Geometrically Consistent Pedestrian Trajectory Extraction for Gait Recognition

Yasushi Makihara, Gakuto Ogi, Yasushi Yagi
Osaka University
{makihara, ogi, yagi}@am.sanken.osaka-u.ac.jp

Abstract

In the gait recognition community, silhouette-based gait representations such as gait energy image have been widely employed for the last decade. In order to obtain good quality of gait features, it is essential to get a well aligned silhouette sequence, which is, however, not necessarily easy with imperfect silhouettes and also with scale and position changes due to perspective projection. We therefore propose a gait recognition-oriented approach to pedestrian trajectory extraction, i.e., bounding box sequence for each pedestrian. More specifically, we firstly developed an interactive tool to get camera calibration parameters for a target scene geometry without on-site workload. We then introduce a geometric constraint to better keep the consistency of bounding boxes among frames w.r.t. the pedestrian's height and foot bottom points on the ground plane. Subsequently, we apply analytical dynamic programing (DP) repeatedly to find multiple pedestrian's trajectories on the ground plane, where data and transition scores are computed based on semantic segmentation results and color histogram similarity. Moreover, since DP just considers the smoothness between adjacent frames, we approximate the trajectory by a piece-wise linear trajectory to make it more globally smooth. Experimental results show that the proposed method enables us to make better aligned gait features and consequently improves gait recognition accuracy.

1. Introduction

Gait [43] is a behavioral biometric that has advantages over other biometrics such as the face, irises, or finger veins because (i) gait is available even when the subject is at a distance from a camera because it can be recognized from a relatively low-resolution image sequence [41], and (ii) a gait feature can be obtained without subject cooperation because people unconsciously exhibit their own walking styles in general. Because of these advantages, gait recognition is suitable for many potential applications such as surveillance, forensics, and criminal investigation [7, 16, 32].

Gait recognition enjoys a rich body of literature mainly from two streams: model-based approaches [2, 6, 8, 25, 56, 57, 63, 65] and model-free (appearance-based) approaches [3, 4, 11, 19, 35, 42, 45, 58, 60]. Gait recognition researchers have geared to the appearance-based approaches for the last decade because they work well even for relatively low-resolution images and incur low computational costs, while the model-based approaches generally suffer from the difficulty of human model fitting and high computational cost. In particular, silhouette-based representations such as gait energy images (GEIs) [11], frequency-domain features (FDFs) [35], chrono-gait images [58], and Gabor GEIs [53], are dominant in the gait recognition community because of their simple yet effective properties. Moreover, gait recognition researchers have paid their effort to gain the robustness against various covariates such as clothing [3, 4, 15, 26, 44], carrying status [9, 52, 54], view [17, 20, 22, 24, 31, 35, 47–49, 62], and walking speed [1, 10, 21, 23, 30, 37, 38, 50].

(a) (b) (c) (d)

Figure 1. Bounding boxes and constructed GEIs. (a), (b): bounding boxes of a pedestrian for adjacent frames by a conventional tracking method, (c), (d): GEIs constructed by the conventional tracking method and the proposed method, respectively. We can see that GEI by the conventional tracking method (c) is blurry, while that by the proposed method (d) is well aligned.

In order to accomplish a gait recognition procedure, it is naturally required not only feature extraction and matching but also pedestrian detection, tracking, and segmentation (i.e., silhouette extraction). The above mentioned gait recognition studies, however, mainly focus only on the feature extraction and matching parts and overlooked the effect of the pre-processing parts such as pedestrian detection, tracking, and segmentation. More specifically, the previous methods assume relatively simple settings for

978-1-5386-7180-1/18/$31.00

pre-processing, i.e., a single pedestrian in a scene, well-designed background for easy silhouette extraction such as a chroma-key background which are also seen in the publicly available gait databases [13, 34, 45, 64]. As such, relatively simple pre-processes, e.g., background subtraction and max region filtering, will do, and some works such as recent deep learning-based approaches [48, 49, 62] even start from extracted gait features such as GEI [11]. There is therefore still a large gap between gait recognition research and practical use in real scenarios.

There are some studies on the pre-processing parts for gait recognition, particularly for silhouette extraction part, since it is essential for the quality of the silhouette-based gait representation such as GEI [11], although the number of such studies are limited. Liu et al. [29] and Hofmann et al. [14] proposed silhouette refinement methods by a population hidden Markov model and alpha matting process, respectively, while Wang et al. [59] and Makihara et al. [36] proposed graph cut-based silhouette extraction using standard gait models. In addition, Matovski et al. [39] introduced a silhouette quality to judge whether the system proceeds to recognition step or continue to capture a gait sequence until getting better silhouettes. Moreover, recent excellent studies on deep learning-based semantic segmentation such as RefineNet [27] and Mask R-CNN [12] could be exploited as a silhouette extraction module as well as a pedestrian detection module for gait recognition.

As for tracking modules, many studies have been done based on analytical dynamic programming (DP) [55], lifted multi-cut model [51], circulant feature maps [61], and super pixel-based video segmentation [40]. While they may be accurate enough for many applications such as surveillance, intrusion detection, etc., they may be insufficient for gait recognition, in particular, when relying on aligned and size-normalized silhouette sequence such as GEI [11] a.k.a. an averaged silhouette over one gait cycle [28]. Assume that a bounding box sequence for adjacent frames is obtained as depicted in Fig. 1 (a)(b) for examples and that we employ GEI [11] as a gait feature. This seems good enough at a glance but we notice that the bottom of the bounding box is inconsistent between the adjacent frames. If we align and size-normalize silhouettes based on such bounding boxes, we may obtain a blurry GEI as shown in Fig. 1 (c), which results in low gait recognition accuracy. In fact, what we need for better gait recognition is a more consistent bounding box sequence among frames in order to get better aligned GEI as shown in Fig. 1 (d).

We therefore propose a gait recognition-oriented approach to pedestrian trajectory extraction under scenes where multiple pedestrian walk by changing their foot bottom points and apparent heights due to perspective projection. For this purpose, we introduce a constraint by a scene geometry to better keep the consistency of bounding boxes among frames and apply DP multiple times to find multiple pedestrian's trajectory on the ground plane. The contributions of this work are three-fold.

1. Geometrically consistent pedestrian trajectory extraction. The proposed method extracts a geometrically consistent pedestrian trajectory (i.e., a bounding box sequence) by imposing a constant height in the world coordinate for each pedestrian and by approximating a trajectory in the spatio-temporal three-dimensional domain (two dimensions of the ground plane and one dimension of the time) as piece-wise linear segments. Unlike the most of general tracking approaches work on an image plane and hence may induce inconsistency in scale and foot bottom/head top positions, the proposed method keeps the better consistency in the scale and the positions.

