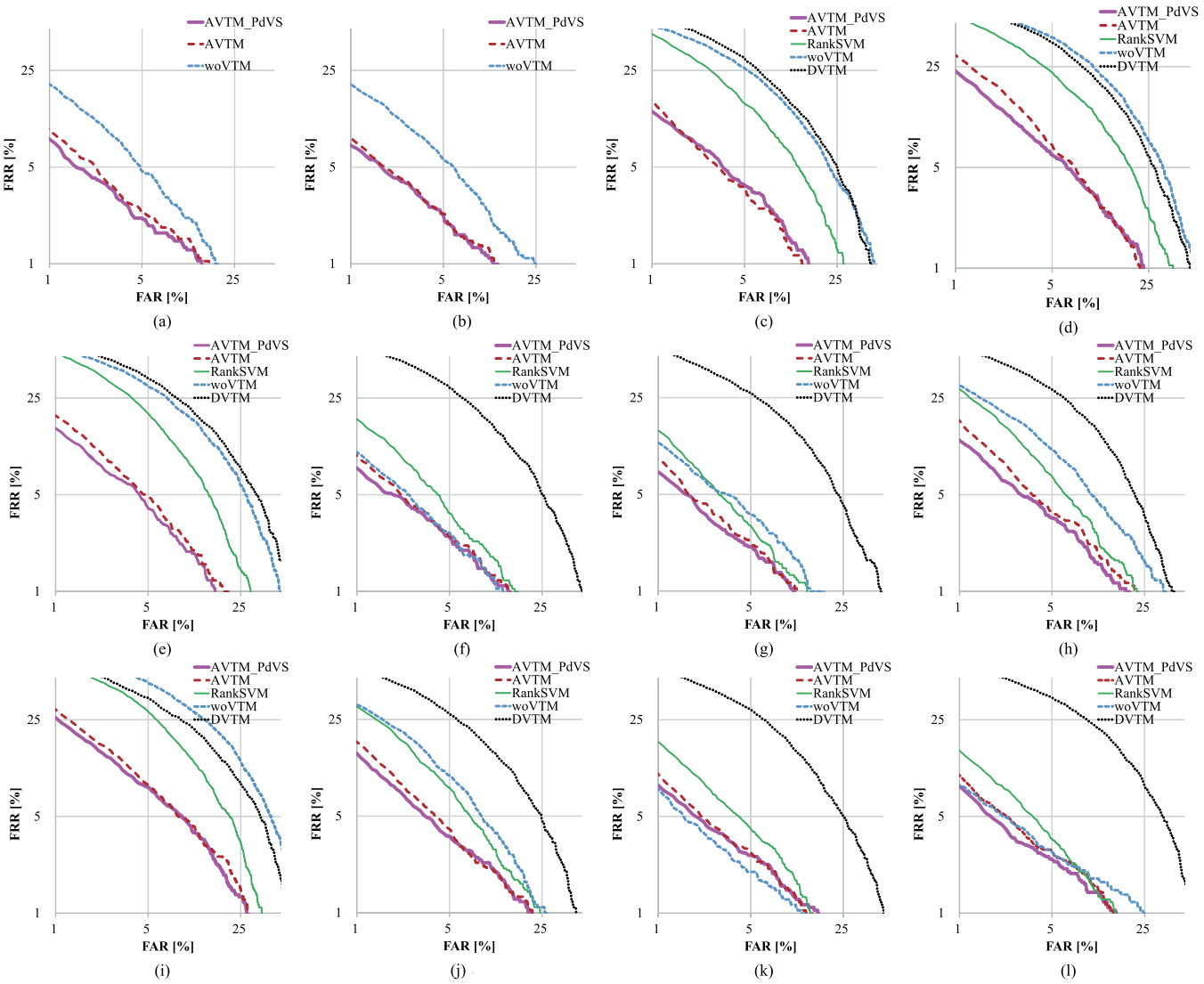

Fig. 14. ROC curves after target-dependent score normalization in different cross-view settings; (G, P) = (θ_G, θ_P) means that the observation angles of the gallery and probe are θ_G and θ_P, respectively. (a) (G, P) = (55, 65). (b) (G, P) = (65, 55). (c) (G, P) = (75, 55). (d) (G, P) = (85, 55). (e) (G, P) = (55, 75). (f) (G, P) = (65, 75). (g) (G, P) = (75, 65). (h) (G, P) = (85, 65). (i) (G, P) = (55, 85). (j) (G, P) = (65, 85). (k) (G, P) = (75, 85). (l) (G, P) = (85, 75).

