Extraction and Analysis of Multiple Periodic Motions in Video Sequences

Alexia Briassouli, Narendra Ahuja
Beckman Institute
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{ briassou, ahuja }@vision.ai.uiuc.edu

Abstract

The analysis of periodic or repetitive motions is useful in many applications, such as the recognition and classification of human and animal activities. Existing methods for the analysis of periodic motions first extract motion trajectories using spatial information, and then determine if they are periodic. These approaches are mostly based on feature matching or spatial correlation, which are often infeasible, unreliable, or computationally demanding. In this paper, we present a new approach, based on the time-frequency analysis of the video sequence as a whole. Multiple periodic trajectories are extracted and their periods are estimated simultaneously. The objects that are moving in a periodic manner are extracted using the spatial domain information. Experiments with synthetic and real sequences display the capabilities of this approach.

Index Terms

Periodic Motion Analysis, Time-Frequency Distributions, Short Term Fourier Transform

I. Introduction

Periodicity characterizes many movements, including human and animal motions. For example, a beating heart, flying birds, running and walking, all exhibit periodicities. Also, the motions

of wheels, windmills, swings, and pendulums all have a periodic nature. The characteristics of such motions play a fundamental role in activity recognition and object tracking [1], [2].

Most past approaches for periodic motion analysis [3], [4] involve first the detection of object locations in successive images, then the identification of their trajectories and, finally, the analysis of these trajectories. Current spatio-temporal methods for general motion analysis may use level-sets, statistical shape priors [5], [6], or compute the optical flow [7] under appropriate smoothness constraints.

In this paper, we present a new approach for the analysis of multiple periodic motions over a static background, motivated by the observation that repetitive patterns have distinct signatures in frequency space. The extraction of these frequency domain signatures leads to the estimation of the periodic motions, which can then be used for motion segmentation in the spatial domain. The proposed method does not make any assumptions on the object shapes or the smoothness of their trajectories, so it can be applied over a wide range of videos. The main parts of the proposed approach are as follows.

1) The pixel values of each video frame are projected onto the x and y axes, giving two 1D signals of the projections over time (Sec. II). The periodic motions in the x and y directions are evident in these projections.
2) The periodic x and y velocities are used to create a frequency-modulated (FM) signal (Sec. IV), which is itself periodic.
3) The time-varying power spectrum of each FM signal is computed to identify the multiple time-varying frequencies present in it (Sec. III-A).
4) The power spectrum maxima are proportional to the periodic trajectories. Thus, the periods of individual objects are extracted from them via standard spectral analysis (Sec. V).
5) Once all the periods occurring in the video are estimated, spatial correlation is used for object segmentation.

A. Previous Work

Numerous methods exist for analyzing repetitive motions, and can be divided into two categories according to the image features used. The first large category consists of methods that are based on point feature correspondences between successive frames, while the second one uses correlations between regions instead of points.

1) Point Correspondence Methods: Point feature correspondences have been used in many applications, e.g. for the extraction of leg and arm trajectories, for gait analysis [8], [9]. In controlled environments, point correspondences can be found by manually placing reflective markers on the moving objects and tracking their positions throughout the video sequence [10], [11]. This leads to a precise trajectory for each moving object, from which its period can be estimated. When it is not possible to obtain point correspondences manually (e.g. in medical images), they need to be extracted from the acquired images. However, feature correspondence methods are susceptible to errors caused by changing illumination, reflectance, or even invisibility due to occlusion, and are also computationally intensive. This makes these methods unreliable and not generally applicable.

2) Region Correspondence Methods: In region correspondence methods, successive frames of a video sequence are correlated in the spatial domain, resulting in the "similarity plot" [4]. When an object moves in a repetitive way, with period T, its periodicity also appears in the "similarity plot". Periodicity is viewed strictly as oscillations around fixed positions, so these methods cannot detect periodicities superposed on other motions, such as translations, as is the case in walking, running. For these cases, pre-processing is required: the sequence needs to be aligned, i.e. the periodically moving objects have to be localized and brought to the same position in each frame [4], [10], before they are correlated to detect the periodicities. However, the localization of the periodically moving objects also requires prior knowledge or pre-processing. Finally, region correlations are sensitive to noise and may give unreliable results for videos of poor quality.

B. Motivation

Besides the intuitive motivation of exploiting the frequency-compatible nature of the problem of periodic motion analysis, the proposed work is aimed at overcoming the aforementioned limitations of current methods [4], [12].

1) Frequency-domain approaches involve spatially global, instead of local, analysis [13], which makes them robust to local occlusions and illumination variations. Also, frequency based motion estimation is reliable near the boundaries of moving objects where local, intensity-based algorithms may not perform satisfactorily [14].
2) There is no need for explicit feature matching as in spatial methods [15]. Thus, problems related to detecting and matching point features or regions, such as sensitivity to varying

reflectance, illumination, temporary occlusions, and poor video quality, are avoided.

3) Frequency domain motion estimates are extracted using the data from the entire video frame, so they are robust to global illumination changes [16], unlike spatial methods [17].
4) Efficient algorithms [18], [19], [20] (and hardware) are available for frequency transform computations. We compute only 1D transforms, of the x-y image projections, which are even more computationally efficient than if 2D transforms were used.

