Towards Robust Gait Recognition

(Invited Paper)

Yasushi Makihara
The Institute of Scientific and Industrial Research, Osaka University
Email: makihara@am.sanken.osaka-u.ac.jp

***Abstract*—Gait recognition is a method of biometric person authentication from his/her unconscious walking manner. Unlike the other biometrics such as DNA, fingerprint, vein, and iris, the gait can be recognized even at a distance from a camera without subjects' cooperation, and hence it is expected to be applied to many fields: criminal investigation, forensic science, and surveillance. However, the absence of the subjects' cooperation may sometimes induces large intra-subject variations of the gait due to the changes of viewpoints, walking directions, speeds, clothes, and shoes. We therefore develop methods of robust gait recognition with (1) an appearance-based view transformation model, (2) a kinematics-based speed transformation model. Moreover, CCTV footages are often stored as low frame-rate videos due to limitation of communication bandwidth and storage size, which makes it much more difficult to observe a continuous gait motion and hence significantly degrades the gait recognition performance. We therefore solve this problem with (3) a technique of periodic temporal super resolution from a low frame-rate video. We show the efficiency of the proposed methods with our constructed gait databases.**

I. Introduction

There is a growing necessity in modern society for identification of individuals in many situations, such as from surveillance systems and for access control. For this end, many biometric authentication methods are proposed using a wide variety of physiological cues (e.g., fingerprint, finger or hand vein, iris, and face) or behavioral cues (e.g., offline and online signature, eye movement, and gait). Among these, gait recognition has recently gained considerable attention because gait is a promising cue for surveillance systems to ascertain identity at a distance from a camera without subjects' cooperation.

The absence of the subjects' cooperation in gait recognition may, however, sometimes induce large intra-subject variations of gait features due to the changes of viewpoints, walking directions, speeds, clothes, and shoes. Since it is generally difficult to enroll all the possible variations of the gait features for individual recognition targets (uncooperative subjects for testing such as perpetrators and suspects) in advance, gait recognition under cross conditions is often required.

On the other hands, it is likely possible to collect a variety of the gait features for non-recognition targets (cooperative subjects for training such as students in a laboratory and colleagues in a company), and hence we exploit them to draw useful knowledge about how the gait features change across the conditions. More specifically, we opt for generative approaches to transform a gait feature under a condition into that under another condition by transformation models. From this context, we introduce methods of cross-view gait recognition with an appearance-based view transformation model (VTM) [1] and also cross-speed gait recognition with a kinematics-based speed transformation model (STM) [2].

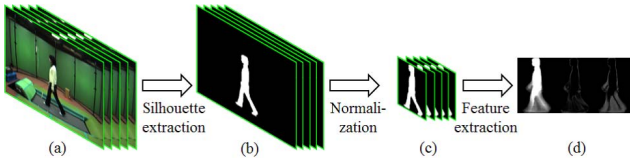


Fig. 1. Feature extraction procedure. (a): Original image sequence, (b): Silhouette sequence, (c): Size-normalized silhouette sequence, (d): Frequency-domain feature. In (d), the left, center, right images stand for 0, 1, and 2-times frequency amplitude spectra.

Moreover, CCTV footages are often stored as low frame-rate videos due to limitation of communication bandwidth and storage size. Since only sparsely sampled phases (gait stances) are observed in such low frame-rate videos, sets of the sparse phases for a matching pair may be different each other. This means the gait recognition using low frame-rate videos also leads to a kind of cross-condition gait recognition, namely, cross-phase gait recognition. In order to solve this problem, we again opt for a generative approach: periodic temporal super resolution (PTSR) from a low frame-rate video based on both multiple-periods observations and exemplars of high frame-rate videos from the training subjects [3].

II. Feature extraction

A. Preprocessing

Since we employ appearance-based, more specifically, silhouette-based representation for a gait feature, the first step to extracting the gait feature is silhouette extraction. Given an original image sequence (Fig. 1(a)) and a background image sequence, a silhouette sequence (Fig. 1(b)) is extracted by background subtraction-based graph-cut segmentation [4]. A bounding box sequence for the silhouette sequence is computed and a size-normalized silhouette sequence (Fig. 1(c)) is then generated by scaling the height of the bounding box and by registering the silhouette center. In the following sections, we represent the silhouette value (1: foreground, 0: background) in the size-normalized image sequence at position (x, y) at the n-th frame as $f(x, y, n)$.

