RELIABLE SEGMENTATION OF PEDESTRIANS IN MOVING SCENES

Yang Ran, Qinfen Zheng, Isaac Weiss and Larry. S Davis
Center for Automation Research,
University of Maryland,
College Park, MD 20742-3275, USA
{rany, qinfen, weiss, lsd}@cfar.umd.edu

ABSTRACT

This paper describes a periodic motion based pedestrian segmentation algorithm for videos acquired from moving platforms. Given a sequence of bounding boxes containing the detected and tracked walking human, the goal is to analyze the low dimension structure by considering every object sample as a point in the high dimensional manifold space and use the learned structure for segmentation. In this work, unlike the traditional top-down dimension reduction (manifold learning) methods such as Isomap and locally linear embedding (LLE) [9], we introduce a novel bottom-up learning approach. We represent the human stride as a cascade of models with increasing parameter numbers. These parameters describe the dynamics of pedestrians from coarse to fine. By applying the learned manifold structure, we can predict the location of body parts, especially legs, with high accuracy at every frame. The segmentation in consecutive images is done by EM clustering. With the accuracy for prediction using twin-pendulum model, EM is more likely to converge to global maximums. Experimental results for real videos are presented. The algorithm has demonstrated a reliable performance for videos acquired from moving platforms.

1. INTRODUCTION

1.1 Motivation

Many surveillance applications involve analyzing video data of humans activities. To accurately describe the location, action, and appearance of humans as well as activities under video surveillance, we need to segment the people from the scene. Objects in image sequences could be considered as points in high dimension space. They form a manifold with an intrinsic structure. Researchers have proposed many methods [1,2] to understand the manifolds of different modalities and then apply the learned dimension-reduced structure to various applications. We are particularly interested in segmenting pedestrians from video data captured by moving cameras. A good segmentation will allow us to better estimate the limb sizes, posture, and orientations, to identify people by their walking characteristics, and to understand activity [4].

In recent years, manifold representation of complicated modalities has become popular because of its non-linearity. To illustrate our work in such a framework, consider the walking cycle in a high dimension space. Even with appeared non-rigidity, self occlusion and significant local variation, it is expected that the manifold has a low dimensional structure constrained by periodicity and the specific pattern of gait. The periodic motion signature of walking humans has been widely used in gait recognition and related applications [1,2,3,4,8] and acts as the intrinsic constraint connecting every point in the space. Yet few methods attempt segmentation. This could be partially explained by the fact that those learning methods do not make use of prior knowledge. When enough knowledge of the manifold structure is available, a bottom-up method could be used for better and faster learning. In our case, one can use a background subtraction to segment moving objects from a static background. This will be very much likely to fail because of camera motion.

1.2 Related Work

Among the many pedestrian segmentation methods, motion signature analysis is a simple and promising approach. Periodic motion is robust clue in these situations. A good review for this topic can be found in [1,2]. Several methods have been proposed for measuring the periodicity of human motion. Allmen and Dyer [1] presented a 3-D based detection scheme in curvature space. Polana and Nelson [2] used a Discrete Fourier Transform (DFT) based approach. Efros and Berg [2] identified the cyclic motion in the optical flow domain. A method closely related to this paper can be found in [3]. The authors used correlation of image pairs to calculate a similarity matrix. Every entry in the matrix represents the similarity between two images of the same object. The periodic property appears as darker lines parallel to the diagonal line detected by Short Time Frequency Analysis. One problem with the above approach is too sensitive to object misalignment as well as changing background. Other method relying on contour extraction is also sensitive to the quality of silhouette, which is a challenging task especially when the camera is moving.

1.3 Brief Algorithm Overview

This paper uses a bottom-up approach to learn the low dimension structure of the pedestrian manifold as shown in Figure 1. The method starts with a finite frequency probing method to extract the walking period as the first Degree Of Freedom (DOF). According to the different

periodicity, pixels are clustered into torso, background, and legs. A twin-pendulum model is fitted in a coarse to fine manner to the whole sequence as more and more DOFs are available. With the increase in the number of parameters, the simplified model (structure) incrementally better resembles the high dimension manifold. After reduction to only a few DOFs, we use an EM method to segment the pedestrians at every frame using the model fitting results. The method is *efficient* due to low computation cost.