A characteristic difficulty occurring with frequency-domain motion estimation methods, is the "localization" problem [21], [22]. Although the global nature of frequency-domain motion estimation avoids local errors (e.g. at object boundaries), it does not provide information about the spatial location of the moving objects. Thus, further processing is required, using information from the spatial domain, to correctly assign the motion estimates to pixels.

C. Contributions

The work presented in this paper makes the following major contributions:

1) Unlike any of the previous methods, it extracts multiple periodic motions simultaneously (see Step (3) in Sec. I). In the existing literature, each object or feature point trajectory is extracted separately, and then examined to detect periodicities.
2) All frames are processed simultaneously, making the method robust to deviations from strict periodicity, since it captures the repetitions over the entire sequence. This is also proven mathematically for the cases where the period or the velocity magnitude is corrupted by Gaussian noise. Thus, strong, even though imperfect, periodic motions, which would need additional frame-by-frame, local analysis by the spatial methods, are detected.
3) The computational cost is lower than that of the spatial methods for the following reasons:
    a) Frame by frame spatial processing is limited to a small number of full-frame correlations for object segmentation (see Step (5) in Sec. I).
    b) Spatial pre-processing steps, such as segmentation of the periodically moving objects, are not necessary since the proposed approach does not need to align the video frames before detecting periodicities [4], [10].
    c) The frequency domain computations themselves are efficient.
    d) Most of the analysis involves 1D signals, which have inherently lower cost.

4) The proposed approach can serve as an example for formulating joint spatial and frequency based solutions to other problems.

II. Mathematical Formulation

In this section, we present the problem formulation for multiple periodically translating objects in a scene taken by a static camera (i.e. the background remains the same in all frames). The M periodically translating objects have luminance $s_i(\bar{r})$, $1 \leq i \leq M$ at pixel $\bar{r} = (x, y)$, and displacement $\bar{b}_i(k) = (b_i^x(k), b_i^y(k))$ from frame 1 to k. The displacement is a periodic function of time with periods T_i^x and T_i^y in directions x and y. The Fourier Transform (FT) of each object i is $S_i(\bar{\omega}) = Z_i(\bar{\omega})e^{j\Theta_i(\bar{\omega})}$, where $\bar{\omega} = (\omega_x, \omega_y) = [2\pi m/N_1, 2\pi n/N_2]^T$, $m, n \in Z$, is the 2D frequency, $N_1 \times N_2$ the image size, $Z_i(\bar{\omega})$ the FT magnitude, and $\Theta_i(\bar{\omega})$ the FT phase. When referring to video sequences, we use k to denote time, because it corresponds to video frames, so it is discrete.

Each video frame in the spatial domain is represented as the sum of the background, denoted $s_b(x, y)$, and the M moving objects, $s_i(x, y)$, $1 \leq i \leq M$, so frame 1 is:

$$a(x, y, 1) = s_b(x, y) + \sum_{i=1}^{M} s_i(x, y) + v_{noise}(x, y, 1). \quad (1)$$

The number M of moving objects is unknown: it is extracted simultaneously with the multiple periodic trajectories, by simply counting them. Thus, no prior knowledge of the number of moving objects is needed. The additive noise term $v_{noise}(x, y, 1)$ is commonly used in the literature [23] to represent the measurement noise present in frame 1.

In reality, the image formation involves masking, rather than addition of pixel values, in the object areas. A more precise model can be acquired by removing the background from each frame. Numerous methods have been developed for background removal [24], [25]. Methods that model the background as mixtures of Gaussians [26], [27] give very good results, but require a training stage and have a higher computational cost. Simpler methods, such as median filtering, provide acceptable accuracy, at a very low computational cost and limited memory requirements. In the experiments presented in this paper, the background is removed via median filtering, which gives very good results, since all parts of the background are uncovered in different frames during the video. In poor quality sequences, where the video is very jittery and

the illumination changes significantly, background information from "local" frames is used to remove it more effectively. After background removal, the first frame is modeled as

$$a(x,y,1) = \sum_{i=1}^{M} s_i(x,y) + v_{noise}(x,y,1), \tag{2}$$

and frame k, $1 \leq k \leq N$ becomes $a(x,y,k) = \sum_{i=1}^{M} s_i(x - b_i^x(k), y - b_i^y(k)) + v_{noise}(x,y,k)$. Its 2D spatial FT is

$$A(\omega_x,\omega_y,k) = \sum_{i=1}^{M} S_i(\omega_x,\omega_y) e^{-j(\omega_x b_i^x(k) + \omega_y b_i^y(k))} + V_{noise}(\omega_x,\omega_y,k), \tag{3}$$

so the motion parameters $b_i^x(k)$ and $b_i^y(k)$ appear in a sum of weighted exponentials. To identify the periodic motions, we project each frame in the x and y directions respectively:

$$q_x(x,k) = \sum_{y} \sum_{i=1}^{M} [s_i(x,y,k) + v_{noise}(x,y,k)] = \sum_{i=1}^{M} s_i^x(x - b_i^x(k)) + v_{noise}^x(x,k), \tag{4}$$

$$q_y(y,k) = \sum_{x} \sum_{i=1}^{M} [s_i(x,y,k) + v_{noise}(x,y,k)] = \sum_{i=1}^{M} s_i^y(y - b_i^y(k)) + v_{noise}^y(y,k), \tag{5}$$

where

$$s_i^x(x,k) = \sum_{y} s_i(x,y,k), \quad s_i^y(y,k) = \sum_{x} s_i(x,y,k) \tag{6}$$

$$v_{noise}^x(x,k) = \sum_{y} v_{noise}(x,y,k), \quad v_{noise}^y(y,k) = \sum_{x} v_{noise}(x,y,k). \tag{7}$$

