Identifying and Segmenting Human-Motion for Mobile Robot Navigation using Alignment Errors

Wael Abd-Almageed
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
Email: wamageed@umiacs.umd.edu

Brian J. Burns
Artificial Intelligence Center
SRI International Inc.
Menlo Park, CA 94025
Email: burns@ai.sri.com

Larry S. Davis
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
Email: lsd@umiacs.umd.edu

***Abstract*— This paper presents a new human-motion identification and segmentation algorithm, for mobile robot platforms. The algorithm is based on computing the alignment error between pairs of object images acquired from a moving platform. Pairs of images generating relatively small alignment errors are used to estimate the fundamental frequency of the object's motion. A decision criterion is then used to test the significance of the estimated frequency and to classify the object's motion. To verify the validity of the proposed approach, experimental results are shown on different classes of objects.**

I. Introduction

Deciding if a moving object is a human or not is an important task in many vision-based robotic applications, such as visual navigation [1] and visual surveillance [2]. Military robotic vehicles, for example, need to determine if the moving object is a human in order to asses the threat posed by the object. On the other hand, this determination is also important for civilian robotic vehicles for safety purposes.

When the effective image resolution is relatively high, shape-based techniques, such as [1], [3], [4] and [5], can be used for initial human detection. Using this class of approaches, accurate silhouette must be first extracted and then matched against either a shape database or a set of shape models. Since it is usually difficult to extract such an accurate silhouette, particularly in cluttered environments, shape-based methods usually produce false detections.

Motion information can be used to reduce the number of false detections produced by shape-based methods. On the other hand, at relatively low object image resolutions, it becomes virtually impossible to detect humans based on their shape. At such scales, motion information is the only cue for human identification.

In stationary-camera environments, background-image subtraction (combined with morphology and connected component analysis) is usually used for finding independently moving objects. Sequences of images of the detected objects can then be analyzed for identification and segmentation purposes. This task becomes more problematic when the video stream is acquired from a moving platform, such as a mobile robot. In this case, more sophisticated approaches are needed to segment the moving object from the background in order to classify the moving object. The emphasis of this paper is identifying and segmenting human motion from videos acquired from a mobile robot platform.

The goal of the proposed method is as follows: *Given a sequence of images of a tracked object acquired from a mobile robot, we need to identify (or verify) if the moving object is a human, and if so we need to segment its motion in both the time and spatial domains*. The tracked object can be either the output of a shape-based human detector that needs to be verified, or the output of a moving object detector that needs to be identified. The proposed approach is based on computing the alignment error between pairs of object images. The alignment error is then used to estimate the fundamental frequency of the motion. Based on the estimated frequency, a decision is made on whether or not the object is a human. Experimental results show the accuracy of the estimated frequency which is effectively used to identify and segment human motion.

This paper is organized as follows. Section II discusses the previous work in the area of human motion identification. The proposed method for period estimation is presented in Section III. In Section IV, the estimated period is used for segmenting the human motion. Experimental results validating the proposed methods are introduced in Section V. The paper is concluded and directions for future research are given in Section VI.

II. Related Work

The approaches to identifying human activity from video can be broadly divided into two major categories; shape-based methods and motion-based methods. For a comprehensive review of all approaches, the reader is referred to [6]. Motion-based methods that are closely related to our approach are highlighted in this Section.

A very distinct property of human motion is its periodic nature. A walking or running human, particularly over short time intervals, has constant speed and period. Several approaches have been developed to human motion based on the periodic nature of the motion. In most of these approaches, the leg motion is usually described by a moving pendulum as shown in Figure 1.