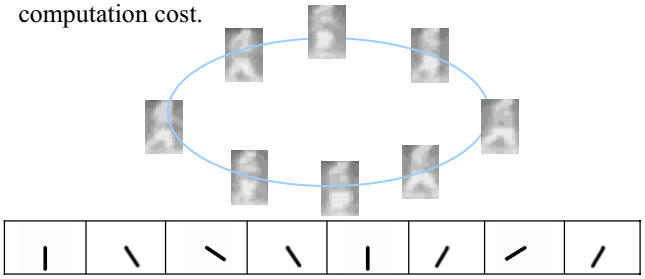

Figure 1. Pedestrian walking manifold: a complete cycle and the corresponding phase compared to a pendulum

We assume that the moving objects have been detected and tracked for at least two periods. They are specified by bounding boxes in each frame. During the observed time period, the human poses and sizes remain almost the same. This paper is organized as follows. The second section starts to analysis the pedestrian manifold with only one DOF: period. Section 3 extracts the motion signature with more DOFs. We propose a twin-pendulum model featured by 5 DOFs to describe pedestrian walking. With totally 6 DOFs, this work finally segments the object and the result is shown and analyzed in section 4 and 5.

2. PERIODICITY ANALYSIS BY FINITE FREQUENCIES PROBING

2.1 Periodic Signal Probing

The key idea for understanding human walking lies in the periodicity. Many researchers have done considerable work on gait analysis [8]. When we are asked how a human walks, the first answer we come up with might be how fast he/she walks, or the time/frames he/she spends in one stride cycle. The specific motion parameter, i.e., the period, reveals the nature of walking. This corresponds to using a predefined periodic signal, such as a cosine, to represent every point in the image stack.

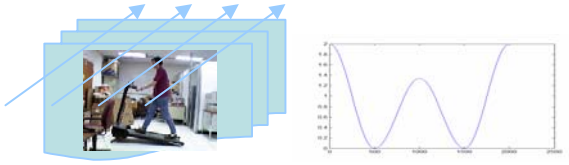

Figure 2. Illustration of Probing and reference signal

Based on this observation, our method uses a reference signal to probe the original target signal at every pixel along time axis. The probing is defined as follows.

Definition 1*: Probing is a process of matching a periodic reference signal to the target signal to obtain a measure of the period's amplitude.*

We start with a quasi-periodic signal $W(t) \approx W(t+nT)$, where T is the quasi-period of the signal. If we use a temporal window to truncate the sample of W(t), we will get a vector $\overline{W}(t) = [W(t), W(t+\tau)....W(t+(N-1)\tau)]$. Given a reference signal W' and under additive Gaussian noise, the following *a posteriori* probability is maximized

$$P_{W|H_T}(W(t)\,|\,H_T) = \frac{1}{(2\pi\sigma^2)^{N/2}}\left(-\frac{\|W(t)-s(t,T)W'(t)\|^2}{2\sigma^2}\right) \quad (1)$$

where s(t,T) is a scaling function and H_T is the hypothesis that period is T. Partial differentiation gives

$$\frac{\partial}{\partial s(t,T)}\left(-\frac{\|W(t)-s(t,T)W'(t)\|^2}{2\sigma^2}\right) = 0 \Rightarrow s(t,T) = \frac{W(t)W'(t)}{\|W(t)\|\|W'(t)\|} \quad (2)$$

Using Bayes rule we have

$$P_{H_T|W}(H_T\,|\,W(t)) = \frac{P_{s(t)\|H_T}P_{H_T}}{P_{W(t)}} \quad (3)$$

The best estimate of the quasi-period is the frequency which maximizes the cross correlation $C(W,W')$.

2.2 Probing with Finite Frequency Sets

The output of the detecting/tracking module gives bounding boxes for one object. After cropping, we have a stack of rectangles with the same size and object center. A probing function with period ω and waveform k is defined as $\Phi_k(\omega) = k(\omega t)$. In practice, we have limited length and size and the discrete version is:

$$C(k,\omega) = \sum_{x=1}^{X}\sum_{y=1}^{Y} cor(\Phi_k(\omega), W(n,x,y)) \quad (4)$$

Our goal is to calculate the overall correlation of the signal W by summing up the value at each location (x, y) with the same reference signal Φ at frequency ω. The period is defined from (3) and (4) as the ω in $(\omega_1, \omega_2, \ldots, \omega_m)$, which maximizes the averaged response function (4).

$$period = \arg\max P_{H_T|W}(H_T\,|\,W(t)) = \arg\max_{\varpi\in\{\varpi 1,\varpi 2...\varpi m\}} C(k,\omega)$$
$$= \arg\max_{\varpi\in\{\varpi 1,\varpi 2...\varpi m\}} \sum_{x=1}^{X}\sum_{y=1}^{Y} cor(\Phi_k(\omega), W(n,x,y)) \quad (5)$$

2.3 Periodicity Detection

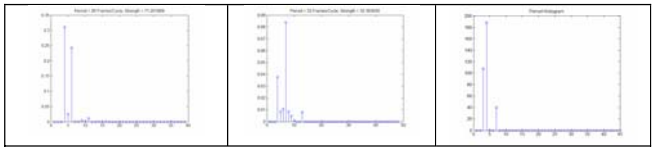

Figure 3. Histogram of maximum response frequencies

To detect the period precisely, we histogram the frequencies corresponding to each pixel's maximum period response to extract the first order statistics. As shown in the figure 3, typical histograms after smoothing have well-pronounced twin peaks at the right frequency corresponding to period and half period. Many related frequency estimation and tracking techniques could be found in [5]. Fisher's Test is used here and is proved to be robust in performance and fast in implementation.