The new 1D representations $q_x(x,k)$ and $q_y(y,k)$ of the video sequence contain the M horizontal and vertical projections of the periodically varying objects as a function of time. The x and y motions are often dependent on each other (e.g. in a pendulum's motion), but can still be estimated separately. The fact that they are dependent in a physical sense does not affect the estimation process. Their dependence simply appears in the final result: for example, in the case of a pendulum (Sec. VII-B), the extracted period in the y direction is twice as much as the period in the x direction, as it is expected for such motions. When each frame is projected on the x and y axis (after background removal), it is possible that the projected pixels of one object will hide the projections of the other object, partially, or even completely. However, even this difficulty can be dealt with, by developing other methods that combine the spatial and frequency domain representations of the sequence without using projections. This is beyond the scope of the current paper, but is a topic of future research.

Since the object displacement between successive frames varies with time, the corresponding FT phase changes per frame of the resulting signals $q_x(x,k)$ and $q_y(y,k)$ are non-stationary functions of time k. Specifically, the FT of $q_x(x,k)$ (the same analysis applies to $q_y(y,k)$) is $Q_x(\omega_x,k) = \sum_{i=1}^{M} S_i^x(\omega_x) e^{-j\omega_x b_i^x(k)} + V_{noise}^x(\omega_x,k)$, where $S_i^x(\omega_x)$ is the 1D FT of each $s_i^x(x,k)$. For constant inter-frame displacement b_i^x, there is a constant phase change in $Q_x(\omega_x,k)$ from frame to frame. This phase change corresponds to a line in the ω_x-k space, whose slope is determined by b_i^x, so it can be used to estimate b_i^x [23], [13]. However, time-varying displacements, e.g. in periodic motions, cannot be estimated from the simple FT, since it will have a time-varying spectral content. Thus, the FT of the video needs more analysis to estimate motions with non-constant velocities.

III. Time-Frequency Distributions

The FT of the video sequence with (periodically) time-varying displacements has a time-varying spectrum, i.e. it is a non-stationary signal. Time-frequency distributions (TFDs) capture the variations of the frequency content of the signal with time, essentially by computing the spectra after windowing the 1D signal around the time of interest [28], [29]. Numerous types of time-frequency distributions have been developed and, as noted in the literature [28], a different one is best suited for each application.

One large category of TFDs consists of the linear TFDs, like the Short-Term Fourier Transform (STFT) [30], while the other large category is that of quadratic distributions [31], [32], such as the Wigner Ville Distribution (WVD). Linear TFDs are characterized by the linearity property: the TFD of a linear combination of signals is equal to the linear combination of the signals' TFDs. Quadratic TFDs can be interpreted as the two-dimensional distribution of signal energy over time and frequency. The most common quadratic TFD is the WVD, which windows the data with the data signal itself. The WVD has the significant drawback of having many cross-terms, due to the multiplication of the signal $s(k)$ with time-shifted versions of itself. For multi-component signals, the cross-terms make it impossible to extract the time-varying spectra.

We empirically verified this by applying the WVD for the detection of multiple periodic motions in a few sequences. Indeed, the result displayed so many occluding cross-terms that it was impossible to separate the multiple motion trajectories. On the other hand, for the same sequences, the STFT captured the signals' time-varying spectrum well. In addition, the STFT is

simpler to compute, so we chose to use it in our experiments.

A. Short Term Fourier Transform

The main concept of the STFT is that, to capture the spectral variation with time, one should compute the FT of the signal only locally in time, and not over the entire time axis. This can be achieved by windowing the original signal in time with an appropriate low-pass function. The spectrum of the resulting signal around time k then approximately represents the spectral content of the signal at that time instant. For a 1D signal $s(k)$ in discrete time, the STFT is defined as

$$STFT_s(k, \omega; h) \equiv \sum_{\tau=-\infty}^{+\infty} s(\tau + k) h^*(\tau) e^{-j\omega\tau}, \tag{8}$$

where $h(k)$, sometimes called a lag window function, is usually a lowpass function representing the "analysis window". The window function controls the relative weight imposed on different parts of the signal, thus defining an inherent tradeoff between time and frequency resolutions. If $h(k)$ has higher values near the center of the interval (the observation point k), the STFT estimates quantities that are local in the time domain. Thus, a window that is compact in time leads to higher time resolution, whereas a window that is spread out in time leads to higher frequency resolution [33]. We use a Gaussian $h(k)$, due to its smooth windowing properties [34].