Cutler and Davis [2] use power spectral density analysis methods to analyze 2D patterns obtained by computing the

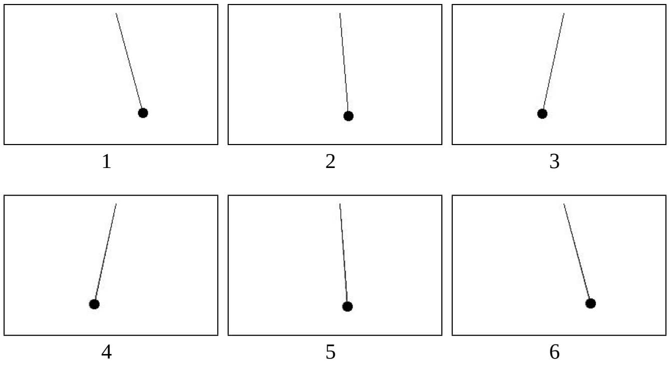


Fig. 1. Pendulum Motion; An Example of Periodic Motion

cross correlation between all pairs of object images. Ran et al. [7] correlate 1D spatio-temporal signals with a set of reference signals. The reference signals have a finite set of frequencies that represent the human gait. Periodic motion will generally resonate with one of the pre-defined reference signals, which indicates a human motion.

In [8], Liu and Collins compute 2D spatio-temporal projections of object sequences in order to generate motion patterns, similar to the mathematical Frieze patterns. In order to estimate the direction of motion, they classify the generated patterns into one of pre-defined seven Frieze pattern groups, representing different motion directions.

Little and Boyd [9] use least squares linear prediction to find the dominant frequency in a sequence of multi-dimensional feature vectors describing the moving object. Also, in [10] and [11] Boyd and Little use what is called Video Phase Locked Loops in order to estimate the fundamental gait frequency and gait phase.

In [12] and [13] Shutler et al. introduced velocity moments, a spatio-temporal modification to the well known geometric and Zernike [14] moments. Velocity moments were used to recognize human gait. Niyogi and Adelson [15] search the 2D spatio-temporal slices, in the lower half of the moving object, to find braided patterns that indicate a walking human.

III. Period Estimation

All of the methods outlined in Section II, except for [2], reported results on videos acquired from a stationary camera. They used either a human walking on a treadmill as a test subject, the data set collected by Little and Boyd in [9] or both. Clearly, this facilitates the motion segmentation process prior to applying the period estimation method. Also, all of these methods are very sensitive to initial object alignment prior to the identification process.

Many methods can be used to detect moving objects. Here, we use the object detection mechanism used by Cutler and Davis in [2]. The detected objects are tracked through the video sequence using the tracker of Zhou et al. [16]. The period estimation algorithm is then applied to the sequence of bounding boxes of the tracked object.

We start the analysis by first estimating the period of the motion of the object. A significance test is applied to the estimated period to decide if the tracked object is a human or not. If it is a human, a foreground-background segmentation process is performed across the motion sequence. In other words, we test the hypotheses

$$\begin{aligned} &\mathcal{H}_0 : \text{Object is human} \\ &\text{and} \\ &\mathcal{H}_1 : \text{Object is non-human} \end{aligned} \tag{1}$$

and accept the null hypothesis if and only if the significance test is satisfied.

A pixel, (x, y), in the bounding box of the moving object belongs to either the object or the scene background. Therefore, it may undergo one of two different types of motion; the general camera motion resulting from the robot motion and the periodic motion performed by the human. Given two images, I_t and $I_{t+\dot{p}}$, separated (in time) by the true motion period, $\dot{p}$, image pixels that do not belong to the object (i.e. that belong to the background) must obey the global transformation, $\mathbf{A}_{t,t+\dot{p}}$, such that

$$I_{t+\dot{p}}\left(\mathbf{A}_{t,t+\dot{p}}(x, y)\right) = I_t(x, y). \tag{2}$$

For a slowly moving camera, the transformation $\mathbf{A}_{i,j}$ can be a six-parameter affine transform and is estimated by applying the phase correlation method of [17] to images I_i and I_j. It is important to note here that $\mathbf{A}$ is estimated based on the entire image not only the object area.

On the other hand, for object pixels we can expect that object images separated in time by exactly the true period (or multiples of the true period) will be exactly the same (see Figures 3.1 and 3.6,) i.e.

$$I_{t+\dot{p}}\left(\mathbf{T}_{t,t+\dot{p}}(x, y)\right) = I_t(x, y) \tag{3}$$

where $\mathbf{T}_{i,j}$ is a translation-only transformation between images I_i and I_j. The transformation $\mathbf{T}$, is estimated using the normalized cross correlation introduced in [18]. Figure 2 illustrates the relationships described by Equations 3 and 2.