2.4 Alignment

As discussed above, we want to explore the periodicity along temporal axis for every pixel to vote for the global period. Thus an accurate alignment is required. Notice that the videos are from moving platforms and hence changing background causes the traditional correlation based alignment algorithm to fail. Pedestrians' walking is an articulated activity containing non-rigid motion and self-occlusion. If we treat all pixels equally, the change in the background will pollute the correlation function if the two frames are far apart.

In this paper, instead of treating all pixels equally as before, we only consider 'stable' pixels. If we revisit the period probing process, we may find that pixels could be divided into 3 major categories:

1. Pixels from background with irregular change;
2. Pixels from lower body part with periodic change;
3. Pixels from torso part with no dramatic change.

By focusing on the last class we are able to solve the alignment problem. If we can make use of pixels which are common for every frame, the same procedure could be used because these torso pixels can be regarded as evolving under pure translation between frames (if the pedestrian does not change his/her pose much). Those pixels belonging to the torso part will be clustered together because they do not behave periodically and only change within a small range of gray value.

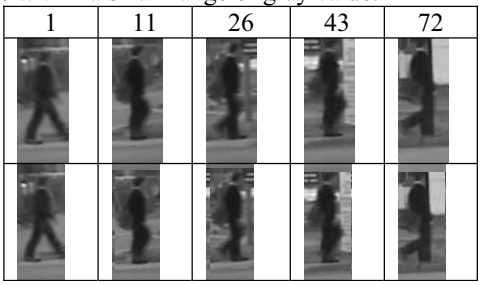

Figure 4. Alignment result for walking sequence 5

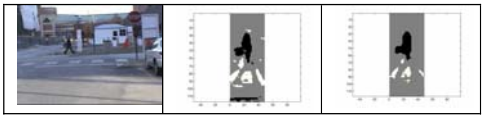

Figure 5. Pixels clustering, the black ('stable') pixels are those segmented torso pixels in the first and last iterations.

From the above figure, only torso pixels are used for alignment. In pratice, due to unavoidable pose change, and stride variation, alignment could not be achieved in one step. So we use an iterative way to find the optimal solution. The process ends when overall change between adjacent iterations is below a threshold or maximal iteration number is reached.

3. EXTRACTING MOTION SIGNATURE

3.1 Model Selection

The combination of human motion analysis and biometrics has become a promising approach for visual surveillance. The human gait has such characteristics which make it a particularly useful behavioral feature for personal identification at a distance. For gait analysis, one main issue is human segmentation from complex background. Many model-based method has been proposed [1,2] such as stick figure, 2-D contour or volumetric models.

We use a simplified version of 2-D stick model [1]. It is observed that walking has both periodicity and symmetry due to the legs and arms. Figure 6(a) [2] suggests a twin-pendulum model. Figure 6(b) illustrates the model we use [2]. Each leg is represented by two segments with same width, different angle and length at various times, i.e. $(l_1, \theta_1, l_2, \theta_2, w)$ while the center of the twin-pendulum is middle bottom point of the extracted torso in Fig 5.

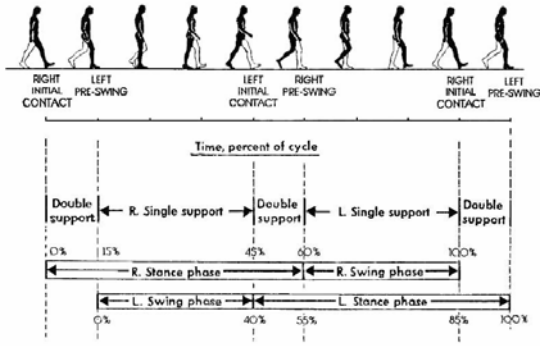

Figure 6. a) Illustration of period/symmetry of walking.

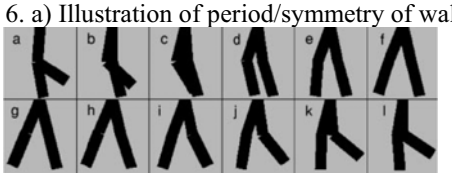

Figure 6. b) Twin pendulum model of a walking human

We investigated the similarity between pedestrians. After simple smoothing, they have high similarity. This makes it possible to predict the location of the legs using a small set of parameters. In our experiment, every parameter is trained using K-means method for the mean value and variance at each frame.