B. Frequency-domain feature

The second step is detection of gait period, namely, time duration for a pair of left and right steps. Focused on a specific position (x, y) of the size-normalized silhouette sequence $f(x, y, n)$, it is viewed as a one-dimensional signal along the temporal axis. Since a set of signals from all the positions

978-1-4799-2190-4/13 $26.00 © 2013 IEEE
DOI 10.1109/ACPR.2013.211

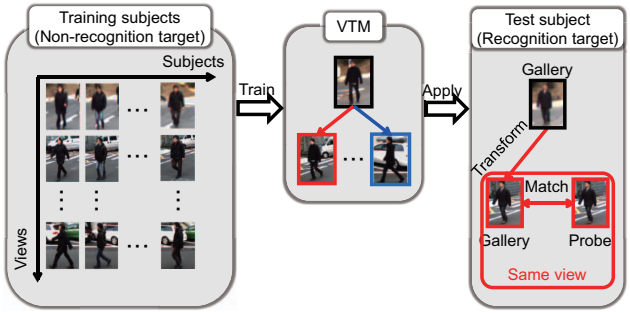


Fig. 2. Framework of VTM.

construct a multi-dimensional signal along the temporal axis, we compute the gait period P by maximizing the normalized autocorrelation $C(\tau)$ of the multi-dimensional signal with τ frame shift as

$$P = \arg \max_{\tau \in [\tau_{min}, \tau_{max}]} C(\tau) \tag{1}$$

$$C(\tau) = \frac{\sum_{x,y} \sum_{n=1}^{N-\tau} f(x,y,n) f(x,y,n+\tau)}{\sqrt{\sum_{x,y} \sum_{n=1}^{N-\tau} f(x,y,n)^2} \sqrt{\sum_{x,y} \sum_{n=1}^{N-\tau} f(x,y,n+\tau)^2}}, \tag{2}$$

where N is the number of total frames in the sequence, τ_{min} and τ_{max} are the minimum and the maximum periods, respectively.

Next, we compute discrete Fourier transformation (DFT) along the temporal axis for each position (x, y) independently and subsequently compute amplitude spectra as

$$A(x,y,k) = \left| \sum_{n=1}^{P} f(x,y,n) e^{-j\omega kn} \right|, \tag{3}$$

where $\omega (= 2\pi/P)$ is a base angular frequency for the gait period P, and $A(x, y, k)$ is an amplitude spectrum at position (x, y) for k-times frequency. We stack all the elements of the amplitude spectra for low frequencies ($k = 0, 1, 2$) into an unfolded vector $\boldsymbol{a} \in \mathbb{R}^M$ and refer to the vector as the frequency-domain feature (FDF) [5]. Since the FDF is free from phase of the start frame and also normalized by the period, it avoids troublesome frame synchronization and time normalization during matching and feature transformation process.

An example of the FDF is visualized as shown in Fig. 1(d). Focused on a side-view case, while 0-times frequencies mainly reflects body shape, 1- and 2-times frequency reflect left-right asymmetric and symmetric motions, respectively.

III. View-Invariant Gait Recognition

A. Framework

We first introduce a framework of view-invariant gait recognition using a VTM as shown in Fig. 2. In the training phase, given gait features of multiple non-recognition targets from multiple views, the VTM from one view to another view is trained. In the test phase, given a pair of gallery and probe gait features of recognition targets from different views, the gallery gait feature is transformed so as to be the same view as those of the probe gait feature and then a pair of the probe gait feature and the transformed gallery gait feature is matched under the same view.

B. Formulation

We formulate a VTM in the frequency domain in a way similar to that in [6]. Note that we apply the model to the FDF extracted from the image sequence while that in [6] directly applied it to a single image.

We first quantize views into L views. Let $\boldsymbol{a}_{\theta_l}^i$ be an M-dimensional feature vector for the l-th view of the i-th training subject. Supposing that the feature vectors for L views of I subjects are obtained as a training set, we can construct a matrix whose row indicates view changes and whose column indicates each subject; and so can decompose it by singular value decomposition as

$$\begin{bmatrix} \boldsymbol{a}_{\theta_1}^1 & \cdots & \boldsymbol{a}_{\theta_1}^I \\ \vdots & \ddots & \vdots \\ \boldsymbol{a}_{\theta_L}^1 & \cdots & \boldsymbol{a}_{\theta_L}^I \end{bmatrix} = \begin{bmatrix} Q_{\theta_1} \\ \vdots \\ Q_{\theta_L} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}^1 & \cdots & \boldsymbol{v}^I \end{bmatrix}, \tag{4}$$