IV. Instantaneous Frequency Estimation

In order to analyze the periodic displacements, we first construct a frequency modulated (FM) signal whose frequency is modulated by the time-varying object displacements between successive frames over the entire sequence [35], [36] (step (2) in Sec. I). In our problem, these displacements (and velocities) are periodic, so the corresponding frequencies are also periodic functions of time. By using TFDs, the time-varying frequencies of the FM signal can be extracted, and consequently the spatial trajectories can be found.

To construct this FM signal, we use the technique called constant μ-propagation [35], which is essentially the FT at frequency μ. This is obvious in the equation that follows, where the

signal $q_x(x,k)$ is multiplied by $e^{j\mu x}$ and the resulting signal is summed over all x:

$$\begin{aligned} d_x(\mu,k) &= \sum_x q_x(x,k)e^{j\mu x} = \sum_x \sum_{i=1}^{M} [s_i^x(x-b_i^x(k)) + v_{noise}^x(x,k)]e^{j\mu x} \\ &= \sum_{i=1}^{M} e^{j\mu b_i^x(k)} S_i^x(\mu) + V_{noise}^x(\mu,k). \end{aligned} \tag{9}$$

$S_i^x(\mu)$ is the value of the FT of object i's ($1 \leq i \leq M$) horizontal projection at frequency μ, and $V_{noise}^x(\mu,k)$ is the value of the FT of the additive noise's horizontal projection, again at frequency μ. Similarly, we get $d_y(\mu,k) = \sum_y q_y(y,k)e^{j\mu y} = \sum_{i=1}^{M} e^{j\mu b_i^y(k)} S_i^y(\mu) + V_{noise}^y(\mu,k)$, where $S_i^y(\mu)$ is the FT (at frequency μ) of object i's ($1 \leq i \leq M$) vertical projection and $V_{noise}^y(\mu,k)$ is the FT of the additive noise's vertical projection (at frequency μ). We shall focus on $d_x(\mu,k)$, since the same analysis applies to $d_y(\mu,k)$.

The sum of Eq. (9) can be written as $d_x(\mu,k) = \sum_{i=1}^{M} d_i^x(\mu,k)$, where $d_i^x(\mu,k)$ corresponds to object i and is given by $d_i^x(\mu,k) = S_i^x(\mu)e^{j(\mu b_i^x(k))} + V_{noise,i}^x(\mu,k)$, where the noise $V_{noise}^x(\mu,k) = \sum_{i=1}^{M} V_{noise,i}^x(\mu,k)$, and $V_{noise,i}^x(\mu,k)$ is such that:

$$S_i^x(\mu)e^{j(\mu b_i^x(k))} + V_{noise,i}^x(\mu,k) = A_i(\mu)e^{j(\mu b_i^x(k) + n_{noise,i}^x(\mu,k))}. \tag{10}$$

Eq. (10) is written this way simply to demonstrate the effect of noise on the motion estimation, so we do not need to actually estimate $A_i(\mu)$ and $n_{noise,i}^x(\mu,k))$. However, we know that they must satisfy Eq. (10), so:

$$S_i^x(\mu)\cos(\mu b_i^x(k)) + \Re[V_{noise,i}^x(\mu,k)] = A_i(\mu)\cos(\mu b_i^x(k) + n_{noise,i}^x(\mu,k)), \tag{11}$$

$$S_i^x(\mu)\sin(\mu b_i^x(k)) + \Im[V_{noise,i}^x(\mu,k)] = A_i(\mu)\sin(\mu b_i^x(k) + n_{noise,i}^x(\mu,k)), \tag{12}$$

where $\Re[V_{noise,i}^x(\mu,k)]$ indicates the real part of $V_{noise,i}^x(\mu,k)$ and $\Im[V_{noise,i}^x(\mu,k)]$ its imaginary component. The object displacements $b_i^x(k)$ appear in the phase of the resulting signal, so they can be found by estimating its time-varying phase $\zeta_i(\mu,k) = \mu b_i^x(k) + n_{noise,i}^x(k)$. The time derivative of a signal's phase gives its frequency, so we have $\omega_i^x(\mu,k) = \frac{\partial \zeta_i(\mu,k)}{\partial k} = \mu\frac{\partial b_i^x(k)}{\partial k} + \frac{\partial n_{noise,i}^x(k)}{\partial k} = \mu u_i^x(k) + \xi_i^x(k)$, where $u_i^x(k)$ is the x-velocity of object i at time instant k and $\xi_i^x(k)$ the time-derivative of the noise $n_{noise,i}^x(k)$. The noise term $\xi_i^x(k)$ adds random variations to the velocities $u_i^x(k)$, which on average do not degrade these estimates, since they do not follow a specific curve (like the periodic motions do). This is also shown in our experiments with real sequences, where the trajectories are extracted reliably, despite the presence of measurement noise, varying

illumination, and a non-steady camera. A noise reduction pre-processing step [37], [38] can also be used to improve the video quality, and thus reduce the influence of $\xi_i^x(k)$.

The TFDs of the FM signal give its time-varying spectrum, which traces curves that follow the motions in it. Since each signal is windowed (Eq. (8)), there is spectral leakage between neighboring frequencies at each time instant. Consequently, the TFDs give a power spectrum that forms "ridges", whose peaks give estimates of the dominant frequencies $f_i(t)$, $1 \leq i \leq M$, caused by the object motions. The number M of moving objects can be simultaneously extracted, by simply counting the number of peaks in the power spectrum at each time instant t.