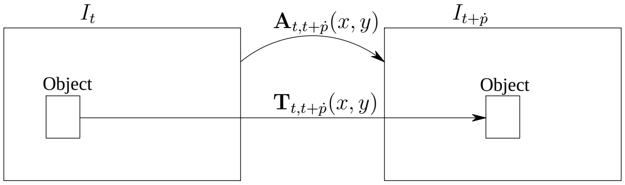

I_t
$\mathbf{A}_{t,t+\dot{p}}(x,y)$
$I_{t+\dot{p}}$
Object
$\mathbf{T}_{t,t+\dot{p}}(x,y)$
Object

Fig. 2. Transformations between I_t and $I_{t+\dot{p}}$

The probability that a pixel, (x, y), belongs to the background, B, can be expressed as

$$P\left((x, y) \in B|\mathbf{A}_{i,j}\right) \approx 1 - |I_i(x, y) - I_j(\mathbf{A}_{i,j}(x, y))|. \tag{4}$$

Also, the probability that the pixel belongs to the foreground, F, can be expressed as

$$P\left((x, y) \in F|\mathbf{T}_{i,j}\right) \approx 1 - |I_i(x, y) - I_j(\mathbf{T}_{i,j}(x, y))|. \tag{5}$$

Therefore, we define the cost function $C(x, y)$ that the pixel (x, y) belongs to the foreground as

$$\begin{aligned} & C((x, y) \in F) \\ & \quad = \min(P((x, y, i) \in F|\mathbf{T}_{i,j}), (1 - P((x, y, i) \in F|\mathbf{A}_{i,j}))). \end{aligned} \tag{6}$$

For all pixels in the object bounding box, O, we compute the alignment error between I_i and I_j, given Equation 6 as follows

$$\epsilon_{i,j} = \int_{\forall x \in O} \int_{\forall y \in O} C((x, y) \in F) \; dx \; dy. \tag{7}$$

In a video stream of length T, a moving object can only exhibit an arbitrary period from the finite range $[p_{min},\ p_{max}]$. In theory,

$$\begin{aligned} & p_{min} = 2 \; frames \\ & \text{and} \\ & p_{max} = \lfloor \frac{T}{2} \rfloor \; frames. \end{aligned} \tag{8}$$

To estimate the motion period, we compute the average alignment error (per image pair per pixel) for every period in the specified range. The average error is computed over all possible cycles and normalized by the area of the bounding box. The estimated period is that one that minimizes the average alignment error, i.e.

$$\begin{aligned} \hat{p} = \arg_p \min \frac{1}{T/p} \int_1^{T/p-1} \int_1^{p} \epsilon_{t,t+p} \quad dn \; dt, \\ p = [p_{min},\ p_{max}]. \end{aligned} \tag{9}$$

If the moving object is a human, the estimated period must produce significantly smaller alignment error with respect to the other periods. On the other hand, if the moving object is not a human (e.g. a vehicle), the alignment error of the estimated period will not significantly vary from other periods. To examine the significance of the estimated period, we use the standard Fisher significance test given by Equation 10

$$\frac{\epsilon(\hat{p}) - \mu_\epsilon}{\sigma_\epsilon} \geq \delta \tag{10}$$

where $\epsilon(\hat{p})$ is the average alignment error of the estimated period, μ_ϵ and σ_ϵ are the mean and the standard deviation of the average alignment error and δ is a manually-set (typically 1-2) threshold value. If the test is satisfied, we accept the null hypothesis, $\mathcal{H}_0$ of Equation 1. Otherwise, we accept the alternate hypothesis, $\mathcal{H}_1$.

IV. Motion Segmentation

Discriminating figure from background has long been an area of interest in computer/robot vision [19] [20]. It has significant importance in applications such as shape-based verification. It is also important for estimating accurate appearance models that can be used as a feedback to improve the tracker performance or can be used later in activity recognition applications. In a mobile camera platform these methods become error prone for many reasons, such as the continuously changing background and the camera motion.