by appearance change. Because of the generalization error, the AVTM cannot synthesize the same gait feature with a target view from a gait feature with a different view from test subjects who are not included in the training data. Therefore, the cross-view recognition accuracy with the VTM is essentially degraded from that of the same view. We call this degradation *generalization error-related degradation*. Conversely, cross-view recognition accuracy with direct matching (woVTM) is also degraded from that of the same view, because the appearance of the compared gait features differs. We call this degradation *appearance-related degradation*. When appearance variations caused by view differences are small, the appearance-related degradation is smaller than the generalization error-related one; therefore, recognition accuracy in the woVTM is sometimes better than that in the AVTM. However, when appearance changes caused by view differences are large, the appearance-related degradation is much larger than the generalization error-related one. Therefore, the AVTM achieves better accuracy than the woVTM in these situations.

The AVTM also achieves better accuracy than the RankSVM in the verification tasks with score normalization and identification tasks owing to the arbitrary nature of the proposed framework. As reported in [31], RankSVM achieves a high accuracy in a cross-view setting where the target views and training views are the same, but it has difficulty in achieving high accuracy in a cross-view setting where the target and training views differ.

However, the accuracy of the AVTM is always worse than that of woVTM and RankSVM in verification tasks without score normalization. Because an AVTM is generated using a limited number of training data, fitting to the trained AVTM must be different for each test subject, and this tends to produce dissimilarity scores including large subject-dependent bias. This subject-dependent bias results in lower accuracy of the AVTM in verification tasks without score normalization even though the AVTM achieves high accuracy for identification tasks. In contrast, the AVTM_PdVS achieves high accuracy even in this task. The AVTM_PdVS selects a