where Q_{θ_l} is the $M \times I$ submatrix, and $\boldsymbol{v}^i$ is the I-dimensional column vector. Note that a whole matrix $[Q_{\theta_1}^T, \ldots, Q_{\theta_L}^T]^T$ is regarded as a set of basis of the eigen space for the all-view feature vector $[\boldsymbol{a}_{\theta_1}^T, \ldots, \boldsymbol{a}_{\theta_I}^T]^T$, while that the vector $\boldsymbol{v}$ is regarded as a point in the eigen space.

Once the eigen space of all-view feature vector is obtained, the view transformation process is regarded as a kind of missing data reconstruction. Given a feature vector $\boldsymbol{a}_{\theta_{ref}}$ of a test subject from reference view θ_{ref}, a corresponding point $\hat{\boldsymbol{v}}$ in the eigen space is computed by the least square as

$$\hat{\boldsymbol{v}} = \arg\min_{\boldsymbol{v}} \left\| Q_{\theta_{ref}} \boldsymbol{v} - \boldsymbol{a}_{\theta_{ref}} \right\|^2 = {Q_{\theta_{ref}}}^{+} \boldsymbol{a}_{\theta_{ref}}, \tag{5}$$

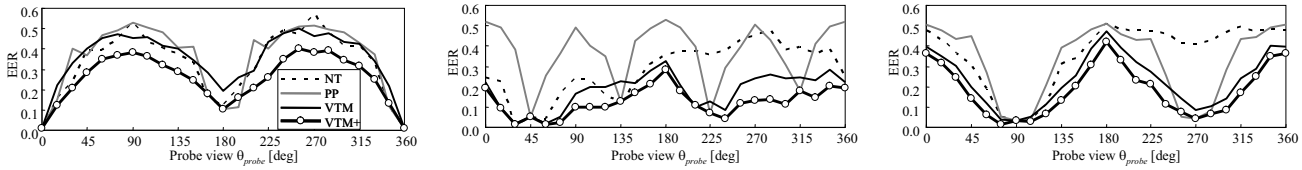
where ${Q_{\theta_{ref}}}^{+}$ is a pseudo inverse matrix of $Q_{\theta_{ref}}$. Now, we can back-project the vector $\hat{\boldsymbol{v}}$ into any other view θ_i as

$$\hat{\boldsymbol{a}}_{\theta_i} = Q_{\theta_i} \hat{\boldsymbol{v}}. \tag{6}$$

Moreover, view transformation accuracy improves when we add reference feature vectors from virtual views based on geometric assumptions. We refer the readers to [7] for more details.

C. Results

We use walking image sequences of 20 subjects (10 for training and 10 for probes) from 24 views (15 deg. azimuth interval) for the experiments. We first show transformed features with our VTM in Fig. 3. We can see that the transformed features looks similar to the probe ones for each viewpoint. We subsequently show equal error rate (EER) of the cross-view gait recognition in a verification scenario (one-to-one matching), which is a tradeoff error between false acceptance rate (FAR) of imposters and false rejection rate (FRR) of the genuine. The benchmarks are direct matching with no transformation (NT), perspective projection of sagittal plane (PP) [8], VTM, and VTM with reference viewpoint addition (VTM+, proposed). It turns out that the proposed method achieves the lowest EER for almost all the cross-view combinations.



(a) Gallery: 0 deg (b) Gallery: 45 deg (c) Gallery: 90 deg

Fig. 4. EERs of cross-view gait recognition.

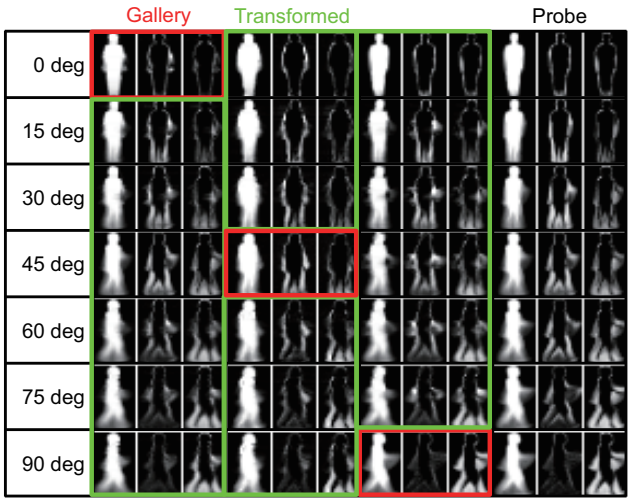


Fig. 3. View transformed features. In the first three columns, while gait features with red boxes are original gait features, those with green boxes are transformed gait features from the original gait features. We can see the transformed gait features are similar to the probe gait features (right) to some extent.