Non of these methods exploits the periodic nature of the object's motion. Utilizing the estimated period can significantly improve the figure-background discrimination. Zhao and Davis [21] recently introduced a non-parametric, iterative method for segmenting the figure from the background. They iteratively estimate the probability $\hat{P}_{fg}(x, y)$ for all pixels in the motion detection area based on a randomly selected subset of image pixels. Equation 11 combines $\hat{P}_{fg}(x, y)$ with the cost function of the correct period in order to obtain an improved object image I_O.

$$I_O(x, y) = \hat{P}_{fg}(x, y) \; C_{\hat{p}}(x, y) \tag{11}$$

where

$$\begin{aligned} & C_{\hat{p}}(x, y) = \\ & \min(P((x, y) \in F|\mathbf{T}_{t,t+\hat{p}}), (1 - P((x, y) \in B|\mathbf{A}_{t,t+\hat{p}}))). \end{aligned} \tag{12}$$

In other words, Equation 12 performs spatial segmentation as in [21] followed by a segmentation based on the estimated period, which minimized the background segmentation errors as will be shown in Section V.

V. Experimental Results

Assuming that the frame rate of the imaging system is 30 frame per second and that the average human stride is approximately one meter per step, the theoretical lower bound of the period range, i.e. $p_{min} = 2$ frames, means that the human velocity is 30 meter per second, which is physically unrealistic. To limit the search range, we use $p_{min} = 6$ frames (i.e. 10 meters per second) which corresponds to the velocity of the fastest human on earth.

Figure 3 shows the odd frames of a 48-frame image sequence acquired from the mobile robot. The image resolution of the sequence is 320x240. In this sequence the camera motion is a combination of a left rotation and a forward motion. The Figure also shows the varying nature and the complexity of the background. Figure 4 shows the bounding boxes of two motion cycles obtained by tracking the moving object. It is important to note the varying width of the bounding boxes, as a result of the articulating nature of the object and the vehicle motion, and the amount of foreground/background intermixing. The continuously changing background cause other methods described in Section II to fail to estimate the correct period. The physical motion period is 34 frames/cycle. However, due to the bilateral symmetry of humans, the true period is 17 frames/cycle.

The average motion alignment error, ϵ, for the range of potential periods ($p_{min} = 6$, $p_{max} = 24$) is shown in Figure 5. The global minimum of the alignment error at $p = 17$ indicates a correct estimate of the true period.

Segmentation results of both the spatial-only method of Zhao and Davis [21] and the motion-improved segmentation of Section IV are shown in Figure 6. It is clear from the Figure that the spatial-only methods produce more segmentation errors because of only using color information of the given frame. On the other hand, incorporating the estimated period enhances the object segmentation. Figure 6.5 indicates that the color-based segmentation methods fail when the background changes rapidly. However, because the rapidly changing background does not exhibit any periodic behavior, incorporating the estimated period in the segmentation process improves the results as shown in Figure 6.6.

In Figure 7, we show the average alignment error for a moving car. The alignment error does not satisfy the significance criterion of Equation 10. Therefore, the tracked object is classified as a non-human object because no significant periodic motion was detected.

Figure 9 is another example of an image sequence acquired from a moving camera. Again, the camera motion is a combination of a left rotation and a forward motion. Along with the moving human in the sequence, there are a few moving vehicles that complicates the scene background. The global minimum, at $p = 14$ of the average alignment error, as shown in Figure 8 coincides with the true motion period of 14 frames/cycle.

The proposed approach was applied to five human sequences and five car sequences. It correctly estimated the motion period in the five human sequences and rejected the null hypothesis in the five car sequences.

VI. Conclusions and Future Work

In this paper, a new approach for human motion identification and segmentation was presented. The video streams of the moving objects are acquired from a moving vehicle. Motion identification is based on computing the probability that a moving pixel belongs to the general motion of the scene or the translational, periodic motion of the object. All possible pairs of object images, given a specific period, are aligned and the average alignment error is computed for the given period. The process is repeated for a range of potential periods. An estimate of the true period is the one that minimizes the average alignment error if and only if it satisfies a significance test. Otherwise, the object is considered non-human, such as a vehicle or a false alarm.