2. An interactive camera calibration tool without on-site workload. A camera calibration process is essential for the proposed geometrically consistent pedestrian trajectory extraction but troublesome on the other hand since it generally requires to capture some calibration targets on-site. We therefore developed an interactive tool to calibrate a camera based on three orthogonal sets of parallel lines from a captured image itself without on-site workload. The camera calibration results rely on human perception on the three orthogonal sets of parallel lines in the scene geometry, and hence it may not be accurate enough for some applications such as three-dimensional reconstruction but still sufficient for the above mentioned geometric constrains for pedestrian trajectory extraction.

3. Gait recognition accuracy improvement by better aligned gait feature. We can construct a better aligned silhouette sequence by the proposed method, and hence obtain well-aligned gait features. This yields higher gait recognition accuracy compared with the gait features extracted from bounding box sequences by conventional general tracking methods.

2. Geometrically consistent pedestrian trajectory extraction

2.1. Overview

In this subsection, we give an overview of the proposed method along with Fig. 2. Given an original image, we obtain camera calibration parameters by the developed interactive tool and then convert the original images into normalized images based on the camera calibration parameters. Next, we extract body parts segments from the normalized images. We then extract a pedestrian trajectory by DP in the spatio-temporal three-dimensional space (two dimensions of the ground plane and one dimension of time), where transition score and data score are computed from the normalized images and the body parts segments, respectively. Subsequently, the extracted trajectory is deleted and

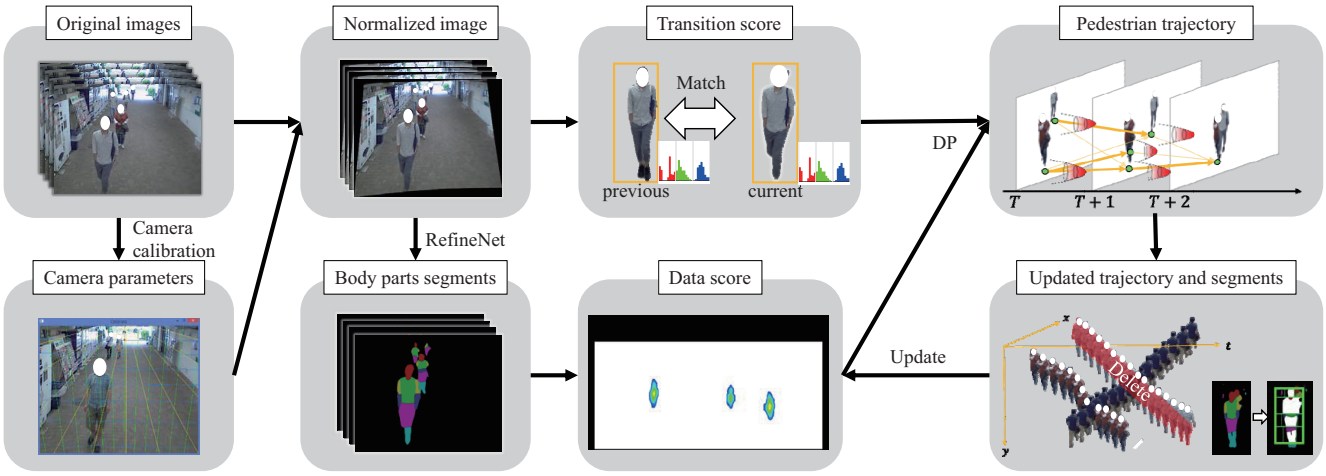


Figure 2. Overview of the proposed method. Facial regions are deleted or blurred due to privacy issues throughout this paper.

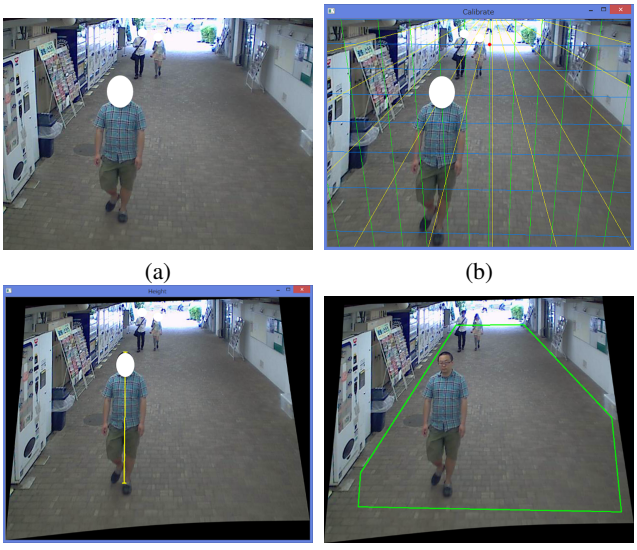


Figure 3. Easy and rough camera calibration tool. (a) An original image, (b) three orthogonal sets of parallel lines for camera parameter estimation up to scale (the red point represents the center of image, the blue, green and yellow lines represent horizontal, vertical and depth lines, respectively), (c) a normalized image and known height specification for scale estimation, (d) ROI specification on the normalized image plane by a polygon.

the data score for DP is updated to extract a trajectory of another pedestrian. As such, multiple pedestrian trajectories are extracted. Once we obtain the pedestrian trajectories, we extract gait features and recognize each pedestrian. We address individual modules in the following subsections.

2.2. An interactive camera calibration tool

Inspired by a camera calibration method in Manhattan world [5], we developed a camera calibration tool based on three orthogonal sets of parallel lines (i.e., horizontal, vertical, and depth lines) which does not require on-site

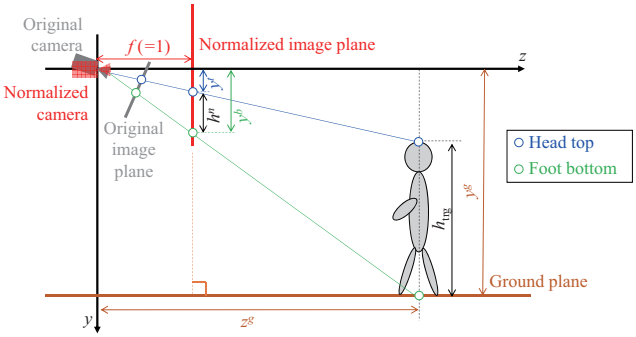
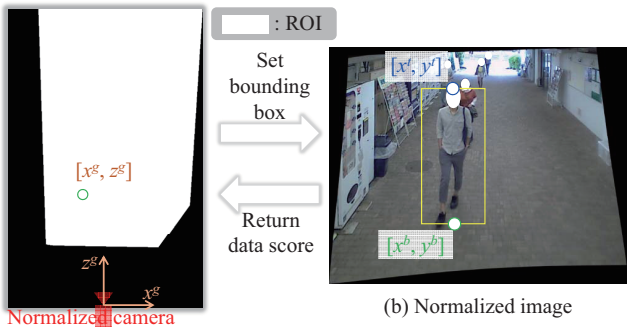


Figure 4. Schematic diagram for calculating the distance between the camera and the ground plane.

workload such as capturing calibration targets. A sufficient number of three orthogonal sets of parallel lines are, however, not always available and hence automatic estimation of camera calibration parameters may not work in some cases. We therefore rely more on human perception capability on the scene geometry than on automatic estimation.