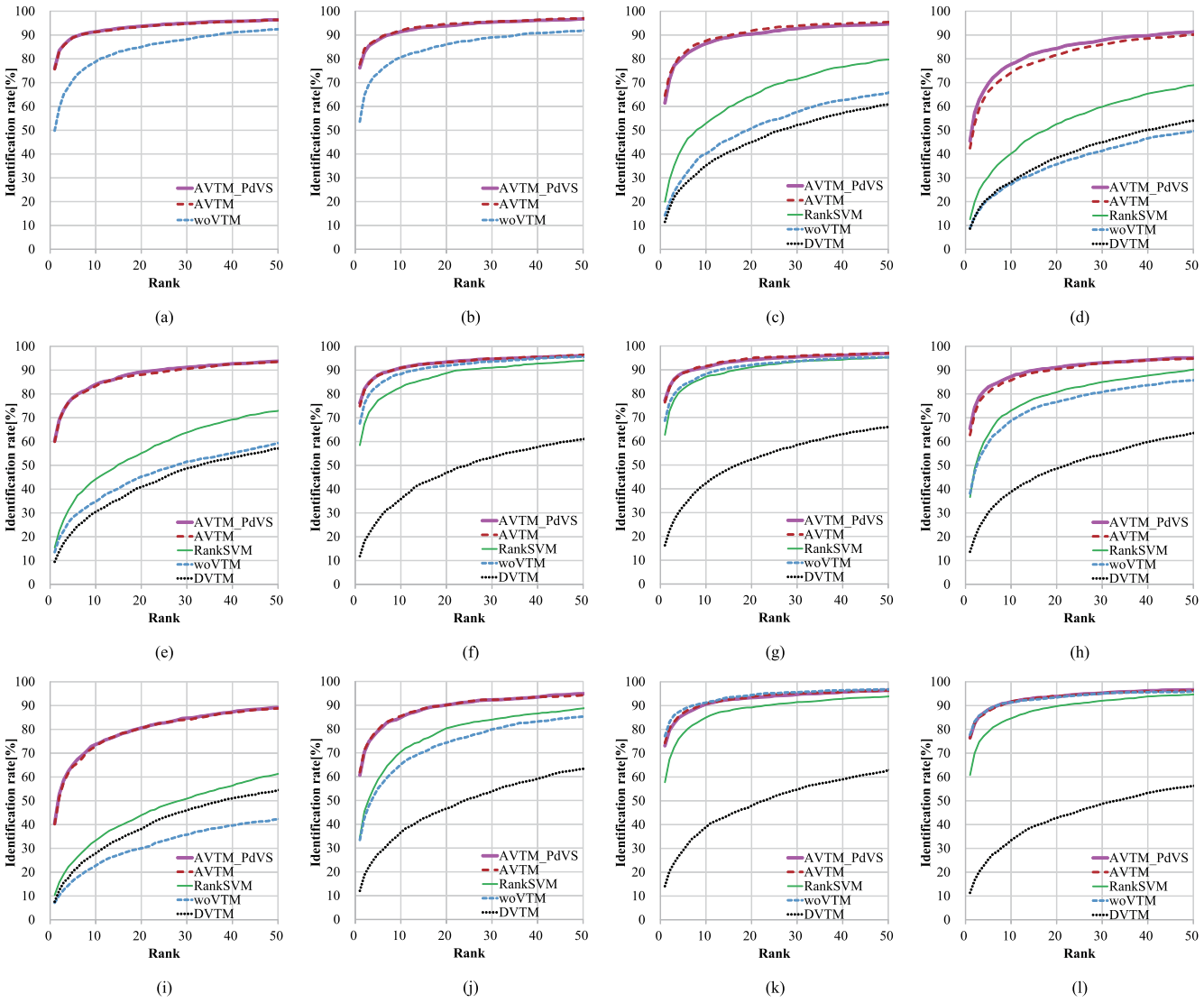

Fig. 15. CMC curves in different cross-view settings; (G, P) = (θ_G, θ_P) means that the observation angles of the gallery and probe are θ_G and θ_P, respectively. (a) (G, P) = (55, 65). (b) (G, P) = (65, 55). (c) (G, P) = (75, 55). (d) (G, P) = (85, 55). (e) (G, P) = (55, 75). (f) (G, P) = (65, 75). (g) (G, P) = (75, 65). (h) (G, P) = (85, 65). (i) (G, P) = (55, 85). (j) (G, P) = (65, 85). (k) (G, P) = (75, 85). (l) (G, P) = (85, 75).

destination view that achieves the lowest transformation error for each body part; this selection has the effect of reducing subject-dependent bias, which leads to improved accuracy in the task. The AVTM_PdVS is also superior to or comparable with other benchmarks in other tasks. In particular, the proposed AVTM_PdVS achieves higher accuracy in cases where the view difference is large. This is because the appropriate destination view for each body part is largely different for the gallery and probe views; that is, the destination views in the original AVTM, as shown in Fig. 12, and the AVTM_PdVS can select a reasonable destination view for each body part.

VI. Discussion

A. Applying the Arbitrary-View Framework to Other Approaches

The experimental results show that the proposed arbitrary-view framework can be efficiently applied to the VTM to generate the AVTM, which achieves higher accuracy than the DVTM and RankSVM, which are trained using data for discrete training views. In this subsection, we show the applicability of the proposed arbitrary-view framework to other approaches. We evaluate the accuracy of RankSVM with our arbitrary-view framework as the arbitrary RankSVM (A-RankSVM), and compare the accuracy with the conventional RankSVM using discrete training data (D-RankSVM). The course dataset was used for the evaluation, with the recognition accuracy for a probe view of 85 deg and gallery views of 55, 65, and 75 deg summarized in Table VI.

From these results, we confirm that the arbitrary-view framework improves the accuracy of the RankSVM since the proposed arbitrary-view framework generates training data with exactly the same views as the target views. Consequently, we can say that the arbitrary-view framework improves accuracy even for other approaches besides the VTM.

B. Recognition Accuracy for Each Part

To analyze the spatial property of view transformation, we evaluated recognition accuracy for each body part.

TABLE III

EQUAL ERROR RATES IN THE VERIFICATION TASK WITHOUT SCORE NORMALIZATION. THE BOLD VALUES INDICATE THE BEST ACCURACY, WHILE VALUES IN ITALICS INDICATE THE SECOND BEST ACCURACY FOR EACH SETTING. NOTE THAT HYPHEN (-) AND N/A DENOTE, RESPECTIVELY, OUT OF SCOPE AND SETTINGS IN WHICH DISCRETE APPROACHES CANNOT BE APPLIED