If the tracked object is a human, a motion segmentation procedure is invoked. Instead of using only 2D spatial information to segment the moving object from the background, the estimated motion models and the estimated period are also utilized to improve the motion segmentation obtained by only using spatial segmentation. The motion segmentation is important for verifying the identity of the moving object using other techniques such as shape-based methods, improving tracker performance and activity recognition applications. Experimental results on a set of videos prove the accuracy of the period estimation method and the improved segmentation of the moving object.

The algorithm described in this paper uses pair-wise alignment errors, which might be prone to alignment outliers. We plan to extend the proposed approach to use global alignment errors to improve the accuracy of the estimated period. It also

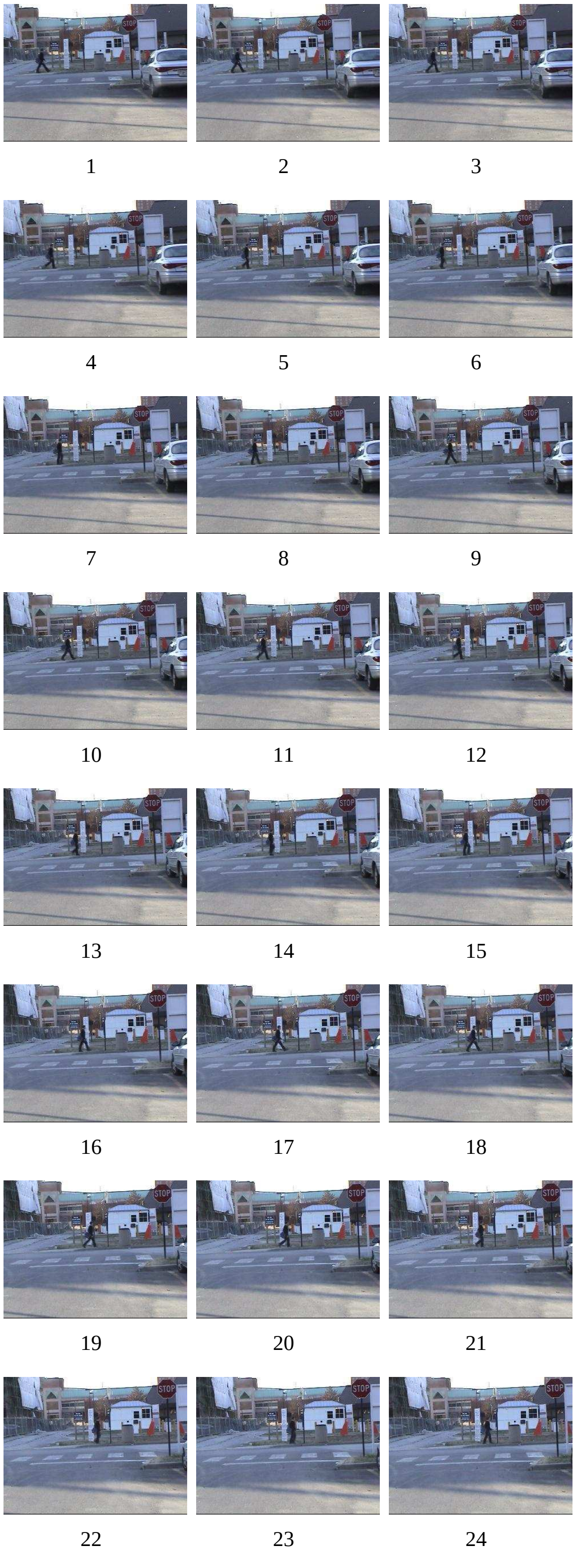


Fig. 3. Odd Frames of The First Test Sequence.