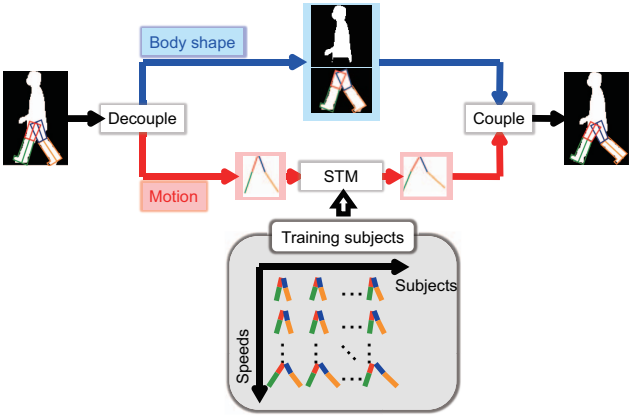


Fig. 5. Framework of STM.

IV. Speed-invariant gait recognition

A. Framework

We first introduce a framework of view-invariant gait recognition using a STM as shown in Fig. 5. We slightly modify the VTM framework so as to reflect a key observation that speed change affect not body shapes but motions, which leads to kinematics-based STM.

A silhouette sequence is first decoupled into the body shape and the motion by fitting a human model. The STM is trained by a set of motions of the multiple non-recognition targets under multiple speeds, and it is then applied to the motion of a recognition target to transform from one speed to another speed. The body shape and the speed-transformed motion is then coupled to generate speed-transformed silhouette sequence.

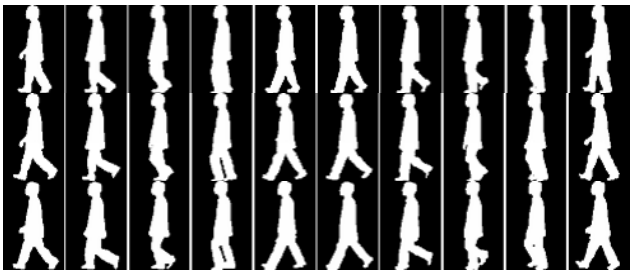


Fig. 6. Silhouette transformation result (middle) from 3 km/h (top) to 7 km/h (bottom).

Because the same formulation is available just by replacing the FDF with the motions, more specifically, a vector of a time-normalized and phase-synchronized joint angle sequence, we omit the detailed formulation of the STM.

B. Results

We conducted experiments of cross-speed gait recognition with the OU-ISIR Gait Database, the treadmill dataset A [9][1]. The number of training subjects is 14, while that of testing subjects is 20. The speed variation used in this experiment is ranging from 2 km/h to 7 km/h at 1 km/h interval.

An example of generated silhouette sequences from 3 km/h to 7 km/h is shown in Fig. 6. We can see that the generated silhouette sequence at 7 km/h is similar to the original probe silhouette sequences at 7 km/h.

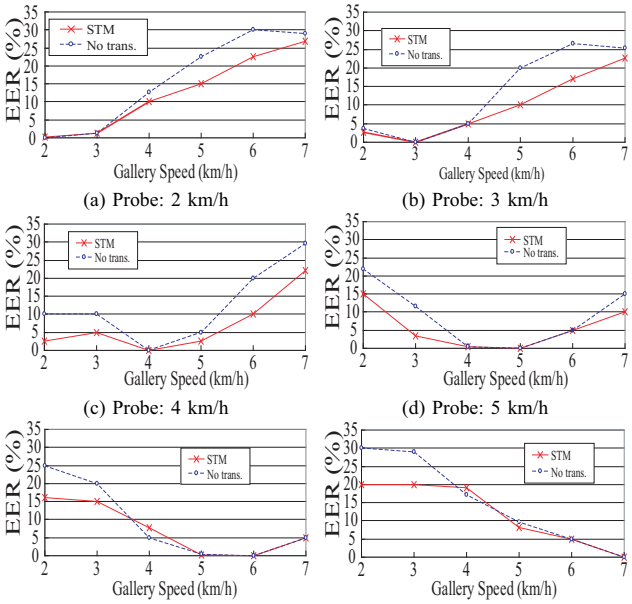
The performance in the verification scenario was evaluated by EERs for all the combinations of the gallery and probe speeds as shown in Fig. 7. As a result, the averaged EERs on different speed combinations were reduced from 15.0% with no transformation to 10.9% (4.1% improvement) with transformation. As a whole, the improvement becomes larger as the speed difference between the gallery and the probe becomes larger, and in particular the largest EER improvements for a specific pair of gallery and probe speeds (2 km/h gallery and 7 km/h probe) is 10.0%.

