

AN EFFICIENT AND ROBUST HUMAN/VECHICLE CLASSIFICATION ALGORITHM USING FINITE FREQUENCIES PROBING

ABSTRACT

This paper describes a periodicity analysis based object classification algorithm for infrared videos. Given a detected and tracked object, the goal is to analyze the periodical signature of its motion pattern. Two major object categories are of interest: human and vehicles. Instead of using traditional silhouette period detection method, we propose an efficient and robust solution which is related to the frequency estimation in speech recognition. Periodic reference functions are correlated with the video signal and efficiently tested against the underline hypotheses. Probing over a finite set of periods exhibited by typical human motion enables us to avoid calculation the responses over all frequencies in Fourier domain. , This approach transforms the object classification into a global maximization problem. Experimental results for both infrared and visible videos acquired by ground as well as airborne moving sensors are presented. Detailed performance and robustness analysis regarding to several factors are also presented.

1. INTRODUCTION

1.1 Motivation

Automatic human activity recognition from a video sequence has recently attracted the attention of many researchers [1, 2, 3]. It plays a critical role in many surveillance systems that aim to know what the objects are, and what are they doing [6]. The goal of this work is to classify human and non-human based on their motion pattern difference. To be more specific, we want to exploit motion signatures intrinsic to humans and vehicles and design appropriate classifiers. The periodic nature of human walking has been widely used in gait recognition and related applications [3, 9].

1.2 Related Work

There are many methods in the literature for moving object classification. Among them, motion signature analysis is a simple and promising approach, especially for infrared and airborne video processing, which typically has low image contrast and small object size. Periodical motion signatures are a robust clue in these situations.

Frequency estimation in noise contaminated signal is a well known problem in signal processing [7]. It is well-studied for speech recognition with Gaussian noise assumption. The optimal maximum likelihood estimator (MLE) are obtained by location of the peak in the periodogram. This estimator achieves the Cramer-Rao lower bound for high SNR [8]. However, the computational cost is expensive even with FFT. Besides, there exists a bias when the signal is not ideal sinusoid.

The corresponding work in periodical human motion analysis from 2D video is still an open problem. Several solutions have been proposed for measuring the periodicity of human motion. Allmen and Dyer [11] presented a 3-D based detection in curvature space. Polana and Nelson [12] gave an approach using Discrete Fourier Transform (DFT). Efros and Berg [13] illustrated the cyclic motion from optical flow domain.

A method closely related to this paper can be found in [1]. The authors used pixel-level correlation to calculate the similarity matrix. Every pixel in the matrix represents the temporal similarity between any two images of the same object. The periodical property appears as darker lines parallel to the diagonal line (e.g. in Figure 4) and is detected by a Short Time Frequency Analysis [1]. Another method commonly used is the skeleton from segmented silhouette. For instance, Fujiyoshi and Lipton [2] provide a real-time method by image skeletonization. It uses a 'star' model extracted from a detected silhouette mask to describe the targets' contour distribution. Its evolution over time reveals the underline human body motion. Once the skeleton is extracted, motion cues can be determined for the torso and legs.

There are a couple of drawbacks in above approaches. The similarity matrix based approach [1] requires calculation of a similarity matrix between any two images, which is computationally expensive. Another problem is that the matrix is easily be polluted by mis-alignment of frames as well as changing background over time. It causes the method to be unstable for period detection. The skeletonization method relies on human contour extraction and is sensitive to the quality of silhouette

generation. This process is far from practical use. Silhouette detection is a challenge task especially when video is in low contrast, when the object is small (for example 10*10 pixels as in the applications addressed here), and when the camera is moving.

In this work, the goal is to develop a computationally efficient and robust periodicity motion analysis tool which works well on infrared and airborne videos. In such situations, it is very hard to get a precise mask of the moving objects. For a periodical signal, it will have peaks on the multiples of the period in its power spectral density (PSD). Based on this observation, we hope to use *a priori* signals with specific periods to compare with the original wave and if the two are similar, the cross-correlation will approach to its global maximum. From observation of the pattern of human motion, we extract the correlation characteristic over time. It works as an *a priori*. Furthermore, it also enables us to present a simple algorithm for classification based on hypothesis testing.