V. Multiple Periodic Motion Analysis

A. Multiple Period Detection and Estimation

In the previous section we describe how the time-varying frequency content of the signal is extracted. However, this results in M instantaneous frequencies (IF) $f_i(k)$ at every time instant, which need to be separated. This problem has always been a challenge in time-frequency research, where the components may be separated by using prior information, e.g. the assumption that the IFs are continuous functions of time [39]. In this paper, we use the periodicity properties to separate each component of the time-varying power spectrum. At each time instant k (frame k), we have M pairs of velocity values $v_1^x(k), ..., v_M^x(k)$ and $v_1^y(k), ..., v_M^y(k)$, each a periodic function of time. As before, we examine only the horizontal projection of the trajectories, since the same analysis can be applied to the vertical projection. For object i, $1 \leq i \leq M$, and time instants k, $1 \leq k \leq N$, we have the periodic signal $\bar{v}_i^x = [v_i^x(1), ..., v_i^x(N)]$. The sum of the M periodic signals $\bar{v}_i^x$ of all objects i at each instant k forms the function $\bar{g}_x = [g_x(1), ..., g_x(N)] = \sum_{i=1}^{M} \bar{v}_i^x$. Its values at each frame k ($1 \leq k \leq N$) are given by $g_x(k) = \sum_{i=1}^{M} v_i^x(k)$.

Classic spectral analysis of $\bar{g}_x$ via, for example, the MUSIC algorithm [40], can lead to the estimation of the M periods T_i^x ($1 \leq i \leq M$) in it. The basic assumption for MUSIC to be applicable is that the data can be expressed as a linear combination of terms containing its fundamental frequencies. Since our data $\bar{g}_x$ is a sum of the periodic functions $\bar{v}_i^x$, we can apply this algorithm to extract its fundamental frequencies. MUSIC estimates the fundamental frequencies of the data from the eigen-decomposition of its covariance matrix $R_g = E[\bar{g}_x^H \bar{g}_x]$. The dominant singular values of R_g (and consequently its dominant eigenvalues) correspond to the M frequencies present in the signal $\bar{g}_x$. In the presence of noise, there are $N > M$

frequencies present, so we keep the M highest singular values, and eliminate the $N-M$ lower ones, which have been introduced by the noise. Thus, the multiple periods can be estimated simultaneously, and used to separate the M different periodic trajectories.

B. Periodically Moving Object Extraction

Once the different periods are estimated in the sequence, they can be used to extract the moving objects. Consider frame k, represented by:

$$a(x,y,k) = \sum_{i=1}^{M} s_i(x - b_i^x(k), y - b_i^y(k)) + v_{noise}(x,y,k). \quad (13)$$

Since the motions are periodic, for each object, we have $b_i^x(k) = b_i^x(k+T_i^x)$, $b_i^y(k) = b_i^y(k+T_i^y)$. For simplicity, in this section we will consider $T_i^x = T_i^y = T_i$, but the same analysis can be applied when the motion periods in the horizontal and vertical directions are different. If T_n denotes the period of object n at time $k' = k + T_n$, Eq. (13) becomes:

$$\begin{aligned} a(x,y,k') &= \sum_{i=1}^{M} s_i(x - b_i^x(k'), y - b_i^y(k')) + v_{noise}(x,y,k') = \sum_{i \neq n} s_i(x - b_i^x(k'), y - b_i^y(k')) \\ &+ s_n(x - b_n^x(k+T_n), y - b_n^y(k+T_n)) + v_{noise}(x,y,k') \\ &= \sum_{i \neq n} s_i(x - b_i^x(k'), y - b_i^y(k')) + s_n(x - b_n^x(k), y - b_n^y(k)) + v_{noise}(x,y,k'), \quad (14) \end{aligned}$$

since $s_n(x - b_i^x(k), y - b_i^y(k)) = s_n(x - b_i^x(k+T_n), y - b_i^y(k+T_n))$, i.e. object n is in the same position in frames k and $k' = k + T_n$. Therefore, we can extract the n^{th} object by correlating frames k and $k' = k + T_n$: since only that object is expected to re-appear in the same position in those frames, the correlation values will be highest in the pixels in the object area. If more than one objects have the same period and phase, they are treated as a single entity during segmentation. However, if they have the same period, but different phase, this approach ensures that they will be separated.

C. Periodic Motion Superposed on Translation

As described in Sec. I, one of the main contributions of our method is the fact that it allows the estimation of periodic motions superposed on translations, such as a human or animal walking or running. In these cases, the legs are moving periodically, similarly to a pendulum, but the entire body is also translating. Most existing methods for periodic motions cannot deal with such

motions because they use correlations between frames assuming no translation. Our approach can estimate the object periods even when the periodic motion is superposed on translation, because the TFDs extract the time-varying frequency of the FM signal that we create via μ-propagation.

