21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

View-invariant Gait Recognition from Low Frame-rate Videos

Al Mansur, Yasushi Makihara, and Yasushi Yagi
ISIR,Osaka University, Japan
{*mansur, makihara, yagi*}*@am.sanken.osaka-u.ac.jp*

Abstract

In this paper, we introduce a torus manifold-based temporal super resolution method for gait recognition from low frame-rate videos with view transitions. Given a low frame-rate gait sequence with view transition from an unknown person, we estimate three unknowns: view, phase, and style. We estimate view by walking trajectory and camera information, phase by dynamic programming using multiview exemplar sequences, and style by bilinear model and linear least squares. Once these parameters are known, we can synthesize a high frame-rate sequence corresponding to that unknown person and can use existing methods for gait recognition. Experiments with OU-ISIR multiview gait dataset demonstrate the effectiveness of the proposed method for frame-rates as low as 1 or 2 fps.

1. Introduction

Human gait is a promising biometric feature for surveillance systems that can be efficiently recognized at a distance, and can be applied to uncooperative subjects. With rapid developments of computer vision technology, vision-based gait recognition has recently gained considerable attention from the biometrics field [12, 4].

Gait recognition using a low frame-rate video is a significant problem, and due to the sparsity of the observed gait phases, existing gait recognition methods do not perform well. This problem is commonly faced in CCTV cameras where the video is recorded at a quite low frame-rate (e.g., 1 to 5 fps) due to limited transmission bandwidth and storage capacity.

Several methods have been proposed [2, 9, 3] to overcome the problem in low frame-rate gait recognition. Although high frame-rate videos are used as gallery, Mori et al. [9] used low frame-rate videos as probe sequence. As a result, this approach fails when both probe and gallery are obtained from low frame-rate videos. In [3], temporal interpolation using a level-set approach is proposed as a solution to the low frame-rate gait recognition problem. However, performance of these methods are not satisfactory in quite low frame-rate videos (e.g., 1 or 2 fps). To overcome the limitations of the previous approaches, Akae et al. [2] proposed a periodic temporal super resolution (TSR) method and demonstrated the effectiveness of the method particularly in quite low frame-rate videos (less than 5 fps). However, all of these approaches assumed that a gait sequence will be observed without any view change during a gait cycle. While observation view of the walking person often changes during capturing, a sufficient length of video is required for temporal super resolution. Therefore, such an assumption is often violated in a real surveillance scene. In addition, view change within a gait period degrade gait recognition performance significantly as reported in [1].

In [5], Lee et al. used multilinear models for view-invariant gait recognition. However, this method requires a high frame-rate complete gait cycle and did not consider view change within a sequence. In [6, 7], Lee et al. separates style and content using a torus manifold for human motion tracking under view change. These methods estimate the phase and style using a single frame and hence, does not guarantee consistent style parameters for an unknown test subject.

In this paper, using a torus manifold, we propose a method for gait recognition from low frame-rate videos with view transition. Given few low frame-rate observed silhouettes, we estimate the gait phases using dynamic programming (DP). Then using the estimated phases, style parameters of the unknown person is estimated using linear least squares. This procedure guarantees consistent style parameters and smooth phase evolution. Finally, we synthesize high frame-rate gait sequences with arbitrary views using back projection from the torus manifold based on the reconstructed subject-specific mapping functions.

2. Manifold Learning

The joint manifold representing gait phase and view is modelled with a torus manifold [7] with two orthogonal coordinate axes: one for view and the other for gait phase. The torus manifold is a supervised technique compared to unsupervised techniques (e.g. GPLVM, LLE, ISOMAP) and has several advantages such as low dimensionality, continuity, and unambiguity [7].