More specifically, the three orthogonal sets of parallel lines are drawn based on current external and intrinsic camera parameters up to scale (i.e., a rotation matrix, an image center, lens distortion coefficients) and the user tries to fit the three orthogonal sets of the parallel lines to this scene geometry by modifying the camera parameters in an intuitively understandable manner (e.g., mouse drug operation) in this interactive system as shown in Fig. 3(b). Note that another external camera parameter, i.e., a translation vector, is set to a zero vector because the origin of the world coordinate is set to the same origin as the camera coordinate.

Once the calibration parameters are obtained, we can generate a virtual image plane whose focal length f is 1, the origin coincides with the image center, the lens distortion is corrected, and also which is perpendicular to the ground plane, where vertical lines in the world coordinate are projected as vertical lines in the image coordinate (call it a nor-



(a) Ground plane (b) Normalized image

Figure 5. Mapping between ground plane and image plane.

malized image plane later) (see Fig. 4). We then project the original image into the normalized image plane as shown in Fig. 3(c).

Furthermore, we take an image including a target person whose actual height h_{trg} in the world coordinate is known, in order to know the scale of scene geometry, specifically, the distance y^g from the camera to the ground plane. For this purpose, we firstly specify a foot bottom point $\boldsymbol{x}^b = [x^b, y^b]^T$ and a head top point $\boldsymbol{x}^t = [x^t, y^t]^T$ of the target person on the normalized image manually (i.e., by mouse click in our system) (see Fig. 3(c)), and then compute an apparent height of the target person on the normalized image as $h^n = y^b - y^t$. We can then easily compute the distance y^g by similarity transformation based on the fact that the normalized image plane is perpendicular to the ground plane (see Fig. 4) as

$$y^g = \frac{h_{\mathrm{trg}}}{h^n} y_b. \tag{1}$$

2.3. Bounding box setting from a foot bottom point on the ground plane

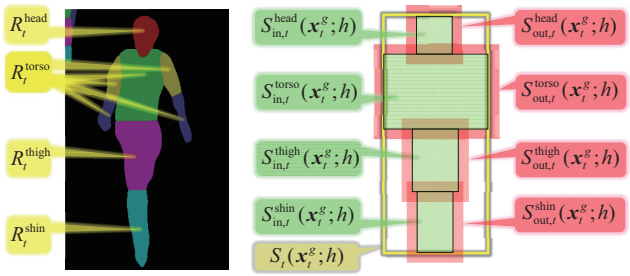
Once the camera calibration parameters are obtained, we can set a bounding box of a person in the normalized image given a foot bottom point $\boldsymbol{x}^g = [x^g, z^g]$ on the ground plane and also his/her assumed height h in the world coordinate (see Figs. 5).

Firstly, based on the fact that the foot bottom point lies on the ground plane (i.e., $y = y^g$) and the depth to the person is z^g, perspective projection gives mapping from a foot bottom point $\boldsymbol{x}^g = [x^g, z^g]^T$ on the ground plane into to a corresponding point $\boldsymbol{x}^b = [x^b, y^b]^T$ on the normalized image plane as,

$$\begin{bmatrix} x^b \\ y^b \end{bmatrix} = \frac{1}{z^g} \begin{bmatrix} x^g \\ y^g \end{bmatrix}. \tag{2}$$

The head top point in the world coordinate is denoted as $[x^g, y^g - h, z^g]^T$, and hence the mapping to the head top point $\boldsymbol{x}^t = [x^t, y^t]^T$ in the normalized image plane is similarly obtained as

$$\begin{bmatrix} x^t \\ y^t \end{bmatrix} = \frac{1}{z^g} \begin{bmatrix} x^g \\ y^g - h \end{bmatrix}. \tag{3}$$



(a) Body part segments (b) ROIs for data score computation

Figure 6. Body parts segments and ROIs for data score computation.

Finally, assuming a predefined aspect ratio of a pedestrian bounding box, we can set the bounding box based on the obtained foot bottom point $\boldsymbol{x}^b$ and head top point $\boldsymbol{x}^t$ as shown in Fig. 5.

Conversely, we can also project the foot bottom point $\boldsymbol{x}^b$ in the normalized image plane into the foot bottom point $\boldsymbol{x}^g$ on the ground plane by Eq. (2). Therefore, once we compute a data score (i.e., a pedestrian likelihood) for each bonding box, which will be described in the following subsection, we can set the data score on the ground plane accordingly (see Fig. 5).

2.4. DP-based tracking on the ground plane

We employ a DP-based batch tracking to extract a pedestrian trajectory. While conventional DP-based approaches were directly applied to an image, i.e., a state space for DP is defined over an image coordinate and the time axis, we define a state space over ROI on the ground plane and the time axis (in total three-dimensional spatio-temporal space). We then set a corresponding bounding box for each point on the ground plane with an assumed pedestrian's height as described in the previous subsection.

This is beneficial in two points. Firstly, a person usually walks on the ground plane at a constant velocity at least locally. Such a locally constant velocity prior is applicable to a pedestrian trajectory on the ground plane, but not necessarily true of that on the image plane due to perspective projection. Secondly, an actual height of a pedestrian does not change during walking except for up-down motion by gait. Such an actual height consistency for each pedestrian trajectory can be easily considered, since multiple state spaces are defined for individual assumed heights separately and the assumed height does not change within each state space over the ground plane. On contrary, it does not make sense to define a state space over the image plane with a fixed height because an apparent height for the same pedestrian

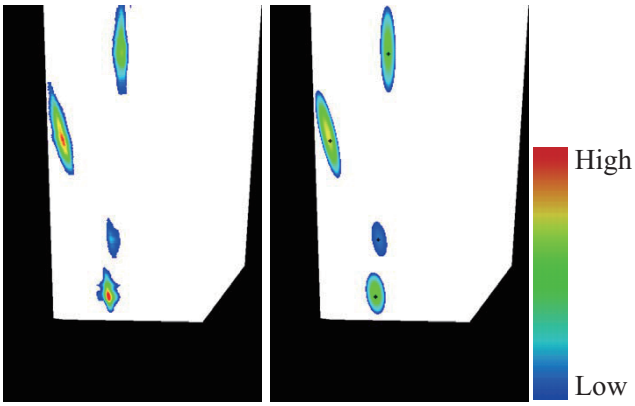


Figure 7. Data scores on the ground plane before and after quadratic approximation, respectively. Note that only higher scores than a threshold are shown with color, while the others are shown with white.

changes frame by frame in the image plane due to perspective projection.