1.3 Assumptions

In this paper, we assume the moving objects have been detected and tracked. They are specified by bounding boxes in each frame. We also assume the alignment has been performed, i.e., objects at different time are scaled and aligned so that they have the same or almost the same size and center location. We make *no* constraints on the background structure. Also notice that we do not assume the availability of silhouette.

1.4 Major Contributions

The main contribution of this paper lies in the novel idea of using a finite frequency set to probe the objects for their periodical signature and use the period and its magnitude to classify objects. A concise signal is derived from the periodical and symmetrical nature of human motion as *a priori* reference. Also, this method is *efficient* due to low computation cost. The period detection is transformed into a global-maximum location process. The *robustness* comes from the good discriminating ability. It does not depend on the silhouette of the objects. It works well for low contrast and small size targets where other methods have difficulties.

1.5 Paper Organization

The paper is organized as follows: Section 2 describes the finite frequencies probing method. Experimental results are provided in Section 3. Section 4 gives the sensitivity analysis. Discussion and summary can be found in Section 5.

2. PERIODICITY ANALYSIS BY FINITE FREQUENCIES PROBING

2.1 Single Periodical Signal Probing

As we stated in the introduction, the idea of using a single signal as a reference to correlate with the original target signal wave comes from the Fourier domain analysis. A simple example gives an illustration. In Fig. 1 (A), a target sample signal consisting of the superposition of two cosines and its power spectrum (DFT coefficients) magnitudes are shown. There are two peaks in the middle image of Row A. Taking the origin point into account, the two distances between them are both 260HZ, which is the period of that signal. This phenomenon could be explained by the principles of Fourier Transform [7].

In the ideal case, the power spectrum of a periodical signal show peaks at the multiples of that period. In the practical case we only have a finite length of signal, which is a bunch of bounding boxes containing target objects. And because of alignment error, noise and changing background within boxes, the signal will be polluted and the period in frequency domain will also be harder to detect as shown in Fig 1 (B). The repeated peaks quickly attenuate.

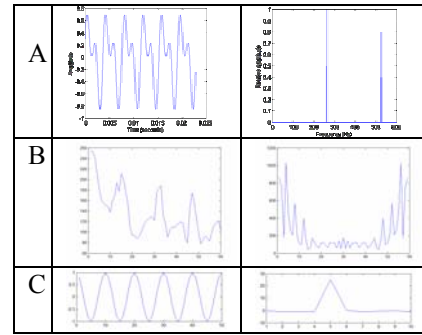


Figure 1. Periodical signals and power spectrums

Notice the fact that in many methods, the underlying thought is that the signal will have particular response at the period or its multiples in frequency domain. Since those algorithms aim at finding the response at given frequency, why cannot we only focus on the response at a finite set of frequencies, which are more likely to be the period of human motion pattern? Before the detailed description, let us first explore a simple example. Let $W(t)$ be the signal to be detected $W(t) = \cos(\omega t)$. Its discrete form is $W(n) = \cos(\omega * n)$. We use another cosine signal $W'(t)$ with a different frequency ω' to correlate with $W(n)$ as a reference signal. The cross correlation is:

$$C(W, W') = \sum_{n=1}^N \cos(\omega n) \cos(\omega' n) \quad (1)$$

where N is the signal length. We give the correlation (or response) at different ω in the column C of Figure 1. From the images it is clear that the response will have a *global* maximum right at ω , even if we only use a signal

with a length of finite periods. We now give the definition for *probing* as:

Definition 1: *Probing is a process of using a reference periodical signal to correlate with the target signal to obtain a measure of similarity in their waveforms.*

Theoretically, starting with a signal $W(t)$ that is assumed to be periodic, or more practically, quasi-periodical, it satisfies: $W(t) \approx W(t+nT)$, where T is the quasi-period of the signal. If we use a window in time to truncate the sample of $W(t)$, we will get a vector $\bar{W}(t) = [W(t), W(t+\tau), \dots, W(t+(N-1)\tau)]$. Given a reference signal W' and additive Gaussian noise assumption, the location of T will be transformed into a maximization of *a posteriori* probability.