A nonlinear mapping function between points on the

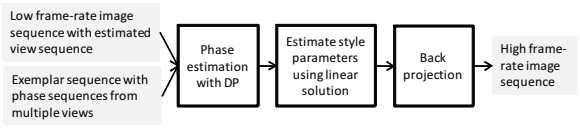


Figure 1. Overview of the phase and style parameter estimation.

torus and the input sequences containing all views and gait phases is learnt using [7]. For the completeness of the paper, we briefly discuss the method here. Given a set of points $\{\boldsymbol{x}_i \in \mathbb{R}^3\}(i = 1, \ldots, M)$ on the torus, and their corresponding observations $\{\boldsymbol{y}_i \in \mathbb{R}^d\}(i = 1, \ldots, M)$, a nonlinear mapping function $g : \mathbb{R}^3 \rightarrow \mathbb{R}^d$ is learnt using radial basis function (RBF) interpolation, and can be written in the form [7]:

$$\boldsymbol{y}_i = D\psi(\boldsymbol{x}_i), \tag{1}$$

where D is a $d \times N$ coefficient matrix and $\psi(\boldsymbol{x}_i) = [k(\boldsymbol{x}_i, \boldsymbol{z}_1), \ldots, k(\boldsymbol{x}_i, \boldsymbol{z}_N)]^T$. Here, $\{\boldsymbol{z}_j\}(j = 1, \ldots, N)$ are a finite set of representative points on the torus as kernel centers, not necessarily corresponding to the data points, and $k(\cdot, \cdot)$ is the RBF kernel. The coefficient matrix D is computed by solving a linear system in the form $[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M] = D[\psi(\boldsymbol{x}_1), \ldots, \psi(\boldsymbol{x}_M)]$.

Given learned nonlinear mapping coefficients $\{D^1, D^2, \ldots, D^{N^p}\}$, corresponding to N^p persons' gait, using SVD, the shape style parameters are decomposed as $[\boldsymbol{d}^1, \boldsymbol{d}^2, \ldots, \boldsymbol{d}^{N^p}] = AS$, where $\boldsymbol{d}^p$ is the dN-dimensional vector representation of the matrix D^p. Now a person dependent generative model for silhouettes at a particular view, v and phase, b can be given as a tensor product form:

$$\boldsymbol{y}^s_{vb} = \mathcal{A} \times_1 a^s \times_2 \psi(\boldsymbol{x}_{vb}), \tag{2}$$

where $\mathcal{A}$ is a third-order tensor with dimensions $d \times S \times N$, where S is the dimensionality of the gait style space. a^s is the style vector that generates an observation $\boldsymbol{y}^s_{vb}$ with style s, view v, and phase b. $\boldsymbol{x}_{vb}$ is a point on torus manifold with view v and phase b, and the notation $\times_n$ indicates the tensor multiplication.

3. Estimation of View, Phase, and Style

To reconstruct a high frame-rate video from a low frame-rate and view transited video, we need to estimate the views, phases, and the style parameters from the observations. The overview of the phase and style parameter estimation process is shown in Fig. 1. Estimation of the correct style parameters depends on the accuracy of the phase estimation. Once the style parameters are found, we can synthesize a high frame-rate video with any phase and any of the training views. The procedure is fast, non-iterative, and accurate.

3.1 Estimation of View

Following [11], we estimate the view direction of each input frame. However, instead of omnidirectional camera, we use limited viewing angle camera in our experiments.

3.2 Estimation of Phase

In the phase estimation step, DP [2] is used to estimate the phases of the low frame-rate input frames. However, instead of using single exemplar sequence from the sagittal view, we use multiple exemplar sequences from different persons with a set of views. This results in a very accurate phase estimation.