3.2 Object Segmentation

After predicting the leg's location at every frame, we use an EM method. EM segmentation models the image intensities as visible variables, Y, classifications as hidden variables, Γ, and the bias field as governed by model parameters, β. We would like to choose the parameters that maximize the log likelihood of the data, log p(Y, Γ|β), but we do not know this likelihood because β 's invisibility renders p(Y, Γ|β) to be a random variable. Thus, although we cannot maximize it, we can *maximize* its *expectation*. This results in the following two iterative steps until convergence to a local minimum [12].

E-Step*: Compute the expectation* $\Sigma_{\Gamma} p(\Gamma|Y,\beta)\log p(\Gamma,Y|\beta)$ *using the current* β.

M-Step*: Find new* $\beta_{(t+1)}$ *to maximize the expectation, assuming* $p(\Gamma|Y, \beta_t)$ *is correct.*

EM method will converge to local minimums in many cases. But in our work, since we make the advantage of the twin-pendulum model as prediction. It is guaranteed to converge to the real location in almost every frame.

4. EXPERIMENTAL RESULTS

We have evaluated the algorithm on several video sets of pedestrians in different backgrounds and lighting conditions, from both static and moving platforms.

In the first experiment, we present the segmentation for a sequence where a pedestrian is walking across a street when the camera is approaching the pedestrian. The video length is 80 frames corresponding to about two cycles of normal walking. The first DOF is the period detected using probing. The DOFs $(l_1, \theta_1, l_2, \theta_2, w)$ are those for the twin-pendulum model fitting the manifold. With more DOFs, the model is more elaborate and is able to describe pedestrian more accurately. The result is in Figure 7.

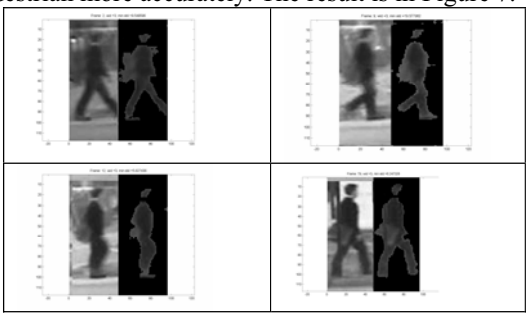

Figure 7. Segmentation for Seq 5, frame. 11, 44, 68, 79

In the second experiment we show the result of the segmentation when a pedestrian is walking on the parking lot. Similar result is shown in Figure 8.

Figure 8. Segmentation for Seq 7.

5. SUMMARY AND DISCUSSION

A periodicity based object segmentation algorithm is reported here in the manifold learning framework. Manifold learning has become an non-linear extension of linear learning. Some representative algorithms include Isomap, LLE and Semidefinite Programming [9]. Almost all of these fall in the top-down learning category. They start from the manifold space and then reduce the dimensionality according to some criteria or cost function. In contrast, this paper proposes a novel bottom-up method in the manifold learning framework. Starting from the simplest degree of freedom, the period in our case, the algorithm gradually increases the number of parameters for the target's low-dimension space. The physical model also progressively advances towards the real structure of the manifold. The motivation behind our method is that h we know more about human gait than about other modalities. Thus top-down method is not necessary since it does not make use of the prior knowledge. The proposed bottom-up method performs better and more efficient than the top-down methods in this case.

6. REFERENCES

[1] D. M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image understanding*, Academic Press, 73(1):82-98, 1999.

[2] Liang Wang, Weiming Hu, Tieniu Tan, "Recent developments in human motion analysis," *Pattern Recognition*, 36(3):585-601, 2003.

[3] R. Cutler, L.S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE T-PAMI*, 22(8):781-796, 2000.

[4] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: Real-time surveillance of people and their activities," *IEEE T-PAMI*, 22(8):809-830, Aug. 2000.

[5] Quinn, B.G, Hannan, E.J. The Estimation and Tracking of Frequency, Cambridge Univ. Press. ISBN 0-521-80446-9 2001

[6] S.M. Seitz, C.R. Dyer, "View-invariant analysis of cyclic motion," *Intl Journal of Computer Vision*, 25:1-23, 1997.

[7] S. Kay, "A fast and accurate single frequency estimator," *IEEE T-SP*, **37**(12):1987-1990, 1989.

[8] L. Lee and W.E.L. Grimson, "Gait analysis for recognition and classification," *Proc. Fifth Intl Conf. on Automatic Face and Gesture Recognition*, 2002.

[9] S. T. Roweis, L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, 290:2323-2326, 2000.