For the case of periodic motion superposed on a translation, the form of the time-varying trajectory is $b(n) = \alpha \cdot n + b_P(n)$, where $1 \leq n \leq N$, α is a constant and $b_P(n)$ is the periodic component of the motion. We create a FM signal via μ-propagation: $z(n) = e^{j\mu(\alpha \cdot n + b_P(n))}$, whose phase is $\mu(\alpha \cdot n + b_P(n))$. Time-frequency analysis estimates its frequency, i.e. its time-derivative:

$$\omega_z^{est}(n) = \frac{\partial(\mu(\alpha \cdot n + b_P(n)))}{\partial n} = \mu\left(\alpha + \frac{\partial b_P(n)}{\partial n}\right). \tag{15}$$

Consequently, the translational component of the motion becomes a simple additive term, whereas the periodicity of $b_P(n)$ is retained in $\frac{\partial b_P(n)}{\partial n}$. This allows us to extract periodicities superposed on translations, without needing to align the video frames. The above can be extended to periodic motion superposed on any kind of time-varying translation, since the periodic component will still be present in $\frac{\partial b_P(n)}{\partial n}$ in that case as well.

D. Object Extraction for Periodic Motion Superposed on Translation

Once the motion periods are estimated, the objects can be extracted using the fact that each periodically moving object will appear the same after an integer number of periods T, i.e. from frame k to frame $k' = k + \lambda \cdot T$ (where $\lambda \in \mathcal{Z}$). Since its center of gravity will also have translated, we first need to estimate the "mean" translation between these two frames [23], [13]. Let frame 1 of the sequence be given by Eq. (2) and frame k by Eq. (3). The ratio of the FTs $A(\omega_x, \omega_y, k)$ and $A(\omega_x, \omega_y, 1)$ is given by

$$\Phi_k(\omega_x, \omega_y) = \frac{A(\omega_x, \omega_y, k)}{A(\omega_x, \omega_y, 1)} = \sum_{i=1}^{M} \gamma_i(\omega_x, \omega_y) e^{-j(\omega_x b_i^x(k) + \omega_y b_i^y(k))} + \gamma_n(\omega_x, \omega_y, k), \tag{16}$$

where $\gamma_i(\omega_x, \omega_y) = \frac{S_i(\omega_x, \omega_y)}{A(\omega_x, \omega_y, 1)}$, and $\gamma_n(\omega_x, \omega_y, k) = \frac{V_{noise}(\omega_x, \omega_y, k)}{A(\omega_x, \omega_y, 1)}$. The inverse FT of $\Phi_k(\omega_x, \omega_y)$ is given by $\varphi_k(x, y) = \sum_{i=1}^{M} \gamma_i(x, y)\delta(x - b_i^x(k), y - b_i^y(k)) + \gamma_n(x, y, k)$, so it has peaks at $(x, y) = (b_i^x(k), b_i^y(k))$, for $1 \leq i \leq M$, which correspond to the "mean" translations of each object's centroid from frame 1 to k. As the experimental results show (Sec. VII-F), compensating for the mean translation retains only the periodic motion component.

It should be noted that the analysis of periodic motion superposed on translation does not require prior knowledge of the types of motion present in the video. The presence of periodic

motions in the video is detected by the TFDs. If there are no periodicities, the TFDs will simply give $T = 0$. Similarly, the presence of translations is detected by the phase correlation stage. If the motion is purely periodic, the translation estimate is equal to zero. After applying the TFD analysis and the phase correlation, we are able to characterize the motions present as purely periodic, purely translational, or periodic superposed on translational.

VI. Sensitivity Analysis for Gaussian Noise

Repetitive motions that appear in nature or man-made applications are often not strictly periodic: their period may fluctuate around a "mean period", and the trajectory's magnitude may also fluctuate around values that re-appear. We examine motion in the x-direction only, as the same analysis applies to the y-direction. Consider an ideal periodic trajectory $b^x(t) = b^x(t+T)$, and a nearly periodic trajectory $b^{x\prime}(t) = b^x(t+T') + \epsilon_2$, where $T' = T + \epsilon_1$, $\epsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$, $\epsilon_2 \sim \mathcal{N}(0, \sigma_2^2)$. We assume Gaussian variations on the trajectory period and magnitude, as this model can be considered a general approximation of all possible random variations, whose effect can be analyzed mathematically.

In the general case, the signal under examination is of the form $d_x(\mu, k)$ of Eq. (9). For simplicity of analysis, here we will consider the case of only one object ($M = 1$), without noise, in continuous time: $d_x(\mu, t) = S(\mu)e^{j\mu b^x(t)}$. The STFT of $d_x(\mu, t)$ is given by:

$$STFT(t,\omega) = \int d_x(\mu, t+\tau)h^*(\tau)e^{-j\omega\tau}d\tau = \int S(\mu)e^{j\mu b^x(t+\tau)}h^*(\tau)e^{-j\omega\tau}d\tau, \qquad (17)$$

where $S(\mu)$ is the FT of the x-projection. For a near-periodic trajectory $x'(t)$, Eq. (17) becomes:

$$STFT'(t,\omega) = \int d'_x(\mu, t+\tau)h^*(\tau)e^{-j\omega\tau}d\tau = S(\mu)\int e^{j\mu(b^x(t+\tau+T+\epsilon_1)+\epsilon_2)}h^*(\tau)e^{-j\omega\tau}d\tau. \quad (18)$$