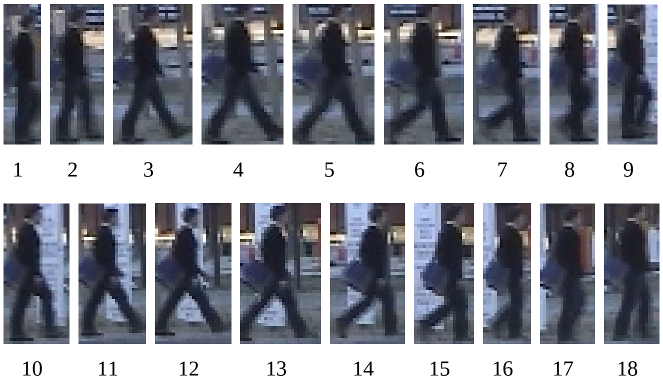


Fig. 4. A Close-up of the Tracked Object in the Odd Frames of the First Test Sequence.

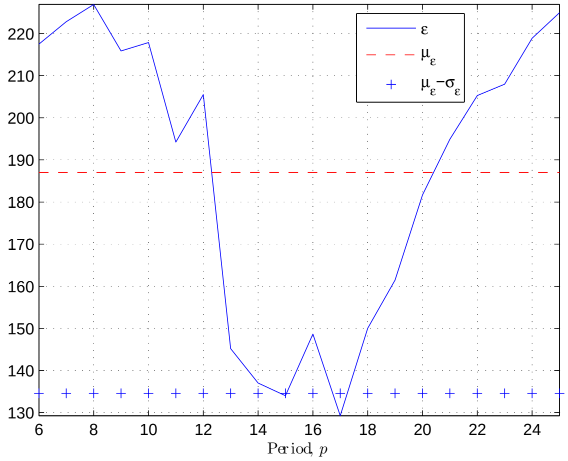


Fig. 5. Average Alignment Error for the Object in Figure 4. The Global Minimum of the Error is at $p = 17$, which Corresponds to the Correct Motion Period.

is based on computing a simple cost function that can be extended to more sophisticated probabilistic models in order to improve the accuracy of the estimated period.

Acknowledgment

This work was supported, in part, by grants from the Defense Advanced Research Projects Agency (DARPA) and the Army Research Laboratory (ARL).

References

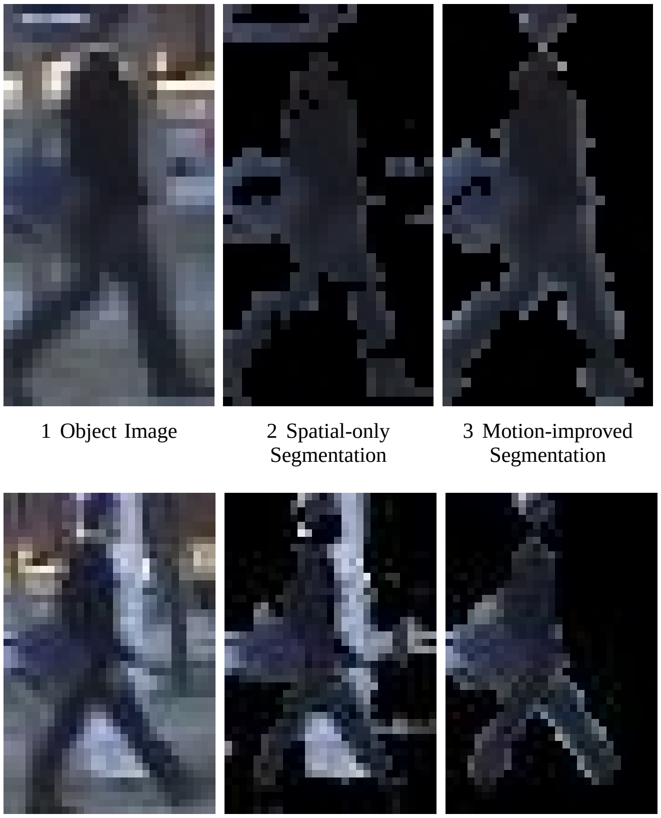
[1] D. Gavrila, "Sensor-based pedstrian detection," *IEEE Intelligent Systems*, vol. 16, no. 6, 2001.