V. Gait recognition at low frame-rate

A. Framework

We first introduce a framework of PTSR for gait recognition at low frame-rate as shown in Fig. 8. Given an input low

[1]Publicly available at http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitTM.html



(a) Probe: 2 km/h (b) Probe: 3 km/h (c) Probe: 4 km/h (d) Probe: 5 km/h (e) Probe: 6 km/h (f) Probe: 7 km/h

Fig. 7. EER for cross-speed gait recognition.

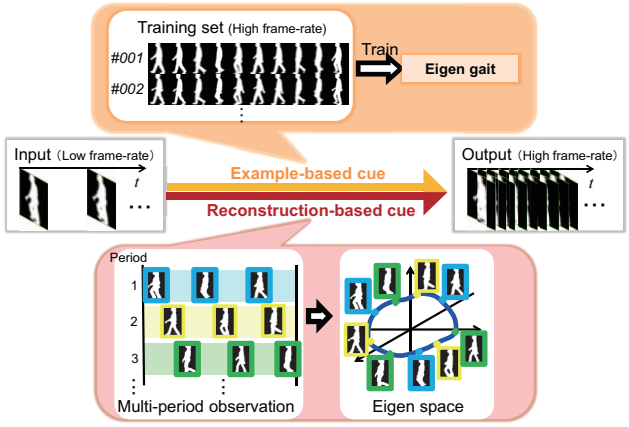


Fig. 8. Framework of PTSR.

frame-rate video, an output high frame-rate video is generated both with reconstruction-based and example-based cues. Since the various phases are obtained through multiple-periods observations, they form a closed curve (manifold) in the silhouette eigen space, which is treated as the reconstruction-based cue. On the other hand, multiple high frame-rate videos from training set (non-recognition targets) construct multiple manifolds and eigen gait (eigen space of the manifolds) constrains, more specifically, regularize the possible solution of the manifold.

B. Formulation

Let an input low frame-rate image sequence is expressed in the M-dimensional eigen space as $Y_Q^{in} = [\boldsymbol{y}_{Q,1}^{in}, \ldots, \boldsymbol{y}_{Q,N}^{in}] \in \mathbb{R}^{M \times N}$ where the accompanying phase sequence $\boldsymbol{s}_Q = [s_{Q,1}, \ldots, s_{Q,N}]^T \in \mathbb{R}^N$ is unknown. Since we represent an output high frame-rate image sequence by a manifold with cubic natural spline in the eigen space, we try estimating a control point matrix of the spline $Y^{cp} = [\boldsymbol{y}_1^{cp}, \ldots, \boldsymbol{y}_{N^{cp}}^{cp}] \in \mathbb{R}^{M \times N^{cp}}$, where N^{cp} is the number of the control points.

Note that given the control point matrix, Y^{cp}, and that approximation in the silhouette eigen space at the i-th phase $s_{Q,i}$ is

$$\hat{\boldsymbol{y}}(Y^{cp}, s_{Q,i}) = Y^{cpT} D^T \boldsymbol{w}(s_{Q,i}), \tag{7}$$

where $\boldsymbol{w}(s_{Q,i})$ is an interpolation coefficient vector and D is a conversion matrix from the control point matrix Y^{cp} to a spline parameter matrix.

On the other hand, an example-based estimator $\hat{Y}_{tr}$ of the control point matrix is expressed as

$$\hat{Y}^{tr} = \bar{Y}^{tr} + \sum_{j=1}^{M_m} \alpha_j E_j^{tr}, \tag{8}$$

where $\bar{Y}^{tr}$ and $\{E_j^{tr}\}(j = 1, \ldots, M_m)$ are the mean and the eigen control point matrices (call them eigen gaits later), which are obtained from high frame-rate image sequences of the training subjects, and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{M_m}]^T$ is a coefficient vector for the eigen gait $\{E_j^{tr}\}$.