Obviously the noise in the trajectory period and magnitude introduces errors in the STFT, which now is a random quantity. Its mean, w.r.t. the random quantities $\epsilon_1,\ \epsilon_2$, is:

$$E_{\epsilon_1,\epsilon_2}[STFT'(t,\omega)] = E_{\epsilon_1}E_{\epsilon_2}[STFT'(t,\omega)] = E_{\epsilon_1}[E_{\epsilon_2}[STFT'(t,\omega)]], \qquad (19)$$

where

$$E_{\epsilon_2}[STFT'(t,\omega)] = E_{\epsilon_2}[S(\mu)\textstyle\int e^{j\mu(b^x(t+\tau+T+\epsilon_1)+\epsilon_2)}h^*(\tau)e^{-j\omega\tau}d\tau]$$
$$= E_{\epsilon_2}[e^{j\mu\epsilon_2}]S(\mu)\textstyle\int e^{j\mu b^x(t+\tau+T+\epsilon_1)}h^*(\tau)e^{-j\omega\tau}d\tau = E_{\epsilon_2}[e^{j\mu\epsilon_2}]F(\epsilon_1), \qquad (20)$$

for $F(\epsilon_1) = S(\mu) \int e^{j\mu b^x(t+\tau+T+\epsilon_1)} h^*(\tau) e^{-j\omega\tau} d\tau$. We first consider the case where $\epsilon_1 = 0$, i.e. $F(\epsilon_1) = STFT(t,\omega)$, and $E_{\epsilon_1,\epsilon_2}[STFT'(t,\omega)] = E_{\epsilon_2}[e^{j\mu\epsilon_2}]STFT(t,\omega)$. Then

$$E_{\epsilon_2}[e^{j\mu\epsilon_2}] = \frac{1}{\sqrt{2\pi}\sigma_2} \int_{-\Delta_2}^{\Delta_2} e^{j\mu\epsilon_2} e^{-\epsilon_2^2/2\sigma_2^2} d\epsilon_2 = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\mu^2\sigma_2^2} \int_{-\Delta_2}^{\Delta_2} \exp\left[-\frac{1}{2}\left(\frac{\epsilon_2}{\sigma_2} - j\mu\sigma_2\right)^2\right] d\epsilon_2. \tag{21}$$

Let $y = \frac{\epsilon_2}{\sigma_2} - j\mu\sigma_2$. Then Eq. (21) becomes:

$$E_{\epsilon_2}[e^{j\mu\epsilon_2}] = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mu^2\sigma_2^2} \int_{-\Delta_2/\sigma_2 - j\mu\sigma_2}^{\Delta_2/\sigma_2 - j\mu\sigma_2} \exp\left(-\frac{1}{2}y^2\right) dy. \tag{22}$$

When $\sigma_2 \to 0$, i.e. when there is little additive noise on the trajectory, the limits of the integral are $\Delta_2/\sigma_2 - j\mu\sigma_2 \to +\infty$ and $-\Delta_2/\sigma_2 - j\mu\sigma_2 \to -\infty$, so Eq. (22) becomes:

$$E_{\epsilon_2}[e^{j\mu\epsilon_2}] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy. \tag{23}$$

Eq. (23) is the integral of the Normal distribution over all values of y, which is well known [41] to be equal to 1. Thus, it is shown that for $\sigma_2 \to 0$, $E_{\epsilon_2}[e^{j\mu\epsilon_2}] \to 1$. This makes sense, since it proves that when there is very little, or no additive noise in the trajectory, the mean of the corresponding STFT is not affected by it. When $\sigma_2 \to \infty$, the limits of the integral are $\Delta_2/\sigma_2 - \mu j\sigma_2 \to -\infty$ and $-\Delta_2/\sigma_2 - \mu j\sigma_2 \to -\infty$. The limits of the integral tend to become the same, so the mean of the STFT tends to become 0 for increasing noise, i.e. noise of infinite variance will significantly degrade the STFT estimate.

We performed Monte-Carlo experiments, where we generated 10000 Normally distributed random variables ϵ_2, with standard deviation σ_2 varying from 0 to 5. The random variables ϵ_2 were used to create random signals $e^{j\epsilon_2}$, since we are interested in estimating the mean $E_{\epsilon_2}[e^{j\mu\epsilon_2}]$. Note that, without loss of generality, we have set $\mu = 1$. For each value of σ_2, we estimate the mean and variance of $e^{j\epsilon_2}$ over the 10000 Monte-Carlo trials. This leads to the results of Fig. 1, where we see how the magnitude of the mean $E_{\epsilon_2}[e^{j\epsilon_2}]$ and the variance $var_{\epsilon_2}[e^{j\epsilon_2}]$ of $e^{j\epsilon_2}$ change for increasing σ_2, and consequently for noisier data. Fig. 1(a) shows that, indeed, as $\sigma_2 \to 0$, the averaged effect of the noise ϵ_2 on the STFT estimate, approaches 1, so the STFT estimate remains close to its true value when the noise is not significant. As the noise increases, the STFT's magnitude will decrease, on average. However, from Fig. 1(a) we see that for $0 \leq \sigma_2 \leq 0.8$, the magnitude remains above 70% of its actual value, i.e. the STFT is not significantly degraded when there is additive noise on the object's trajectory.