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

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

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

The best estimate is T that satisfies (2). Bayesian rule will have

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

Assuming an equal a priori distribution that the signal have period T , the maximization is transformed in (2) and (2') with T . Furthermore, the best estimate of the quasi-period is the frequency which maximizes the cross correlation of $W(t)$ and $W'(t)$. This explains the theoretical background of the intuitive probing method.

2.2 Probing with Finite Frequencies

The real case is more complicated. Two major concerns should be cleared here. One is the availability of the mask used to separate the object from background. The infrared video has very low contrast. The sensors capture heat spectrums beyond visible light, which lowers the color depth of the image. In some case the camera is moving, so we cannot use the background subtraction method [1] in its present form. For infrared data, it has low contrast as the left image in the first row in Figure 3 shows. The second row is a typical airborne surveillance video we are working on. The size of the object of interest is around 10×10 pixels, which cast difficulty in the object mask generation. We will only work on bounding boxes that contain the objects. The periodical motion pattern property applies to only part of the object. So those pixels within the box but not belonging to the object contribute noise and instability to this algorithm. The other concern lies in the computation cost. Our system will require a fast performance of the proposed algorithm. But the correlation matrix based method requires correlation

between any two images, which is much more time consuming.

The output of the detecting/tracking module gives a sequence of bounding boxes for every object. After alignment, we will have a stack of cropped images with same size and centroid, each containing the target object at a different time. Based on such information, a probing function with period ω and wave form k is defined as $\Phi_k(\omega) = k(\omega t)$. The overall cost function is defined over all definition space of $W(t,x,y)$, where $W(t,x,y)$ is the pixel or a corresponding feature at location (x,y) in time t .

$$C(k, \omega) = \int_x \int_y \int_{-\infty}^{\infty} \Phi_k(\omega) \cdot W(t, x, y) dt dy dx \quad (4)$$

In the real case, we will have a limited length and size and the discrete version is:

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

The physical meaning of it is to calculate the overall response of the signal W with the reference signal Φ at a given frequency ω within the bounding box. The period is defined from (3) and (4') as the ω in $\{\omega_1, \omega_2, \dots, \omega_m\}$, which maximizes the cost function.

$$\begin{aligned} \text{period} &= \arg \max_{H_T|W} P_{H_T|W}(H_T | W(t)) = \arg \max_{\omega \in \{\omega_1, \omega_2, \dots, \omega_m\}} C(k, \omega) \\ &= \arg \max_{\omega \in \{\omega_1, \omega_2, \dots, \omega_m\}} \sum_{x=1}^X \sum_{y=1}^Y \text{cor}(\Phi_k(\omega), W(n, x, y)) \end{aligned} \quad (5)$$

2.3 Periodicity Detection

To detect the period more efficiently, we need to select the appropriate wave form of probing function k . Given a signal, the ideal probing function is the signal itself, because the correlation will approach 1, which is the maximum the function can achieve. But as a matter of fact, we are unable to know exactly what the signal looks like *a priori*. Several possible functions are tested and compared. By intuition, the triangle and cosine/sine functions appear to be appealing due to their simple and typical form. And observation of human walking will give a more suitable reference signal.

As research shows, a walking pedestrian will have both period and symmetry due to the waving of the legs and arms. This can be used to design the correct reference signal. Observation in the previous results [1][5] gives us a chance to use the twin peaks resulting from both the period and symmetry. Figure 2 illustrates the nature of these two properties. It is a complete cycle of a walking human. In addition to the similarity between cycles, there is also resemblance between the first and the second halves, shown as two rows in Figure 2.