Continuous DP is applied directly between an input low frame-rate image sequence $Y^{in} = \{\boldsymbol{y}^{in}_i\}(i = 1, \ldots, N^{in})$ with an estimated view sequence $\boldsymbol{v}^{in} = \{v^{in}_i\}$ and $N^{ex} \times N^{view}$ exemplar high frame-rate image sequences $Y^{ex} = \{\boldsymbol{y}^{ex}_{k,l,m}\}$ with corresponding phase sequences $B^{ex} = \{b^{ex}_{k,m}\}$ $(k = 1, \ldots, N^{ex}, l = 1, \ldots, N^{view}, m = 0, \ldots, N^{frame})$, where N^{ex} is the number of exemplar subjects, N^{view} is the number of discrete views, and N^{frame} is the number of frames in a sequence. Moreover, note that the view index l is missing for phase sequences because the phase are synchronized across N^{view} discrete views $\boldsymbol{v}^{ex} = \{v^{ex}_l\}$.

The phase of the input frames, $\boldsymbol{b}^{in} = [b^{in}_1, \ldots, b^{in}_{N^{in}}]$ is found by the DP algorithm described in algorithm 1. Input to the algorithm is low frame-rate image sequence Y^{in} with estimated views $\boldsymbol{v}^{in} = [v^{in}_1, \ldots, v^{in}_{N^{in}}]$ and exemplar sequences Y^{ex} with accompanying phase sequences B^{ex}. Corresponding to N^{ex} different exemplars, we get N^{ex} different phase sequences for the input sequence. Finally, a phase sequence giving minimum cumulative cost is chosen as the optimal phase sequence $\boldsymbol{b}^{in}_{opt}$ among them. Here, Δb_{min} and Δb_{max} are the allowable phase transition range [cycle/frame] dependent on the allowable gait period.

3.3 Estimation of Style Parameters

We estimate the style parameter of the unknown subject by linear solution which guarantees an optimal style parameter set in the least square sense. To do this, Eqn. (2) is rearranged as

$$\boldsymbol{y}^s_{vb} = H_{v,b}\boldsymbol{w}, \tag{3}$$

where $H_{v,b}$ is a rearranged mapping matrix for view v and phase b with dimension $d \times K$ and $\boldsymbol{w} = [w_1, \ldots, w_K]^T$ is a K dimensional vector containing the weights of the K learned style vectors with a constraint $\sum_{i=i}^{K} w_i = 1$. Now given N^{in} input frames, style coefficients are found by solving the following linear system:

$$[\boldsymbol{y}^T_1, \ldots, \boldsymbol{y}^T_{N^{in}}]^T = H\boldsymbol{w}, \tag{4}$$

Alg. 1 Pseudocode for phase estimation.

Input: Y^{ex}, B^{ex}, $\boldsymbol{v}^{ex}$, Y^{in}, and $\boldsymbol{v}^{in}$
Output: estimated phase sequence $\boldsymbol{b}^{in}_{opt}$
begin
for $k = 1$ to N^{ex} **do**
  for $i = 1$ to N^{in} **do**
    $\mu = \arg\min_{l} |v^{ex}_{l} - v^{in}_{i}|$
    for $m = 1$ to N^{frame} **do**
      if $i = 1$ **then** $c_k(1, m) = ||y^{ex}_{k,\mu,m} - y^{in}_{i}||^2$ **else**
      $p^*(i, m) = \arg\min_{n} c_k(i-1, n)$
      (where $n \in \{n | \Delta b_{min} < b^{ex}_{k,m} - b^{ex}_{k,n} < \Delta b_{max}\}$)
      $c_k(i, m) = c_k(i-1, p^*(i, m)) + ||y^{ex}_{k,\mu,m} - y^{in}_{i}||^2$
    end
  end
  $C_k = \min_{m} c_k(N_{in}, m)$
  $q^*_k(N^{in}) = \arg\min_{m} c_k(N^{in}, m)$
  for $i = N^{in}$ to 2 **do**
    $q^*_k(i-1) = p^*(i, q^*_k(i))$
  end
end
$n = \arg\min_{k}\{C_k\}$
for $i = 1$ to N^{in} **do**
  $b^{in}_{opt,i} = b^{ex}_{n,q^*_n(i)}$
end

where $H = [H^T_{v_1,b_1}, \ldots, H^T_{v_{N^{in}},b_{N^{in}}}]^T$ is a matrix formed by concatenating each mapping matrices.