[2] R. Cutler and L. Davis, "Robust Real-Time Periodic Motion Detection, Analysis and Applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, August 2000.

[3] Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 814–827, 7 2003.

[4] H. Nanda and L. Davis, "Probabilistic Template Based Pedstrian Detection in Infrared Videos," in *Proc. IEEE Intelligent Vehicles Symposium*, vol. 1, June 2002, pp. 15–20.

[5] R. T. Collins, R. Gross, and J. Shi, "Silhouette-Based Human Identification from Body Shape and Gait," in *Proc. IEEE Conference on Automatic Face and Gesture Recognition*, 2002.



1 Object Image

2 Spatial-only Segmentation

3 Motion-improved Segmentation

4 Object Image

5 Spatial-only Segmentation

6 Motion-improved Segmentation

Fig. 6. Segmentation Results of Spatial-only Segmentation of [21] and the Motion-improved Segmentation.

[6] D. M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[7] Y. Ran, I. Weiss, Q. Zheng, and L. Davis, "An Efficient and Robust Human Classification Algorithm using Finite Frequencies Probing," in *Joint IEEE Internationl Workshop on Object Tracking and Classification Beyond the Visible Spectrum with CVPR 2004*, 2004.

[8] Y. Liu and R. T. Collins, "Gait Sequence Analysis using Frieze Patterns," in *Proc. European Conference on Computer Vision*, 2002.

[9] J. Little and J. Boyd, "Recognizing people by their gait: The shape of motion," *VIDERE*, vol. 1, no. 2, 1998.

[10] J. Boyd, "Video phase-locked loops in gait recognition," in *Proc. IEEE International Conference on Computer Vision*, 2001.

[11] J. Boyd and J. Little, "Motion from transient oscillations," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.

[12] M. N. Jamie Shutler and C. Harris, "Global statistical description of temporal features," in *IAPRS*, 2000.

[13] J. Shutler and M. Nixon, "Zernike velocity moments for description and recognition of moving shapes," in *British Machine Vision Conference*, Manchester UK, 2001.

[14] M. R. Teague, "Image analysis via the general theory of moments," *Journal of the Optical Society of America*, vol. 70, no. 8, 1979.

[15] S. Niyogi and E. Adelson, "Analysing and recognizing waling figures in xyt," in *Proceedings ofInternational Conferenece on Computer Vision and Pattern Recognition*, 1994.

[16] S. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Trans. on Image Processing*, vol. 11, pp. 1434–1456, November 2004.

[17] B. S. Reddy and B. Chatterji, "An fft-based technique for translation, rotation and scale-invariant image registration," *IEEE Transactions on Image Processing*, vol. 5, no. 8, 1996.

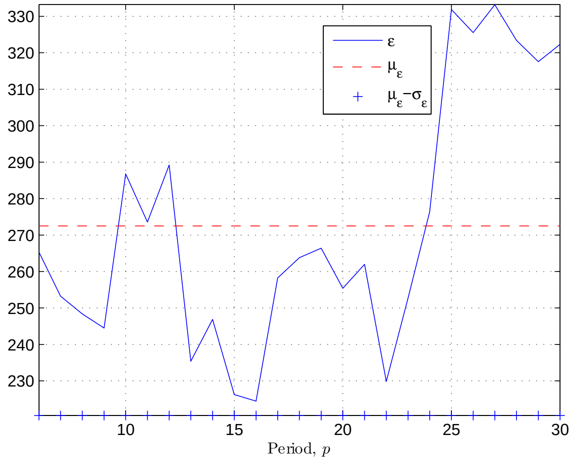


Fig. 7. Average Alignment Error for a Moving Car. The Global Minimum Does not Satisfy the Significance Criterion.