In order to execute DP, we firstly define so-called data score and transition score, and then compute cumulative scores over the entire state space along with the optimal paths. A pedestrian trajectory is finally obtained as the optimal path that maximizes the cumulative score.

Data score

We utilize a pedestrian likelihood of a bounding box on the normalized image plane. The pedestrian likelihood is computed as the degree of overlapping between the predefined body part ROIs and body parts segments by RefineNet [27]. Specifically, we consider four pre-defined parts: head, torso (including arms), thigh, and shin (let them be $\mathcal{P} = \{\mathrm{head}, \mathrm{torso}, \mathrm{thigh}, \mathrm{shin}\}$), and body parts segments obtained RefineNet at the t-th frame are assigned to $\{R_t^p | p \in \mathcal{P}\}$ accordingly, as shown in Fig. 6 (a). After a bounding box $S_t(\boldsymbol{x}_t^g; h)$ on the normalized image plane is set based on a foot bottom point $\boldsymbol{x}_t^g$ on the ground plane at the t-th frame and on an assumed height h, we define ROIs for individual body parts relative to the bounding box as $\{S_{\mathrm{in},t}^p(\boldsymbol{x}_t^g; h) | p \in \mathcal{P}\}$ accordingly, as shown in Fig. 6 (b). Moreover, we introduce a peripheral region to penalize the overlap at the outside of each ROI as $\{S_{\mathrm{out},t}^p(\boldsymbol{x}_t^g; h) | p \in \mathcal{P}\}$ accordingly. We then compute a data score as

$$s_t^{\mathrm{dt}}(\boldsymbol{x}_t^g; h) = \frac{\sum_{p\in\mathcal{P}} |R_t^p \cap S_{\mathrm{in},t}^p(\boldsymbol{x}_t^g; h)|}{\sum_{p\in\mathcal{P}} |S_{\mathrm{in},t}^p(\boldsymbol{x}_t^g; h)|} - \frac{\sum_{p\in\mathcal{P}} |R_t^p \cap S_{\mathrm{out},t}^p(\boldsymbol{x}_t^g; h)|}{\sum_{p\in\mathcal{P}} |S_{\mathrm{out},t}^p(\boldsymbol{x}_t^g; h)|}, \quad (4)$$

where $|\cdot|$ means an area of a region. As such, we compute the data score for each foot bottom point over ROI on the ground plane as shown in Fig. 7 (a).

Transition score

The transition score encodes a transition likelihood from a state at a previous frame to another state at a current frame. For this purpose, we employ color histogram similarity between two bounding boxes. Firstly, we vertically divide the bounding box $S_t(\boldsymbol{x}_t^g; h)$ into N_V sub regions $\{S_{t,i}^{\mathrm{sub}}(\boldsymbol{x}_t^g; h)\}(i = 1, \ldots, N_V)$. We then compute HSV color histograms [46] $\boldsymbol{q}_t^c(\boldsymbol{x}_t^g; h) \in \mathbb{R}^{N_B}$, $c \in \mathcal{C} = \{H, S, V\}$ where N_B is the number of bins for each histogram. Note that voting source for histogram construction is limited to an intersected region of each sub region and body parts segments obtained by RefineNet $R_t = \cup_{p\in\mathcal{P}} R_p$ at the t-th frame, i.e., $S_{t,i}^{\mathrm{sub}}(\boldsymbol{x}_t^g; h) \cap R_t$. Finally, the transition score from a foot bottom point $\boldsymbol{x}_{t-1}^g$ at the $(t-1)$-th frame to another point $\boldsymbol{x}_t^g$ at the t-th frame is defined as

$$s_t^{\mathrm{tr}}(\boldsymbol{x}_{t-1}^g, \boldsymbol{x}_t^g; h) = \sum_{c\in\mathcal{C}} f(\boldsymbol{q}_{t-1}^c(\boldsymbol{x}_{t-1}^g; h), \boldsymbol{q}_t^c(\boldsymbol{x}_t^g; h)), \quad (5)$$

where a function $f(\cdot, \cdot)$ returns an intersection of two histograms [33].

Analytical DP

For the computation efficiency, we employ an analytical DP [55], where data and transition scores are approximated by quadratic forms w.r.t. the state and the optimal path search is done analytically rather than search over the entire state space. We briefly describe the analytical DP in this paragraph and hence may refer the reader to [55] for more details.

We firstly keep data scores above a pre-defined threshold (i.e., discard the others) and find K_t clusters by k-means clustering at the t-th frame. Thereafter, we approximate the data score for the k-th cluster (see Fig. 7 (b)) as

$$\hat{s}_{t,k}^{\mathrm{dt}}(\boldsymbol{x}_t^g; h) = \boldsymbol{x}_t^{gT} A_{t,k,h}^{\mathrm{dt}} \boldsymbol{x}_t^g + \boldsymbol{b}_{t,k,h}^{\mathrm{dt}\,T} \boldsymbol{x}_t^g + c_{t,k,h}^{\mathrm{dt}}, \quad (6)$$

where $A_{t,k,h}^{\mathrm{dt}}$, $\boldsymbol{b}_{t,k,h}^{\mathrm{dt}}$, and $c_{t,k,h}^{\mathrm{dt}}$ are a coefficient matrix, a coefficient vector, and a coefficient scalar for quadratic approximation of the data score for the k-th cluster at the t-th frame for the assumed height h, respectively.

The transition score from the l-th cluster at the $(t-1)$-th frame to the k-th cluster at the t-th frame is similarly approximated as

$$\hat{s}_{t,k,l}^{\mathrm{tr}}(\boldsymbol{x}_{t-1}^g, \boldsymbol{x}_t^g; h) = (\boldsymbol{x}_t^g - \boldsymbol{x}_{t-1}^g)^T A_{t,k,l,h}^{\mathrm{tr}} (\boldsymbol{x}_t^g - \boldsymbol{x}_{t-1}^g) + \boldsymbol{b}_{t,k,l,h}^{\mathrm{tr}\,T} (\boldsymbol{x}_t^g - \boldsymbol{x}_{t-1}^g) + c_{t,k,l,h}^{\mathrm{tr}}, \quad (7)$$

where $A_{t,k,l,h}^{\mathrm{tr}}$, $\boldsymbol{b}_{t,k,l,h}^{\mathrm{tr}}$, and $c_{t,k,l,h}^{\mathrm{tr}}$ are a coefficient matrix, a coefficient vector, and a coefficient scalar for quadratic approximation of the transition score, respectively.