4. Experimental Results

We evaluated the proposed method with real data from the OU-ISIR Gait Dataset [10]. Gait was performed on a treadmill which was simultaneously captured by 25 synchronized cameras, and we used 5 of them in our experiments. These 5 cameras were located at the same height and covered view directions from $-60°$ to $60°$ at $30°$ interval. We use background subtraction-based graph-cut segmentation to extract gait silhouette images. Then scaling and registration of the extracted silhouette images is carried on. The normalized gait silhouette size, and frame rate for each sequence were 44×64 pixels, and 60 fps.

We used 57 training sequences for manifold learning, and 32 exemplar sequences for DP based phase estimation. For evaluation, we used gait sequences from 26 subjects with view change, different from those used in learning the manifold and in exemplars. Low frame-rate gait sequences were constructed by down sampling the 60 fps gait sequences at regular intervals. We used two different phase differed and down sampled version of gait sequences for gallery and probe. The resulting sequences nicely simulates the low frame-rate sequence recorded by surveillance cameras and serve as a dataset for view transited and low frame-rate gait sequence.

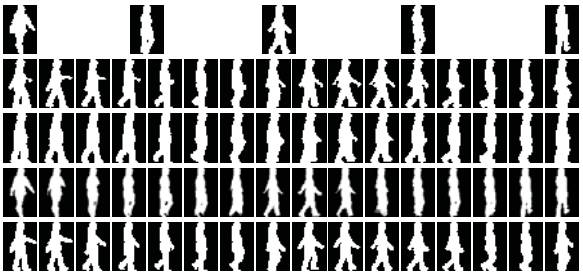


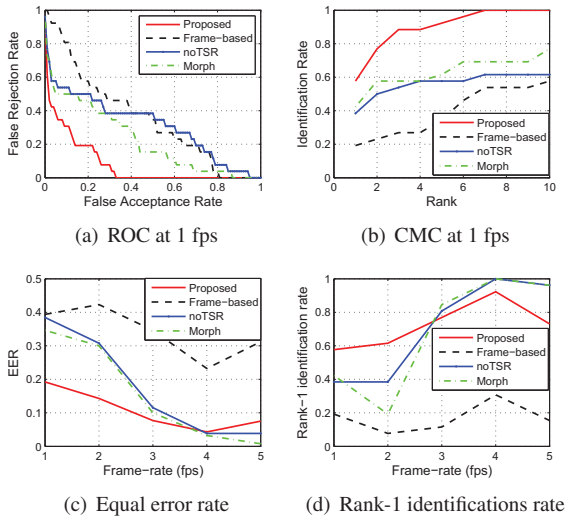
Figure 2. Temporal super resolution results: 1st row: input low frame-rate image sequence (3 fps), 2nd row: TSR by proposed method, 3rd row: TSR using [6], 4th row: TSR using [3], and 5th row: ground truth.

4.1 Temporal Super Resolution

In Fig. 2, we compare the TSR results of the proposed method with those obtained with frame-based method (*Frame-based*) [6], and morphing-based temporal interpolation (*Morph*) [3]. *Frame-based* [6] uses single frame-based phase and style estimation, and hence, we use the average over these style parameters for TSR. The input sequence contains five frames each with different view direction and captured at 3 fps. Using this sequence, we reconstruct 34 frames/gait cycle sequence in sagittal view. We can see that *Frame-based* method cannot successfully estimate the style parameter, and, therefore, the synthesized images differ considerably from the ground truth. *Morph* based reconstruction is also quite different from the ground truth. On the other hand, images synthesized by the proposed method are almost similar to the ground truth images.