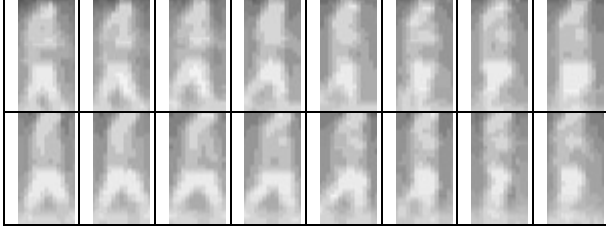


Figure 2. Illustration of period and symmetry of walking

We investigate the similarity between pedestrian sequences in left two images of Figure 3. With simple smoothing, we will notice there are two peaks in every cycle due to period and symmetry mentioned above.

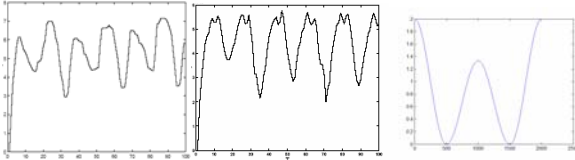


Figure 3. Similarity signals and twin-peak signal

Hence we use a twin-peak reference signal for probing. It is generated by combining two sinusoids as shown in the right image in Figure 3. The first peak, due to period, appears at the multiple of period T and the second, due to symmetry, appears at $(n+1/2)T$. We can also use a single sinusoid or a triangle wave for probing as mentioned before.

The probing results for pedestrians are shown in following figure 4. We first show two samples. One is a ground-based infrared video and the other is airborne data. The second row is the correlation matrix calculated as in [1]. It is easy to see that, although there are darker lines parallel to the diagonal line in the airborne data corresponding to the period, we cannot find any in the correlation matrix for a walking human in infrared video. Yet the probing method detects distinct peak at the periodical frequency as shown in 3rd row in Figure 4.

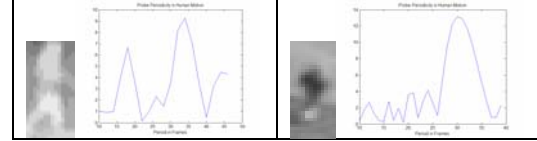
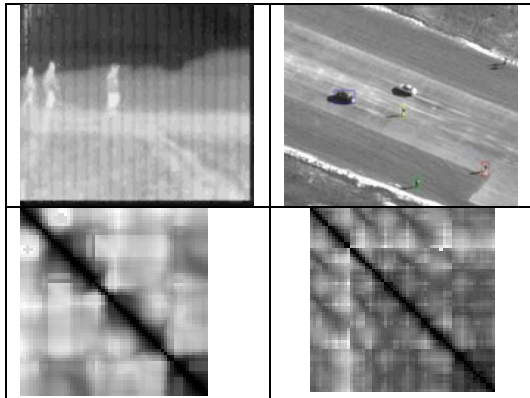


Figure 4 Period detection results

In order to have a better idea how well this method works, we plot the intensity change in figure 5 for a column pixels (a red line in the right image) from the human in figure 4. Although the period only exists in several pixels and is hard to perceive even by human eyes, the proposed algorithm successfully detects it without any difficulty.

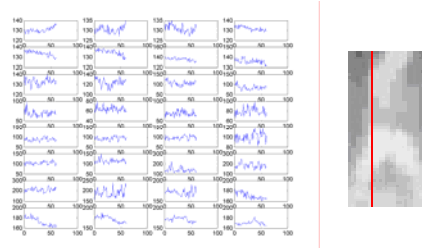


Figure 5. Intensity evolution for pixels in a column

2.4 Object Classification

Notice that the period detection now becomes a simple process as finding global maximum, which could be implemented with many fast and reliable methods. In this work, our aim is to classify two types of objects: humans and vehicles by testing period existence in their motion. It is straightforward to think of a peak-related criterion for decision since the essential pattern is specific for human motion. The decision is made as follows: a candidate object is classified as a human if and only if at some time t , the global peak satisfies

$$\frac{Peak - Mean}{Variance} > TH \quad (7)$$

TH is the threshold for a confident decision. In our experiments, typical values are 2 to 4.

3. EXPERIMENTAL RESULTS

This section describes the classification result of the proposed algorithm. It includes details on the structure of the data as well as the results on testing data sets from both infrared and gray level sensors.