Now, given a cumulative score $\hat{s}_{t-1,l}^{\mathrm{cum}}$ along with the optimal path up to the l-th cluster at the $(t-1)$-th frame, we can compute a cumulative score $\hat{s}_{t,k}^{\mathrm{cum}}$ along with the opti-

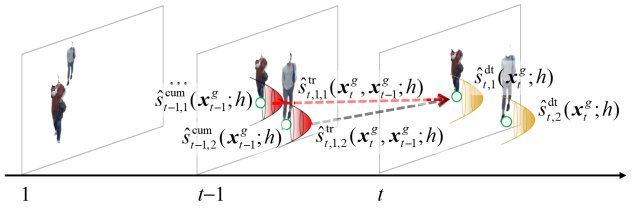
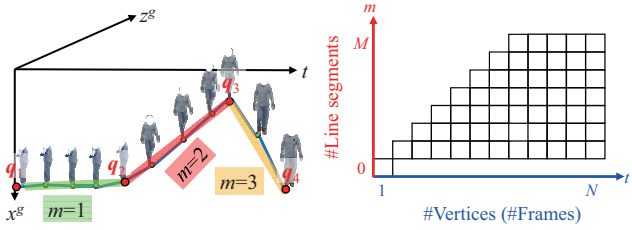


Figure 8. Analytical DP.



(a) Overview (b) State space

Figure 9. Piece-wise linear trajectory extraction.

mal path up to the k-th cluster at the t-th frame as

$$\hat{s}^{\mathrm{cum}}_{t,k}(\boldsymbol{x}^g_t;h) = \min_{l,\boldsymbol{x}^g_{t-1}} \left\{ \hat{s}^{\mathrm{cum}}_{t-1,l}(\boldsymbol{x}^g_{t-1};h) + \hat{s}^{\mathrm{tr}}_{t,k,l}(\boldsymbol{x}^g_{t-1},\boldsymbol{x}^g_t;h) \right\} + \hat{s}^{\mathrm{dt}}_{t,k}(\boldsymbol{x}^g_t;h). \quad (8)$$

Note that selection of the optimal state $\boldsymbol{x}^g_{t-1}$ given the l-th cluster is done analytically since the cumulative score of the quadratic data score and transition score is also quadratic [55], which greatly reduces the computational cost compared with conventional discrete DP with exhaustive search on the state space, i.e., possible positions $\boldsymbol{x}^g_{t-1}$ over the entire ground plane. On the other hand, selection of the optimal cluster l is done in a discrete manner as a conventional DP, but note that the number of clusters K_t is much fewer than the number of possible positions over the ground plane.

As such, we find the optimal state $\boldsymbol{x}^{g*}$ from the final frame or from the boundary region on ROI on the ground plane as well as the optimal assumed height $\boldsymbol{h}^*$ which maximizes the cumulative score, and extract the optimal path as a pedestrian trajectory by back tracking in DP framework.

In order to fine another pedestrian trajectory, we do the followings repeatedly. We delete part segments region used for the already chosen pedestrians, and update the data score as shown in Fig. 2. Subsequently, we choose the optimal trajectory for another pedestrian in the same way.

2.5. Piece-wise linear approximation of pedestrian trajectory

Because the aforementioned DP just considers the transition smoothness between adjacent frames, it does not guarantee more global smoothness on the trajectory. Actually, if we focus on a short moment, a pedestrian usually moves at almost constant velocity. We therefore approximate the trajectory in the three dimensional spatio-temporal space by piece-wise linear one.

Since this piece-wise linear approximation is analogous to a polygonal approximation for a series of contour points in shape analysis community, we adopt the same strategy for a pedestrian trajectory in the three-dimensional spatio-temporal space as shown in Fig. 9 (a). More specifically, we adopt a minimum number problem, i.e., given an N-vertex (-frame) pedestrian trajectory P, approximating it by another trajectory Q with minimum number of line segments so that the line fitting error does not exceed a given maximum tolerance Δ [18]. Nice thing with the minimum number problem is that the number of line segments is automatically determined depending on the complexity of the trajectory, in other words, we need not to define the number of line segments in advance. We briefly describe the algorithm in this subsection and may refer the reader to [18] for more details.

Let us define a pedestrian trajectory $P = \{\boldsymbol{p}_i | \boldsymbol{p}_i \in \mathbb{R}^3, i = 1,\ldots,N\}$ in the three-dimensional spatio-temporal space. Suppose that the trajectory P is approximated by another trajectory $Q = \{\boldsymbol{q}_i | \boldsymbol{q}_i \in P, i = 1,\ldots,M+1\}$ with $(M+1)$ vertices, i.e., by M line segments, a total line segment fitting error E is defined as

$$E = \sum_{m=1}^{M} e^2(\boldsymbol{q}_m, \boldsymbol{q}_{m+1}), \quad (9)$$

where $e^2(\boldsymbol{q}_m, \boldsymbol{q}_{m+1})$ is the sum of squared Euclidean distances from each vertex belonging to the line segment between $\boldsymbol{q}_m$ and $\boldsymbol{q}_{m+1}$.

Let us also defined a discrete two-dimensional state space $\{(n,m) | n = 1,\ldots,N, m = 1,\ldots,M\}$, where M is the possible maximum number of line segments, where each state (n,m) represents the sub-problem of approximation of an n-vertex pedestrian trajectory by m line segments (Fig. 9 (b)). We also introduce a function $D(n,m)$ as a cumulative cost function for the state (n,m). Note that the cumulative cost function value for the initial state $(1,0)$ is set to zero, i.e., $D(1,0) = 0$. The cumulative cost function is then computed with the following recursive expression

$$D(n,m) = \min_{m \le j \le n} \{ D(j, m-1) + d^2(\boldsymbol{p}_j, \boldsymbol{p}_n) \} \quad (10)$$

$$d^2(\boldsymbol{p}_j, \boldsymbol{p}_n) = \begin{cases} e^2(\boldsymbol{p}_j, \boldsymbol{p}_n) & e^2(\boldsymbol{p}_j, \boldsymbol{p}_n) \le \Delta \\ d_\infty & \text{otherwise} \end{cases}, \quad (11)$$

where d_∞ is a sufficiently large positive number (e.g., $d_\infty = N\Delta$). Note that the above equation supposes a situation where the vertices from $\boldsymbol{p}_1$ to $\boldsymbol{p}_j$ are approximated by $(m-1)$ line segments and also the vertices from $\boldsymbol{p}_j$ to $\boldsymbol{p}_n$ are approximated by the m-th line segment. When the error exceeds the pre-defined tolerance Δ, the sufficiently large positive number d_∞ is set so as to be excluded from possible state candidate.