Probe	Gallery	55	65	75	85
55	AVTM_PdVS	-	**5.07**	**6.17**	**8.05**
	AVTM	-	10.09	11.56	14.38
	RankSVM	-	N/A	*9.08*	*11.45*
	DVTM	-	N/A	17.73	17.99
	woVTM	(5.86)	*7.88*	13.77	17.73
65	AVTM_PdVS	**4.88**	-	*4.81*	**6.38**
	AVTM	10.62	-	10.46	12.34
	RankSVM	N/A	-	**4.39**	*6.43*
	DVTM	N/A	-	16.99	16.43
	woVTM	*7.74*	(5.20)	6.06	
75	AVTM_PdVS	**6.55**	**4.83**	-	**4.81**
	AVTM	13.08	10.46	-	10.30
	RankSVM	*10.51*	*5.49*	-	*4.94*
	DVTM	21.31	18.26	-	21.08
	woVTM	14.33	6.17	(5.18)	
85	AVTM_PdVS	**8.53**	**6.00**	**4.71**	-
	AVTM	15.51	12.81	10.51	-
	RankSVM	*12.68*	*6.64*	*5.28*	-
	DVTM	22.60	18.83	17.68	-
	woVTM	18.62	8.72	5.75	(4.97)

TABLE IV

EQUAL ERROR RATES IN THE VERIFICATION TASK WITH SCORE NORMALIZATION. THE BOLD VALUES INDICATE THE BEST ACCURACY, WHILE VALUES IN ITALICS INDICATE THE SECOND BEST ACCURACY FOR EACH SETTING. NOTE THAT HYPHEN (-) AND N/A DENOTE, RESPECTIVELY, OUT OF SCOPE AND SETTINGS IN WHICH DISCRETE APPROACHES CANNOT BE APPLIED

Probe	Gallery	55	65	75	85
55	AVTM_PdVS	-	**3.19**	*4.13*	**5.70**
	AVTM	-	*3.26*	**3.98**	*5.91*
	RankSVM	-	N/A	8.47	10.73
	DVTM	-	N/A	12.71	14.39
	woVTM	(2.33)	5.39	11.97	15.80
65	AVTM_PdVS	**3.14**	-	**2.91**	**4.09**
	AVTM	*3.32*	-	*3.19*	*4.26*
	RankSVM	N/A	-	3.80	5.70
	DVTM	N/A	-	12.55	12.35
	woVTM	4.84	(2.34)	4.18	7.53
75	AVTM_PdVS	**4.60**	**3.35**	-	**3.14**
	AVTM	*4.96*	3.53	-	*3.41*
	RankSVM	9.39	4.45	-	4.18
	DVTM	15.15	12.97	-	15.74
75	woVTM	13.55	*3.50*	(2.30)	3.55
85	AVTM_PdVS	**6.43**	**4.08**	*3.37*	-
	AVTM	*6.54*	*4.45*	3.45	-
	RankSVM	11.65	6.14	4.51	-
	DVTM	15.38	12.45	15.74	-
85	woVTM	18.53	7.05	**2.97**	(2.41)

For this analysis, we used the course dataset, and focused on a setting with gallery and probe views of 55 deg and 85 deg, respectively. We then evaluated the recognition accuracy of the AVTM, the AVTM with destination view selection (AVTM_VS), and the woVTM in verification and identification tasks. The results are summarized in Table VII. From the experiments, we observe that the efficiency of the AVTM/AVTM_VS is highly dependent on the body part.

TABLE V

RANK-1 IDENTIFICATION RATES. THE BOLD VALUES INDICATE THE BEST ACCURACY, WHILE VALUES IN ITALICS INDICATE THE SECOND BEST ACCURACY FOR EACH SETTING. NOTE THAT HYPHEN (-) AND N/A DENOTE, RESPECTIVELY, OUT OF SCOPE AND SETTINGS IN WHICH DISCRETE APPROACHES CANNOT BE APPLIED

Probe	Gallery	55	65	75	85
55	AVTM_PdVS	-	*76.20*	*61.45*	**45.50**
	AVTM	-	**77.72**	**64.54**	*42.69*
	RankSVM	-	N/A	19.98	12.60
	DVTM	-	N/A	11.51	8.69
	woVTM	(86.66)	53.61	14.28	8.94
65	AVTM_PdVS	**75.99**	-	**77.09**	**65.48**
	AVTM	*75.63*	-	*76.36*	*62.76*
	RankSVM	N/A	-	62.71	36.66
	DVTM	N/A	-	16.27	13.70
	woVTM	49.84	(87.34)	68.62	38.18
75	AVTM_PdVS	**60.25**	**76.20**	-	*76.52*
	AVTM	*59.88*	*74.90*	-	76.31
	RankSVM	15.49	58.47	-	60.83
	DVTM	9.47	11.82	-	11.35
	woVTM	13.54	67.63	(87.92)	**77.72**
85	AVTM_PdVS	**40.48**	*60.62*	73.12	-
	AVTM	*40.17*	**61.87**	*74.32*	-
	RankSVM	10.41	34.41	57.79	-
	DVTM	7.58	12.03	14.12	-
	woVTM	7.16	33.42	**77.14**	(85.93)