The energy function is then constructed by considering the four aspects: (1) data fitness between the interpolation $\hat{\boldsymbol{y}}(Y^{cp}, s_{Q,i})$ and the input $\boldsymbol{y}_{Q,i}^{in}$, (2) fitness between the control point matrix Y^{cp} and the exemplar-based estimator $\hat{Y}^{tr}$, (3) smoothness of the periodic manifold $\boldsymbol{y}_s(s; Y^{cp})$ in the silhouette eigen space, and (4) smoothness of the phase evolution $\boldsymbol{s}_Q$ based on the linear phase evolution prior. The actual form of the function is

$$\begin{aligned} E(Y^{cp}, \boldsymbol{\alpha}, \boldsymbol{s}_Q) = & \\ & \frac{1}{N^{in}} \sum_{i=1}^{N^{in}} \left\| Y^{cpT} D^T \boldsymbol{w}(s_{Q,i}) - \boldsymbol{y}_{Q,i}^{in} \right\|^2 + \\ & \lambda_t \frac{1}{N^{cp}} \left\| Y^{cp} - \left(\bar{Y}^{tr} + \sum_{j=1}^{M_m - 1} \alpha_j E_j^{tr} \right) \right\|^2 + \\ & \lambda_m \frac{1}{N^{cp}} \left\| B Y^{cp} \right\|^2 + \\ & \lambda_s \frac{1}{N^{in}} \sum_{i=1}^{N^{in}-1} \left(s_{Q,i+1} - s_{Q,i} - \frac{1}{P'} \right)^2, \end{aligned} \tag{9}$$

where the first, second, third and fourth terms are the data term, fitness term to the exemplar-based estimator, smoothness term for the manifold and the smoothness term for phase evolution respectively. B is a coefficient matrix for calculating the manifold curvature, and $P'(= fP)$ is the global period (in terms of frames) assuming linear phase evolution. Finally, the control point matrix Y^{cp}, the coefficients for the eigen gaits $\boldsymbol{\alpha}$, and phase sequence $\boldsymbol{s}_Q$ are estimated so as to minimize the energy function.

C. Results

The proposed method was evaluated with walking image sequences from the OU-ISIR Gait Database, the treadmill dataset D [9]. We used 200 sequences from 100 subjects for evaluation, while we used 170 sequences from 85 subjects for training.

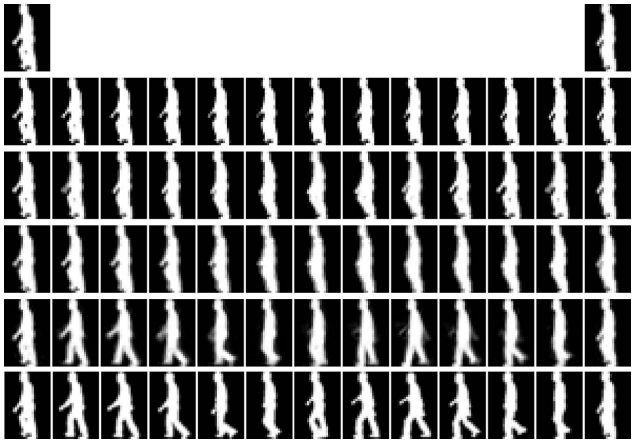
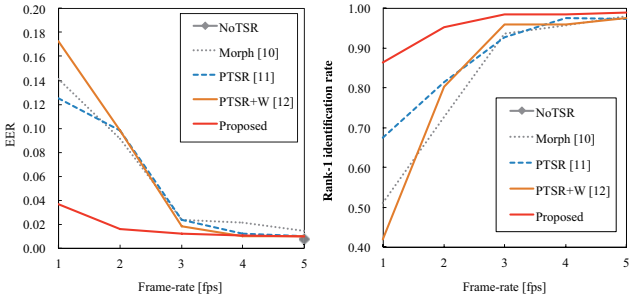


Fig. 9. Result of temporal super resolution. The top row: input low frame-rate video (1 fps), the 2nd row: Morph [10], the 3rd row: PTSR [11], the 4th row: PTSR+W [12], the 5th row: proposed method, and the bottom row: ground truth (60 fps) video (every 4 frames).