Finally, we select the minimum m^* among a set of the number of line segments whose cumulative cost function value at the last vertex is less than the sufficiently large positive number, i.e., $\{m|D(N,m) < d_\infty\}$.

As such, we obtain m^* line segments, and approximate the foot bottom points $\hat{\boldsymbol{x}}_t^g$ on the ground plane by each line segment. Thereafter, we can again project it into the bounding box sequence in the normalized image plane as described before.

2.6. Gait recognition

Once a segmentation sequence (i.e., a silhouettes sequence) is given by RefineNet [27] and a bounding box sequence is given for each pedestrian, we simply apply a conventional approach to silhouette-based gait recognition. In short, we firstly register and size-normalize the silhouette sequence based on the bounding box sequence. We then detect a gait cycle from the normalized silhouette sequence, and average the silhouette over one gait cycle to get GEI [11] as a gait feature. Finally, given a pair of GEIs for matching, we compute Euclidean distance between them as a dissimilarity measure, which we execute person verification and identification by.

3. Experiments

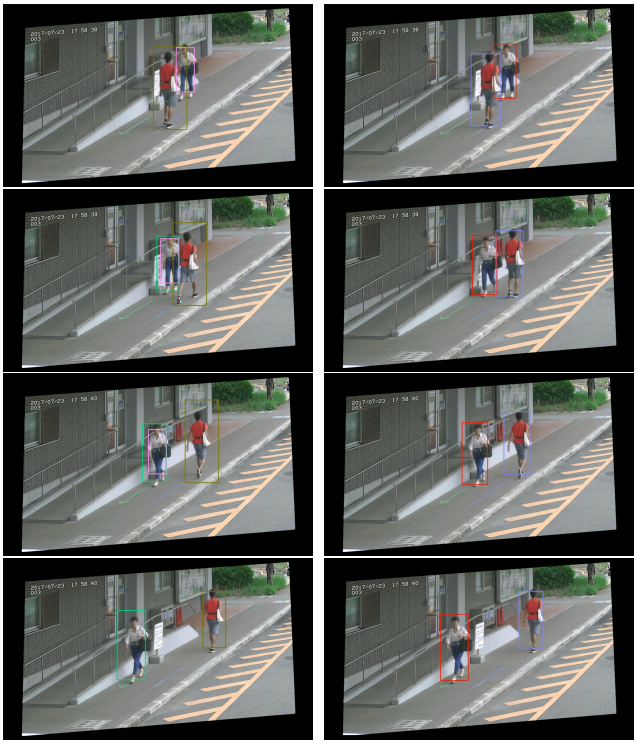
3.1. Data set

We conducted our experiments using a gait database collected by ourselves. Each subject was asked to walk along a pre-defined course around buildings. The original videos were captured by 960 $\times$ 540 pixels at 10 fps. Each subject agreed with the use of captured data for the research purpose. As such, two walking image sequences of 75 subjects were collected, where some subjects overlapped with other subjects on the images.

3.2. Qualitative evaluation

In this subsection, we qualitatively compare the extracted trajectories between the proposed method and a conventional tracking method by Milan et al. [40]. Figure 10 shows how the extracted bounding box sizes are consistent along with time axis. As for Milan et al. [40], we notice that the size of bounding box of a pedestrian with an orange T-shirt, gets rapidly larger just after overlapping with another pedestrian, and then suddenly gets smaller. On the other hand, the proposed method successfully keeps the consistency of bounding box size because it imposes that an actual height in the world coordinate is constant for each pedestrian.

We also checked the quality of extracted gait features, GEIs If we set bounding boxes based on body parts segments by RefineNet [27] frame-by-frame, the foot bottom points are inconsistent (Fig. 1 (a)(b)). As a result, the extracted GEI is blurry as shown in Fig. 1 (c). On the other hand, since the proposed method better keeps the consistency on foot bottom points, the extracted GEI is also well aligned as shown in Fig. 1 (d).



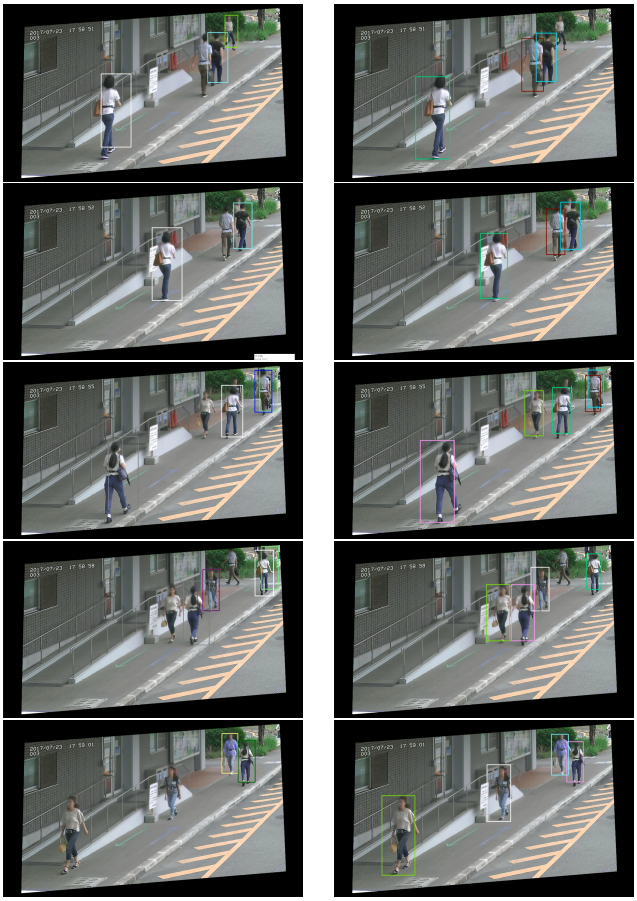
(a) Milan et al. [40] (b) Proposed

Figure 10. Bounding box size consistency. Each color indicates each extracted trajectory.

Table 1. EER [%] and rank-1, rank-5, rank-10 identification rates (denoted as Rank-1, Rank-5, Rank-10) [%]. Bold indicates the best accuracy.

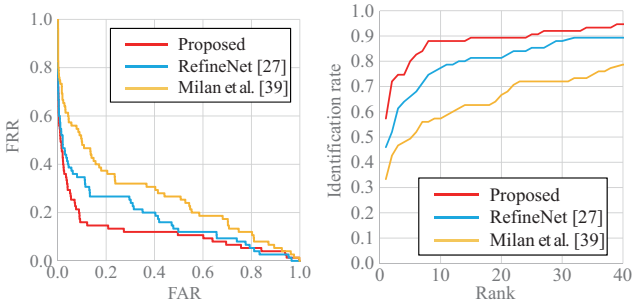
Method	EER	Rank-1	Rank-5	Rank-10
Milan et al. [40]	32.0	33.3	49.3	57.3
RefineNet [27]	26.7	46.0	66.0	77.0
Proposed	**14.7**	**72.0**	**80.0**	**88.0**

We then evaluate the frequency of ID switch and mis-detection in Fig. 11. As for Milan et al. [40], we observe several mis-detection (e.g., a pedestrian following another pedestrian with black T-shirt) and also ID switches (e.g., a pedestrian with grey bounding box is later regarded as another pedestrian with a green bounding box). On the other hand, the proposed method seldom suffers from such mis-detection and ID switches even after overlapping each other thanks to DP-based tracking framework. Moreover, the bounding box size is well consistent throughout the frames for each pedestrian.