TABLE VI

ACCURACY COMPARISON OF THE ARBITRARY RANKSVM AND CONVENTIONAL RANKSVM

View pair (θ_G, θ_P)	Method	EER wSN [%]	EER woSN[%]	Rank-1 [%]
(55, 85)	D-RankSVM	12.68	11.65	10.41
	A-RankSVM	10.62	8.67	28.97
	Improvement	2.06	2.98	18.56
(65, 85)	D-RankSVM	6.64	6.14	34.41
	A-RankSVM	6.42	5.25	60.77
	Improvement	0.22	0.89	26.36
(75, 85)	D-RankSVM	5.28	4.51	57.79
	A-RankSVM	4.45	3.24	77.04
	Improvement	0.83	1.27	19.25

TABLE VII

RECOGNITION ACCURACY OF EACH PART IN THE SETTING $(\theta_G, \theta_P) = (55, 85)$. THE BOLD VALUES INDICATE THE BEST ACCURACY, WHILE VALUES IN ITALICS INDICATE THE SECOND BEST ACCURACY FOR EACH SETTING

Part	Method	EER wSN [%]	EER woSN[%]	Rank-1 [%]
Q1	woVTM	*22.11*	**18.49**	**7.37**
	AVTM	24.41	20.50	2.77
	AVTM_VS	**21.55**	*19.93*	*7.11*
Q2	woVTM	38.13	37.39	1.36
	AVTM	*24.74*	**17.83**	*5.18*
	AVTM_VS	**21.03**	*18.57*	**8.37**
Q3	woVTM	*19.21*	17.86	5.23
	AVTM	19.94	*12.37*	*10.41*
	AVTM_VS	**13.18**	**10.48**	**20.29**
Q4	woVTM	29.80	28.71	1.83
	AVTM	*21.98*	*12.94*	*10.83*
	AVTM_VS	**10.46**	**9.57**	**23.85**

For Q1 (head region), woVTM achieves higher accuracy than the AVTM/AVTM_VS. This is because the shape of the head region is almost round and hence, its appearance does not change that much under view differences. In contrast, the

AVTM/AVTM_VS achieves higher accuracy than the woVTM for the other parts. WoVTM for Q2 (chest and arm region) yields the worst results of all the regions and the benchmarks. A possible reason for this poor accuracy may be related to arm swing. The gait feature associated with arm swing changes markedly with a view change, and hence, direct matching results in poor accuracy. However, this poor accuracy is improved by applying the AVTM/AVTM_VS. For Q3 and Q4, the AVTM_VS achieves the best accuracy of the three methods. These parts include the leg/foot region, which has efficient dynamic information associated with its movement during walking. These parts are very important for recognition, whereas gait features associated with these parts are easily affected by a view change. Therefore, the accuracy of the woVTM for Q4 is lower than that for the head region despite achieving high accuracy for same-view settings. However, by applying the AVTM_VS, accuracy is improved and the best accuracy is achieved. The AVTM_VS dramatically improves recognition accuracy in regions that include view-variant dynamic information.

VII. Conclusion

In this paper, we proposed an AVTM for cross-view gait recognition. Using this arbitrary-view framework, we eliminated the discrete nature of previously proposed VTMs, and generated an AVTM. Through the experiments, we showed that the AVTM achieves higher accuracy than the DVTM, woVTM, and RankSVM in verification tasks with score normalization and identification tasks. Moreover, we extended our AVTM to the AVTM_PdVS by incorporating a PdVS scheme based on the observation that transformation errors are dependent not only on body parts, but also on destination views. While the performance of the AVTM was sometimes worse than that of RankSVM and woVTM in verification tasks without score normalization because of the inhomogeneous subject-dependent bias of the dissimilarity scores, the AVTM_PdVS achieved higher accuracy than the comparative methods including the AVTM in most of the settings for all tasks.