3.1 Structure of Testing Data

The datasets have two major categories. One is from infrared sensors in ground-based platforms. This set consists of 10 sequences. It consists of more than 20 clips (15 pedestrians and 5 vehicles). The foreground objects include typical scenes ranging from a parking lot, a road to other urban scenarios. Notice that there are different objects with various sizes and poses. The other dataset is from gray level airborne video. There are 10 human

sequences and 5 vehicle sequences. All data are captured at a speed of 30 frames per second. We have *no* constraint on the background and all the video are from real outdoor activities. For both sequences, the detection and tracking algorithm that we use is provided by the work in [1,6,10]. The outputs of their methods are the objects within bounding boxes. This gives us great convenience for the probing experiments.

3.2 Experiments on Infrared Test Set

Here we test our algorithm on the infrared data set. As we state in the above sections, we are interested in classifying two objects types: pedestrian and vehicles. In figure 6, we present the probing process at a given time. The video length is 60 frames, which corresponds to around two cycles of a normal walking pedestrian.

In the above figure, the result is shown for detecting the periodical motion for different targets. The blue line is the plotted response and the red line is the mean of the response over all frequencies. These are humans walking through an urban scene at different poses. All of them have a dominant twin-peak which shows that our algorithm has good detection ability for pedestrians. Using the criterion in Eq (7), they could be robustly classified.

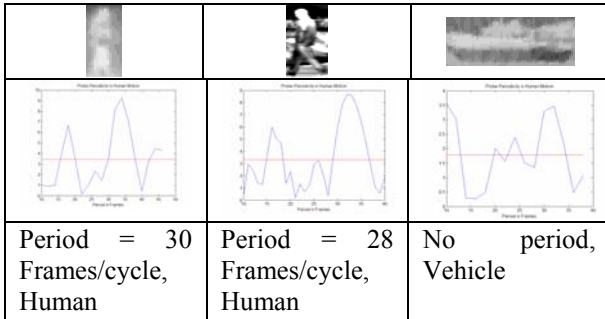


Figure 6. Probing for different objects in infrared data

For a more reliable classification purpose, we may depend on the consistency of the periodicity property over time. In Figure 8, a continuous human period detection result is given during a long time interval. The left image is the superimposed response for 5 different times in a 100-frame pedestrian walking sequence. Not only is the twin-peak pattern distinct, but also the period consistently falls into a narrow range around 25 frames/cycle, which is shown in the right image. By checking the consistency, the object is classified with high confidence as a pedestrian.

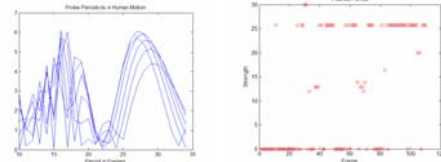


Figure 7. Continuous period detection from infrared data. Left) overlapped response for different time; Right) plot of detected periods along different time.

3.3 Experiments on Gray Level Airborne Data

The same experiment is performed over the gray level airborne data. In the airborne sequence, our targets are the moving humans and vehicles on the ground. Since the camera is far away from the objects, the object size falls into a very small bounding box of 10*10 pixels. Besides, because of the view angle of the camera, those objects viewing from top cast more challenge to the probing. As the result shows in Figure 8 and 9, the proposed method manages to classify the targets as human or vehicle.

Figure 8 gives classification results for two human and a car. Similarly, we also want to know the consistency in infrared data. In figure 9, the continuous probing result is drawn for a human and vehicle. Unlike the case for human, those detected periods of a vehicle are randomly distributed in a wide range. This temporal feature for consistence could be used as a stricter rule for classification.

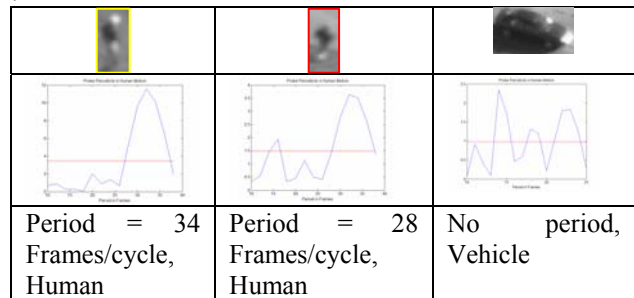


Figure 8 Results for airborne surveillance video

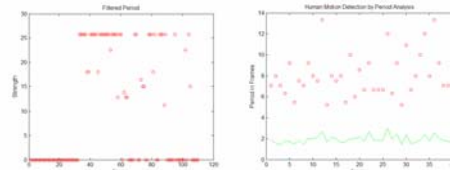


Figure 9 Continuous period detection for human (left) and car(right) in airborne data.