4.2 Gait Recognition

In this experiment, we evaluate the proposed method in terms of gait recognition performance and compare it with the results obtained by *Frame-based* [6], *Morph* [3], and without temporal super resolution (*noTSR*). In our experiments, both the gallery and the probe are low frame-rate sequences. At first, we reconstruct high frame-rate probe and gallery sequences in case of the proposed method, *Frame-based*, and *Morph*. In addition, for the proposed method, and *Frame-based* method, we reconstruct probe and gallery sequences in 5 different views. Matching is done using average silhouette feature [8], and distance between probe and gallery is computed between the corresponding views of probe and gallery. Finally we consider the average distance for gait recognition. In case of *noTSR*, only the



(a) ROC at 1 fps (b) CMC at 1 fps

(c) Equal error rate (d) Rank-1 identifications rate

Figure 3. Gait recognition performance.

given frames are used for computing the single average silhouette. The performance is measured in terms of receiver operating characteristics (ROC), with equal error rate (EER), and cumulative match characteristic (CMC) with rank-1 identification rate (see Fig. 3). From these results, it is clear that the proposed method outperforms the baseline methods at quite low frame-rates (e.g., 1 to 3 fps).

5. Conclusion

This paper introduced a TSR based gait recognition method for low frame-rate videos with view transitions. This TSR requires estimation of view, phase, and style parameters from an input sequence. View is simply estimated from the camera information, and for phase estimation, we employed multiple exemplar high frame-rate sequences with known phases. Finally, we used linear least square solution for style parameter estimation. The proposed method can synthesize high frame-rate sequence from a view transited sequence with a frame-rate as low as 1 fps. Experimental results using OU-ISIR gait dataset verified the effectiveness of the proposed method in terms of gait recognition performance and visual quality of the reconstructed gait sequences.

Acknowledgement

This work was partially supported by Grant-in-Aid for Scientific Research (S) 21220003 and "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.

References

[1] N. Akae, Y. Makihara, and Y. Yagi. The optimal camera arrangement by a performance model for gait recognition. In *Automatic Face and Gesture Recognition*, pages 292–297, 2011.

[2] N. Akae, A. Mansur, Y. Makihara, and Y. Yagi. Video from nearly still: an application to low frame-rate gait recognition. In *CVPR*, Rhode Island, USA, June 2012.

[3] M. S. Al-Huseiny, S. Mahmoodi, and M. S. Nixon. Gait learning-based regenerative model: A level set approach. In *ICPR*, pages 2644–2647, 2010.

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

[5] C.-S. Lee and A. Elgammal. Towards scalable view-invariant gait recognition: Multilinear analysis for gait. In *AVBPA*, pages 395–405, NY, USA, 2005.

[6] C. S. Lee and A. Elgammal. Simultaneous inference of view and body pose using torus manifolds. In *ICPR*, volume 3, pages 489–494, Hong Kong, Aug. 2006.

[7] C.-S. Lee and A. Elgammal. Tracking people on a torus. *TPAMI*, 31(3):520–538, March 2009.

[8] Z. Liu and S. Sarkar. Simplest representation yet for gait recognition: Averaged silhouette. In *ICPR*, volume 1, pages 211–214, 2004.

[9] A. Mori, Y. Makihara, and Y. Yagi. Gait recognition using period-based phase synchronization for low frame-rate videos. In *ICPR*, pages 2194–2197, 2010.

[10] Ou-isir gait database. http://www.am.sanken.osaka-u.ac.jp/GaitDB/index.html.

[11] K. Sugiura, Y. Makihara, and Y. Yagi. Gait identification based on multi-view observations using omnidirectional camera. In *ACCV*, pages 452–461, 2007.

[12] C. Wang, J. Zhang, J. Pu, X. Yuan, and L. Wang. Chrono-gait image: a novel temporal template for gait recognition. In *Proceedings of the 11th European conference on Computer vision: Part I*, ECCV'10, pages 257–270, 2010.