(a) Milan et al. [40] (b) Proposed

Figure 11. ID switches and mis-detection.



(a) ROC curves (b) CMC curves

Figure 12. ROC and CMC curves.

3.3. Quantitative evaluation

In this subsection, we address the gait recognition accuracy quantitatively. We consider Milan et al. [40] and frame-by-frame bounding box setting for RefineNet result [27] as benchmarks. We evaluated the proposed method as well as the two baselines in both verification (one-to-one matching) and identification (one-to-many matching) modes.

For the verification mode, given a pair of inputs, we accept it as the same subject's pair if the dissimilarity measure between them is below an acceptance threshold, otherwise we reject it (i.e., regard it as a different subjects' pair). Here, we consider two types of error rates as performance measures: false acceptance rate (FAR) of the different subjects' pairs, and false rejection rate (FRR) of the same subjects' pairs. Because the FAR and FRR change as the acceptance threshold changes, we evaluated a trade-off between the FAR and FAR by a receiver operating characteristics (ROC) curve. In addition, we extract an equal error rate (EER) of the FAR and FRR as a typical performance measure.

For the identification mode, we match a probe to all the galleries and make a rank list based on dissimilarity measures (i.e., galleries with smaller dissimilarity measures get smaller ranks). We evaluated rates of true match galleries included up to rank-n by a cumulative matching characteristic (CMC) curve.

ROC curves CMC curves are shown in Fig. 12. In addition, EERs , and rank-1, rank-5, and rank-10 identification rates are summarized in Table 1. We can see the proposed method yields the lowest errors and the highest identification rates among the benchmarks. More specifically, compared with Milan et al. [40] and RefineNet [27], the proposed method reduces EERs by 17.3% and 12.0%, respectively, and improves rank-1 identification rates by 38.7% and 26.0%, respectively. Consequently, we conclude that the proposed method steadily improves both verification and identification performances.

4. Conclusion

This paper described a method of pedestrian trajectory extraction considering scene geometry to get well aligned gait features. More specifically, we developed an interactive tool for camera calibration without on-site workload. We then applied discrete and analytical DP on the ground plane by using recent semantic segmentation and color histogram similarity. Moreover, we introduced piece-wise linear approximation to the pedestrian trajectory to gain more global smoothness. Experiments using scenes where multiple pedestrians sometimes overlap each other and change their apparent heights due to perspective projection, showed the effectiveness of the proposed method.

Since we focus on pre-processing part for gait recognition so far, we employed a quite simple approach for matching part in this paper. In order to show the effectiveness of the proposed method in the state-of-the-art framework, we will evaluate the proposed method in conjunction with recent CNN-based approaches [48, 49, 62] in the future.

Acknowledgement

This work was supported by JSPS Grants-in-Aid for Scientific Research (A) JP18H04115.

References

[1] M. R. Aqmar, K. Shinoda, and S. Furui. Robust gait recognition against speed variation. In *Proc. of the 20th International Conference on Pattern Recognition*, pages 2190–2193, Istanbul, Turkey, Aug. 2010.

[2] G. Ariyanto and M. Nixon. Marionette mass-spring model for 3d gait biometrics. In *Proc. of the 5th IAPR International Conference on Biometrics*, pages 354–359, March 2012.

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

[4] K. Bashir, T. Xiang, and S. Gong. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31(13):2052–2060, Oct. 2010.

[5] J.-C. Bazin, Y. Seo, C. Demonceaux, P. Vasseur, K. Ikeuchi, I. Kweon, and M. Pollefeys. Globally optimal line clustering and vanishing point estimation in manhattan world. In *Proc. of the 25th IEEE Conf. on Computer Vision and Pattern Recognition*, pages 638–645, Jun. 2012.

[6] A. Bobick and A. Johnson. Gait recognition using static activity-specific parameters. In *Proc. of the 14th IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 423–430, 2001.

[7] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon. On using gait in forensic biometrics. *Journal of Forensic Sciences*, 56(4):882–889, 2011.

[8] D. Cunado, M. Nixon, and J. Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90(1):1–41, 2003.

[9] B. Decann and A. Ross. Gait curves for human recognition, backpack detection, and silhouette correction in a nighttime environment. In *Proc. of the SPIE, Biometric Technology for Human Identification VII*, volume 7667, pages 1–13, 2010.

[10] Y. Guan and C.-T. Li. A robust speed-invariant gait recognition system for walker and runner identification. In *Proc. of the 6th IAPR International Conference on Biometrics*, pages 1–8, 2013.

[11] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(2):316– 322, 2006.

[12] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Oct 2017.

[13] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll. The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits. *J. Vis. Comun. Image Represent.*, 25(1):195–206, Jan. 2014.

[14] M. Hofmann, S. M. Schmidt, A. Rajagopalan, and G. Rigoll. Combined face and gait recognition using alpha matte preprocessing. In *Proc. of the 5th IAPR Int. Conf. on Biometrics*, pages 1–8, New Delhi, India, Mar. 2012.

[15] M. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi. Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. *Pattern Recognition*, 43(6):2281–2291, Jun. 2010.

[16] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi. Gait verification system for criminal investigation. *IPSJ Trans. on Computer Vision and Applications*, 5:163–175, Oct. 2013.

[17] A. Kale, A. Roy-Chowdhury, and R. Chellappa. Fusion of gait and face for human identification. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2004 (ICASSP'04)*, volume 5, pages 901–904, 2004.

[18] A. Kolesnikov and P. Fränti. Polygonal approximation of closed contours. In J. Bigun and T. Gustavsson, editors, *Image Analysis*, pages 778–785, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[19] W. Kusakunniran. Attribute-based learning for gait recognition using spatio-temporal interest points. *Image Vision Comput.*, 32(12):1117–1126, Dec. 2014.

[20] 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. of IEEE computer society conference on Computer Vision and Pattern Recognition 2010*, pages 1–8, San Francisco, CA, USA, Jun. 2010.

[21] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Speed-invariant gait recognition based on procrustes shape analysis using higher-order shape configuration. In *The 18th IEEE Int. Conf. Image Processing*, pages 545–548, 2011.