4. SENSITIVITY ANALYSIS

In order to get a complete evaluation of the proposed method, we test the algorithm on several key factors which may influence the result. During the evaluation, a

significant factor is what measurement we use to compare the results. Two questions are of interest here:

- Will it still detect the twin pattern when we change the key experimental factors;
- To what extent will the result be affected in terms of peak-mean-ratio.

Because of the simplicity of the method, we are able to use a variable C defined as the change of the peak-mean-ratio in percentage when we change the conditions.

$$C = \text{abs}(P/M - P'/M') / (P/M) \cdot 100\% \quad (8)$$

4.1 Computational Cost

The first thing we are concerned with is the speed of this method. Unlike other reported methods, we are aiming at a fast or even real-time implementation. A quantitative table is given in table 1. Suppose the bounding boxes after alignment have a width w and a height h , and the length we use is N frames. The probing frequency ranges over n frequencies. Then the required addition and multiplication operations are both $N \cdot w \cdot h \cdot n$. The time is given by: $T = N \cdot w \cdot h \cdot (ADD + MUL) \cdot n$

which is a linear computation time in terms of N or n . Further more, to fasten the speed, we can use multi-resolution probing. A coarse frequencies set is firstly used to roughly locate the global maximum. Then a more dense set is generated around the detected frequency from the 1st step and refines the detection result.

4.2 Alignment

The definition of cost function requires good alignment of the frames for the detected objects. Otherwise the period will be polluted since the correlation in equation (5) is along the temporal axis. Current detecting and tracking algorithms cannot provide error-free alignment for bounding boxes. We need to analyze how well this method performs under different aligning conditions. To get a quantitative comparison, we add a Gaussian noise to a set of calibrated bounding boxes. By increasing the variance, we measure the peak/mean ratio change compared to the original result and show the comparison in the following table as C is defined as in the beginning of this section.

Table 1. Comparison between different alignment noises

σ	0.5	1.0	1.5	2.0	2.5	3.0
Period	34	34	33	34	32	36
C(%)	95.6	78.9	50.2	37.1	12.2	5.8

4.3 Object Size

A robust algorithm should have good detection ability even when the object size is small. This ability will also speed up the system since we can down-sample the object while keeping the correct classification. We give the measurement for one sequence with different down-

sampled sizes. This method shows a strong invariance to the object size. Even when we reduce the size to around 10×10 , we still have good results. Notice that during the sub-sampling, the detected period does not change. We only give the change in peak-mean-ratio in Table 2.

Table 2 Comparison in various sizes, original 100×80

Sub-sample ratio	2	4	6	8
C(%)	93.2	87.9	76.3	70.1

4.4 Video Length

An interesting issue is the minimal length at which we need to analyze the period to obtain sufficient confidence. This is equivalent to considering the size of window we use to truncate the signal. Suppose we estimate the frequency directly from FFT result without any further processing [8]. If the true period is ω , and it falls into two adjacent bins: k and $k+1$,

$$\omega \in [k * 2\pi F_{\text{sample}} / N, (k+1) * 2\pi F_{\text{sample}} / N] \quad (9)$$

where F_{sample} is the sample frequency, we will have a bias up to the width of the bin. Hence it requires longer sequence for higher resolution. But this is determined by the tracking algorithm, which is not always easy to achieve in low quality video for small size objects from moving platforms. In the proposed method, for a normal human, we only need around two to three cycles (60-90 frames for a 30 fps video) to probe out the right period. This goal is much easier to achieve by a common tracker.