Moreover, we also showed that the proposed arbitrary-view framework improves the accuracy of other approaches such as RankSVM, which indicates wider applicability of the arbitrary-view framework. Note that the proposed AVTM_PdVS still achieves higher accuracy than the arbitrary-view version of RankSVM in many settings, which confirms that the proposed AVTM_PdVS is a promising approach for cross-view gait recognition.

References

[1] M. S. Nixon, T. Tan, and R. Chellappa, *Human Identification Based on Gait* (International Series on Biometrics). New York, NY, USA: Springer-Verlag, Dec. 2005.

[2] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. New York, NY, USA: Springer-Verlag, 2008.

[3] A. Kale, A. K. Roy-Chowdhury, and R. Chellappa, "Towards a view invariant gait recognition algorithm," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Jul. 2003, pp. 143–150.

[4] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Proc. 9th Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, pp. 151–163.

[5] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition under various viewing angles based on correlated motion regression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 6, pp. 966–980, Jun. 2012.

[6] Y. Makihara, A. Tsuji, and Y. Yagi, "Silhouette transformation based on walking speed for gait identification," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 717–722.

[7] M. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi, "Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control," *Pattern Recognit.*, vol. 43, no. 6, pp. 2281–2291, Jun. 2010.

[8] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.

[9] 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.

[10] 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, Apr. 2011.

[11] A. F. Bobick and A. Y. Johnson, "Gait recognition using static, activity-specific parameters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Dec. 2001, pp. I-423–I-430.

[12] C. Yam, M. Nixon, and J. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, 2004.

[13] H.-D. Yang and S.-W. Lee, "Reconstruction of 3D human body pose for gait recognition," in *Proc. IAPR Int. Conf. Biometrics*, Jan. 2006, pp. 619–625.

[14] M. Goffredo, I. Bouchrika, J. N. Carter, and M. S. Nixon, "Self-calibrating view-invariant gait biometrics," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 997–1008, Aug. 2010.

[15] G. Ariyanto and M. S. Nixon, "Marionette mass-spring model for 3D gait biometrics," in *Proc. 5th IAPR Int. Conf. Biometrics*, Mar./Apr. 2012, pp. 354–359.

[16] R. Urtasun and P. Fua, "3D tracking for gait characterization and recognition," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2004, pp. 17–22.

[17] G. Shakhnarovich, L. Lee, and T. Darrell, "Integrated face and gait recognition from multiple views," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Dec. 2001, pp. I-439–I-446.

[18] L. Lee, "Gait analysis for classification," Ph.D. dissertation, Dept. Elect. Eng., Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2002.

[19] R. Bodor, A. Drenner, D. Fehr, O. Masoud, and N. Papanikolopoulos, "View-independent human motion classification using image-based reconstruction," *Image Vis. Comput.*, vol. 27, no. 8, pp. 1194–1206, Jul. 2009.

[20] Y. Iwashita, R. Baba, K. Ogawara, and R. Kurazume, "Person identification from spatio-temporal 3D gait," in *Proc. Int. Conf. Emerg. Secur. Technol.*, Sep. 2010, pp. 30–35.

[21] F. Jean, R. Bergevin, and A. B. Albu, "Computing and evaluating view-normalized body part trajectories," *Image Vis. Comput.*, vol. 27, no. 9, pp. 1272–1284, Aug. 2009.

[22] J. Han, B. Bhanu, and A. Roy-Chowdhury, "A study on view-insensitive gait recognition," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3. Sep. 2005, pp. III-297–III-300.

[23] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Sep./Oct. 2009, pp. 1058–1064.

[24] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Multi-view gait recognition based on motion regression using multilayer perceptron," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2186–2189.

[25] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 974–981.