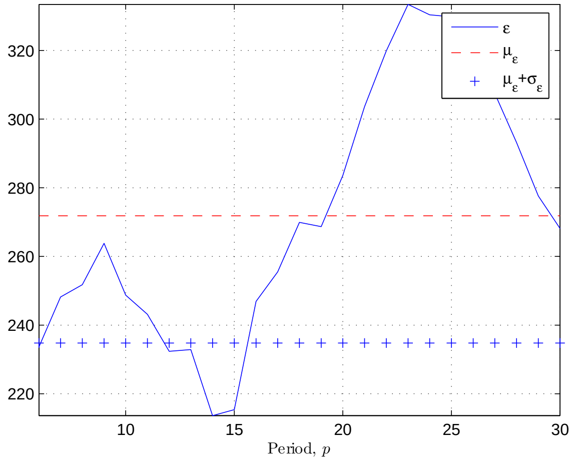


Fig. 8. Average Alignment Error for the Object in Figure 9. The Global Minimum of the Error is at $p = 14$, which Corresponds to the Correct Motion Period.

[18] J. P. Lewis, “Fast template matching,” in *Vision Interface*, 1995.

[19] A. Amir and M. Lindenbaum, “Ground from figure discrimination,” *Computer Vision and Image Understanding*, vol. 1, no. 76, pp. 7–18, 1999.

[20] L. Herault and R. Horaud, “Figure-ground discrimination: A combinatorial optimizatio approach,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 899–914, 1993.

[21] L. Zhao and L. Davis, “Iterative figure-ground discrimination,” in *Proceedings of the International Conference on Pattern Recognition*, 2004.

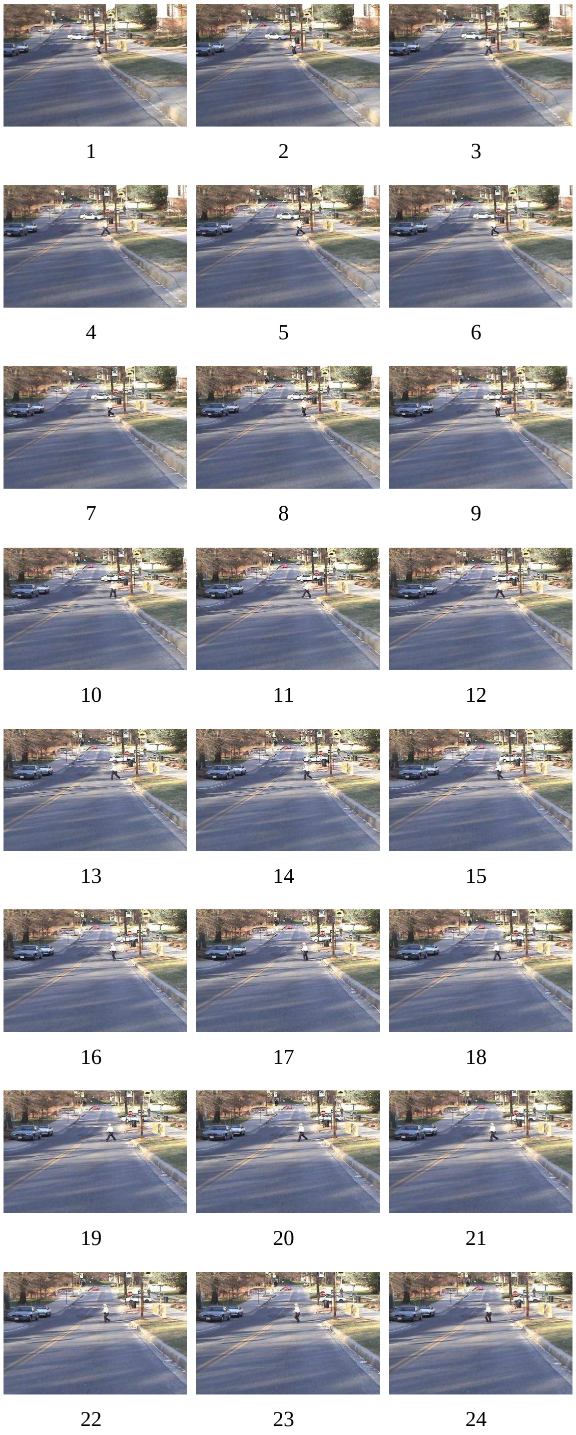


Fig. 9. Odd Frames of the Second Test Sequence.