(a) EER (b) Rank-1 identification rate

Fig. 10. EER and rank-1 identification rate for low frame-rate videos. The horizontal axis represents frame-rates of both the probe and gallery sequences.

The benchmarks are direct matching with no TSR (No TSR), level-set morphing (Morph) [10], PTSR [11], and PTSR using exemplar image sequence and morphing (PTSR+W) [12]. Figure 9 shows results of temporal interpolation or super resolution from a 1-fps image sequence (the first row). The sampling rate of this low frame-rate sequence is such that we could observe the same phase of a gait cycle repeatedly (stroboscopic effect). While the other benchmarks failed under such a situation, the proposed method successfully reconstructed the high frame-rate image sequence.

The EER and rank-1 identification rates of the proposed method along with benchmark methods are shown in Fig. 10, respectively. The proposed method outperforms all of the benchmark methods, and the improvement is much more significant particularly at quite low frame-rates.

VI. Conclusion

We described methods of robust gait recognition with (1) an appearance-based view transformation model, (2) a kinematics-based speed transformation model, and (3) a technique of periodic temporal super resolution from a low frame-rate video. A common idea for all the methods is leveraging the prior knowledge obtained from the variations of the cooperative training subjects to overcome the difficulty of the cross-condition matching in gait recognition.

Future avenues of research are gait recognition under covariate transition (e.g., speed transition or view transition) within an image sequence, and applications of video-based gait analysis to other fields such as medical, health, and sport sciences.

Acknowledgment

This work was partly supported by JSPS Grant-in-Aids for Scientific Research (S) 21220003, Young Scientists (A) 23680017, "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society", Strategic Funds for the Promotion of Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government., and the JST CREST "Behavior Understanding based on Intention-Gait Model" project.

References

[1] 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 Conf. on Computer Vision*, Graz, Austria, May 2006, pp. 151–163.

[2] 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.

[3] N. Akae, A. Mansur, Y. Makihara, and Y. Yagi, "Video from nearly still: an application to low frame-rate gait recognition," in *Proc. of the 25th IEEE Conf. on Computer Vision and Pattern Recognition (CVPR2012)*, Providence, RI, USA, Jun. 2012, pp. 1537–1543.

[4] Y. Makihara and Y. Yagi, "Silhouette extraction based on iterative spatio-temporal local color transformation and graph-cut segmentation," in *Proc. of the 19th Int. Conf. on Pattern Recognition*, Tampa, Florida USA, Dec. 2008.

[5] R. Sagawa, Y. Makihara, T. Echigo, and Y. Yagi, "Matching gait image sequences in the frequency domain for tracking people at a distance," in *Proc. of the 7th Asian Conf. on Computer Vision*, vol. 2, Jan. 2006, pp. 141–150.

[6] A. Utsumi and N. Tetsutani, "Adaptation of appearance model for human tracking using geometrical pixel value distributions," in *Proc. of the 6th Asian Conf. on Computer Vision*, vol. 2, 2004, pp. 794–799.

[7] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Which reference view is effective for gait identification using a view transformation model?" in *Proc. of the IEEE Computer Society Workshop on Biometrics 2006*, New York, USA, Jun. 2006.

[8] A. Kale, A. Roy-Chowdhury, and R. Chellappa, "Towards a view invariant gait recognition algorithm," in *Proc. of IEEE Conf. on Advanced Video and Signal Based Surveillance*, 2003, pp. 143–150.

[9] 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 Trans. on Computer Vision and Applications*, vol. 4, pp. 53–62, Apr. 2012.

[10] M. S. Al-Huseiny, S. Mahmoodi, and M. S. Nixon, "Gait learning-based regenerative model: A level set approach," in *The 20th Int. Conf. on Pattern Recognition*, Istanbul, Turkey, Aug. 2010, pp. 2644–2647.

[11] Y. Makihara, A. Mori, and Y. Yagi, "Temporal super resolution from a single quasi-periodic image sequence based on phase registration," in *Proc. of the 10th Asian Conf. on Computer Vision*, Queenstown, New Zealand, Nov. 2010, pp. 107–120.

[12] N. Akae, Y. Makihara, and Y. Yagi, "Gait recognition using periodic temporal super resolution for low frame-rate videos," in *Proc. of the Int. Joint Conf. on Biometrics (IJCB2011)*, Washington D.C., USA, Oct. 2011, pp. 1–7.