Besides, the cross-correlation of the two signals will attain maximum when the windows (length) is a multiple of the period. Hence there would be residue in Eq (5) if this is not the case. But this error can be suppressed by summation over all pixels, which may raise more topics in future work.

4.5 Frame Rate

Due to the limitation of the sensors, we may be unable to get the full frame rate all the time. Also for similar reason as mentioned in 4.2, down-sampling the frame rate while still providing a good result could speed up the process.

Table 3 Comparison between different sizes

frame rate	20	15	10	5
period	17(34)	16(32)	11(33)	5(25)
C(%)	95.0	87.9	46.3	20.1

In the second row, the number in the bracket is the normalized period in the original frame rate. The experimental table proves the fact that the probing is more sensitive to the down sampling the object size than the frame rate. This could be explained by the periodical motion perception ability our Vision System. The HVS

maintains a powerful ability to integrate the temporal information, or motion.

5. SUMMARY AND DISCUSSION

The novelty in the presented technique lies in its simplicity, the efficiency of the implementation, and the robustness to many factors such as target size and frame rate. It transforms the complicated period detection into an easy global maximum location process. The choice of the probing by *a priori* reference signal within a finite frequency set enables accurate object classification even with short video clip (2-3 seconds) in infrared video as well as data from other sensors. Detailed sensitivity analysis reveals the robust nature of this proposed method. Future research could be done in fusing more complicated motion analysis tools. Integration over time can also be considered to extend the application.

6. REFERENCES

- [1] Ross Cutler, Larry Davis. Robust real-time periodic motion detection, analysis, and applications, Pattern Analysis and Machine Intelligence, IEEE Transactions on, Volume: 22 , Issue: 8 , Aug. 2000 Pages:781 - 796
- [2] Hironobu Fujiyoshi, Alan J. Lipton, Real-Time Human Motion Analysis By Image Skeletonization, Applications of Computer Vision, 1998. WACV '98. Proceedings. Pages:15 - 21
- [3] Hogg, D., Model-based vision: A program to see a walking person, Image and Vision Computing, 1983 vol. 1, Pages. 5-20
- [4] Orriols, X., Binefa, X., Classifying periodic motions in video sequences, Orriols, X.; Binefa, X. Image Processing, 2003. Proceedings. 2003 International Conference on , Volume: 1 , 2003 Pages: 945 - 948
- [5] R. Cutler and L. Davis. View-based detection and analysis of periodic motion. Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on , Volume: 1 , 1998 Pages: 495 - 500
- [6] Ismail Haritaoglu, Davis Harwood, Larry Davis, W4: Real-Time Surveillance of People and Their Activities, Pattern Analysis and Machine Intelligence, IEEE Transactions on , Volume: 22 , Issue: 8 , Aug. 2000 Pages:809 - 830
- [7] David C. Rife and Robert R. Boorstyn, Single-Tone Parameter Estimation from Discrete-Time Observations, Information Theory, IEEE Transactions on , Volume: 20 , Issue: 5 , Sep 1974 Pages: 591 - 598
- [8] Steven Kay, A Fast and Accurate Single Frequency Estimator, Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on , Volume: 37 , Issue: 12 , 1989 Pages:1987 - 1990
- [9] Li, B., Holstein, H., Recognition of human periodic motion-a frequency domain approach, Pattern Recognition, 2002. Proceedings. 16th International Conference on, Volume: 1 , Pages:311 – 314
- [10] Jie Shao, S. Zhou, R. Chellappa, Simultaneous background and foreground modeling for tracking in surveillance video, submitted to ICIP, 2004.
- [11] Steven M. Seitz, Charles R. Dyer, View-Invariant Analysis of Cyclic Motion , Intl Journal of Comp. Vis. 25 Pages :1-23 1997
- [12] Ramprasad Polana and Randal C. Nelson, Recognition of Motion from Temporal Texture, Proc. IEEE Conference on Computer Vision and Pattern Recognition, Champaign, Illinois, June 1992, Pages: 129-134
- [13] A. A. Efros, A. C. Berg, G. Mori, J. Malik, Recognizing Action at a Distance, Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on , Oct. 13-16, 2003 Pages:726 - 733