[26] D. Muramatsu, A. Shiraishi, Y. Makihara, and Y. Yagi, "Arbitrary view transformation model for gait person authentication," in *Proc. IEEE 5th Int. Conf. Biometrics, Theory, Appl. Syst.*, Sep. 2012, pp. 85–90.
[27] D. Muramatsu, Y. Makihara, and Y. Yagi, "Are intermediate views beneficial for gait recognition using a view transformation model?" in *Proc. 20th Korea-Japan Joint Workshop Frontiers Comput. Vis.*, 2014, pp. 222–227.
[28] 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.
[29] J. Lu and Y.-P. Tan, "Uncorrelated discriminant simplex analysis for view-invariant gait signal computing," *Pattern Recognit. Lett.*, vol. 31, no. 5, pp. 382–393, 2010.
[30] N. Liu, J. Lu, and Y.-P. Tan, "Joint subspace learning for view-invariant gait recognition," *IEEE Signal Process. Lett.*, vol. 18, no. 7, pp. 431–434, Jul. 2011.
[31] R. Martín-Félez and T. Xiang, "Gait recognition by ranking," in *Computer Vision* (Lecture Notes in Computer Science), vol. 7572, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Germany: Springer-Verlag, 2012, pp. 328–341.
[32] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with SVMs," *Inf. Retr.*, vol. 13, no. 3, pp. 201–215, Jun. 2010.
[33] A. Utsumi and N. Tetsutani, "Adaptation of appearance model for human tracking using geometrical pixel value distributions," in *Proc. 6th Asian Conf. Comput. Vis.*, vol. 2. 2004, pp. 794–799.
[34] Y. Makihara and Y. Yagi, "Silhouette extraction based on iterative spatio-temporal local color transformation and graph-cut segmentation," in *Proc. 19th Int. Conf. Pattern Recognit.*, Tampa, FL, USA, Dec. 2008, pp. 1–4.
[35] W. T. Dempster and G. R. L. Gaughran, "Properties of body segments based on size and weight," *Amer. J. Anatomy*, vol. 120, no. 1, pp. 33–54, 1967.
[36] R. Sagawa, M. Takatsuji, T. Echigo, and Y. Yagi, "Calibration of lens distortion by structured-light scanning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Edmonton, AB, Canada, Aug. 2005, pp. 832–837.
[37] W. N. Martin and J. K. Aggarwal, "Volumetric descriptions of objects from multiple views," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 2, pp. 150–158, Mar. 1983.
[38] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Target dependent score normalization techniques and their application to signature verification," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 3, pp. 418–425, Aug. 2005.

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 Specially Appointed Associate Professor with the Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan. His research interests are gait recognition, signature verification, and biometric fusion. He is also a member of the Institute of Electronics, Information and Communication Engineers, the Information Processing Society of Japan, and the Institute of Image Electronics Engineers of Japan.

Akira Shiraishi received the B.S. degree in computer and communication science from Wakayama University, Wakayama, Japan, in 2009, and the M.S. degree in information science from Osaka University, Osaka, Japan, in 2011.

Yasushi Makihara received the B.S., M.S., and Ph.D. degrees in engineering from Osaka University, Osaka, Japan, in 2001, 2002, and 2005, respectively, where he is currently an Assistant Professor with the Institute of Scientific and Industrial Research. His research interests are gait recognition, morphing, and temporal super resolution. He is also a member of the Information Processing Society of Japan, the Robotics Society of Japan, and the Japan Society of Mechanical Engineers.

Md. Zasim Uddin received the B.Sc. and M.Sc. degrees from the Department of Computer Science and Engineering, Rajshahi University, Rajshahi, Bangladesh, in 2007 and 2008, respectively. He is currently pursuing the Ph.D. degree with the Department of Intelligent Media, Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan. His research interests include computer vision, medical imaging, machine learning, and gait and action recognition.

Yasushi Yagi is currently the Director of the Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan, where he received the Ph.D. degree, in 1991. In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, Tokyo, Japan, where he was involved in robotics and inspections. He became a Research Associate in 1990, a Lecturer in 1993, an Associate Professor in 1996, and a Professor in 2003 with Osaka University. He served as the Chair of the international conferences, including 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 served as an Editor of the IEEE ICRA Conference Editorial Board (2007–2011). He is an Editorial Member of *International Journal of Computer Vision* and the Editor-in-Chief of the *IPSJ Transactions on Computer Vision and Applications*. He was a recipient of the ACM VRST2003 Honorable Mention Award, the IEEE ROBIO2006 Finalist of T. J. Tan Best Paper in Robotics, the IEEE ICRA2008 Finalist for Best Vision Paper, the MIRU2008 Nagao Award, and the PSIVT2010 Best Paper Award. His research interests are computer vision, medical engineering, and robotics. He is also a fellow of the Information Processing Society of Japan and member of the Institute of Electronics, the Information and Communication Engineers, and the Robotics Society of Japan.