[22] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Cross-view and multi-view gait recognitions based on view transformation model using multi-layer perceptron. *Pattern Recognition Letters*, 33(7):882–889, 2012.

[23] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Gait recognition across various walking speeds using higher order shape configuration based on a differential composition model. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(6):1654–1668, Dec. 2012.

[24] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Gait recognition under various viewing angles based on correlated motion regression. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(6):966–980, 2012.

[25] L. Lee. *Gait Analysis for Classification*. PhD thesis, Massachusetts Institute of Technology, 2002.

[26] S. Lee, Y. Liu, and R. Collins. Shape variation-based frieze pattern for robust gait recognition. In *Proc. of the 2007 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, USA, Jun. 2007.

[27] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5168–5177, July 2017.

[28] Z. Liu and S. Sarkar. Simplest representation yet for gait recognition: Averaged silhouette. In *Proc. of the 17th International Conference on Pattern Recognition*, volume 1, pages 211–214, Aug. 2004.

[29] Z. Liu and S. Sarkar. Effect of silhouette quality on hard problems in gait recognition. *IEEE Trans. of Systems, Man, and Cybernetics Part B: Cybernetics*, 35(2):170–183, 2005.

[30] Z. Liu and S. Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):863–876, 2006.

[31] J. Lu and Y.-P. Tan. Uncorrelated discriminant simplex analysis for view-invariant gait signal computing. *Pattern Recognition Letters*, 31(5):382–393, 2010.

[32] N. Lynnerup and P. Larsen. Gait as evidence. *IET Biometrics*, 3(2):47–54, 6 2014.

[33] L. S. M., X. J. H., and W. S. Evaluation of image similarity by histogram intersection. *Color Research & Application*, 30(4):265–274, 2005.

[34] Y. Makihara, H. Mannami, A. Tsuji, M. Hossain, K. Sugiura, A. Mori, and Y. Yagi. The ou-isir gait database comprising the treadmill dataset. *IPSJ Transactions on Computer Vision and Applications*, 4:53–62, Apr. 2012.

[35] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model in the frequency domain. In *Proc. of the 9th European Conference on Computer Vision*, pages 151–163, Graz, Austria, May 2006.

[36] 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. on Computer Vision and Applications*, 7:(to appear), Jul. 2015.

[37] Y. Makihara, A. Tsuji, and Y. Yagi. Silhouette transformation based on walking speed for gait identification. In *Proc. of the 23rd IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, Jun 2010.

[38] A. Mansur, Y. Makihara, R. Aqmar, and Y. Yagi. Gait recognition under speed transition. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2521–2528, June 2014.

[39] D. Matovski, M. Nixon, S. Mahmoodi, and T. Mansfield. On including quality in applied automatic gait recognition. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3272–3275, Nov 2012.

[40] A. Milan, L. Leal-Taixe, K. Schindler, and I. Reid. Joint tracking and segmentation of multiple targets. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5397–5406, June 2015.

[41] A. Mori, Y. Makihara, and Y. Yagi. Gait recognition using period-based phase synchronization for low frame-rate videos. In *Proc. of the 20th International Conference on Pattern Recognition*, pages 2194–2197, Istanbul, Turkey, Aug. 2010.

[42] H. Murase and R. Sakai. Moving object recognition in eigenspace representation: Gait analysis and lip reading. *Pattern Recognition Letters*, 17:155–162, 1996.

[43] M. S. Nixon, T. N. Tan, and R. Chellappa. *Human Identification Based on Gait*. Int. Series on Biometrics. Springer-Verlag, Dec. 2005.

[44] M. Rokanujjaman, M. Islam, M. Hossain, M. Islam, Y. Makihara, and Y. Yagi. Effective part-based gait identification using frequency-domain gait entrophy features. *Multimedia Tools and Applications*, 74(9):3099–3120, May 2015.

[45] S. Sarkar, J. Phillips, Z. Liu, I. Vega, P. G. ther, and K. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Trans. of Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[46] K. B. Shaik, P. Ganesan, V. Kalist, B. Sathish, and J. M. M. Jenitha. Comparative study of skin color detection and segmentation in hsv and ycbcr color space. *Procedia Computer Science*, 57:41 – 48, 2015. 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015).

[47] G. Shakhnarovich and T. Darrell. On probabilistic combination of face and gait cues for identification. In *Proc. Automatic Face and Gesture Recognition 2002*, volume 5, pages 169–174, 2002.

[48] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *Proc. of the 8th IAPR Int. Conf. on Biometrics (ICB 2016)*, number O19, pages 1–8, Halmstad, Sweden, Jun. 2016.

[49] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. On input/output architectures for convolutional neural network-based cross-view gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2017.

[50] R. Tanawongsuwan and A. Bobick. Gait recognition from time-normalized joint-angle trajectories in the walking plane. In *Proc. of the 14th IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 726–731, Jun. 2001.

[51] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple people tracking by lifted multicut and person re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3701–3710, July 2017.

[52] D. Tao, X. Li, X. Wu, and S. Maybank. Human carrying status in visual surveillance. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 1670–1677, New York, USA, Jun. 2006.

[53] D. Tao, X. Li, X. Wu, and S. J. Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1700–1715, Oct 2007.

[54] D. Tao, X. Li, X. Wu, and S. J. Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1700–1715, Oct. 2007.

[55] S. Uchida, I. Fujimura, H. Kawano, and Y. Feng. Analytical dynamic programming tracker. In *Proc. of the 10th Asian Conf. on Computer Vision*, pages 296–309. Springer Berlin Heidelberg, 2010.

[56] R. Urtasun and P. Fua. 3d tracking for gait characterization and recognition. In *Proc. of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 17–22, 2004.

[57] D. Wagg and M. Nixon. On automated model-based extraction and analysis of gait. In *Proc. of the 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 11–16, 2004.

[58] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan. Human identification using temporal information preserving gait template. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(11):2164 –2176, nov. 2012.

[59] J. Wang, Y. Makihara, and Y. Yagi. People tracking and segmentation using spatiotemporal shape constraints. In *1st ACM Int. Workshop on Vision Networks for Behaviour Analysis*, Vancouver, Canada, Oct. 31 2008.

[60] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, dec. 2003.

[61] M. Wang, Y. Liu, and Z. Huang. Large margin object tracking with circulant feature maps. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4800–4808, July 2017.

[62] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2016.

[63] C. Yam, M. Nixon, and J. Carter. Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*, 37(5):1057–1072, 2004.

[64] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proc. of the 18th Int. Conf. on Pattern Recognition*, volume 4, pages 441–444, Hong Kong, China, Aug. 2006.

[65] G. Zhao, G. Liu, H. Li, and M. Pietikainen. 3d gait recognition using multiple cameras. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 529–534, April 2006.