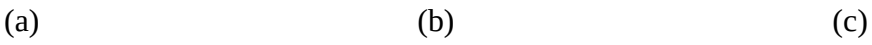

Fig. 1. $\sigma_2 \in [0, 5]$: (a) Magnitude of mean $E_{\epsilon_2}[e^{j\epsilon_2}]$. (b) Variance of $e^{j\epsilon_2}$. (c) Mean error in the STFT.

The variance of $e^{j\epsilon_2}$, shown in Fig. 1(b), increases with the noise, but it is limited by the value 1. This ensures that the STFT's variability, which is introduced by the noise ϵ_2, does not increase indefinitely, even if the noise does, thus making the estimates more reliable.

The mean STFT error is $E_{\epsilon_2}[STFT(t,\omega) - STFT'(t,\omega)] = STFT(t,\omega)(1 - E_{\epsilon_2}[e^{j\mu\epsilon_2}])$. In Fig. 1(c) we show the normalized error value $1 - E_{\epsilon_2}[e^{j\mu\epsilon_2}]$, which demonstrates its behavior for varying σ_2. It is zero for zero noise and, as expected, increases as σ_2 becomes higher. However, it does not exceed the value 1, even when $\sigma_2 \to \infty$. Thus, the noise in the periodically changing object's trajectory magnitude does not affect its STFT estimate significantly, as it introduces an error whose values are bounded, even for infinitely increasing noise.

For the case of additive noise ϵ_1 in the period T, we have:

$$E_{\epsilon_1,\epsilon_2}[STFT'(t,\omega)] = E_{\epsilon_1}[F(\epsilon_1)] \cdot E_{\epsilon_2}[e^{j\mu\epsilon_2}], \tag{24}$$

where

$$\begin{aligned} E_{\epsilon_1}[F(\epsilon_1)] &= \frac{1}{\sqrt{2\pi}\sigma_1} \int\int e^{j\mu b^x(t+T+\tau+\epsilon_1)} h^*(\tau) e^{-j\omega\tau} e^{-\epsilon_1^2/2\sigma_1^2} d\tau d\epsilon_1 \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \int h^*(\tau) e^{-j\omega\tau} A(\tau) d\tau, \end{aligned} \tag{25}$$

with $A(\tau) = \int e^{j\mu b^x(t+\tau+T+\epsilon_1)} e^{-\epsilon_1^2/2\sigma_1^2} d\epsilon_1$. For $\epsilon_1 = 0$, i.e. when the signal period is constant, Eq. (25) gives the STFT of the ideal periodic signal, corresponding to the noiseless case.

$A(\tau)$ depends on the form of $b^x(t)$, but in general, the quantity in Eq. (25) is the same as the STFT of $e^{j\mu b^x(t)}$, after the signal $b^x(t)$ has been "blurred" by the Gaussian function $e^{-\epsilon_1^2/2\sigma_1^2}$.

Thus, Eq. (25) will give the time-frequency power spectrum of this "blurred" signal. Consider, for example, a sawtooth function $b^x(t) = \alpha \cdot t$, which repeats every T time instants. Then,

$$A(\tau) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{j\mu\alpha\cdot(t+\tau+T)} \int e^{j\mu\alpha\cdot\epsilon_1} e^{-\epsilon_1^2/2\sigma_1^2} d\epsilon_1 = \frac{1}{\sqrt{2\pi}\sigma_1} e^{j\mu\alpha\cdot(t+\tau+T)} E_{\epsilon_1}[e^{j\mu\alpha\cdot\epsilon_1}], \qquad (26)$$

which has the term $E_{\epsilon_1}[e^{j\mu\alpha\cdot\epsilon_1}]$, similar to Eq. (22). Thus, for $\sigma_1 \to 0$ we have $E_{\epsilon_1}[e^{j\mu\alpha\cdot\epsilon_1}] \to 1$, as before (Fig. 1), and $A(\tau) \to \frac{1}{\sqrt{2\pi}\sigma_1} e^{j\mu\alpha\cdot(t+\tau+T)}$. In this case, Eq. (25) finds the STFT of $A(\tau)$, and when $\sigma_1 \to 0$, $E_{\epsilon_1}[F(\epsilon_1)] \to STFT(t,\omega)$, so when the deviations in the periodic motion are small, the resulting $STFT'(t,\omega)$ approaches the true $STFT(t,\omega)$.

From Eqs. (22)-(26), we see that the error introduced by deviations in the period or the trajectory does not significantly degrade the periodic motion estimation. Thus, the STFT estimates will, indeed, be robust to deviations from strict periodicity in experiments with real-world sequences, containing repetitive motions that are not perfectly periodic.

VII. Experiments

Experiments are conducted both with synthetic and real sequences that contain multiple periodic motions. The goals of the experiments are to show that the proposed method can:

1) detect multiple periodic motion trajectories.
2) estimate their periods reliably.
3) segment the multiple moving objects, and
4) perform (1), (2) and (3) under deviations from periodicity.

As explained in Sec. II, the mathematical model of Eq. (2) describes the video sequence more accurately than Eq. (1), so the background is eliminated in our experiments.The background is visible in the majority of the frames, so at each pixel, the median of the intensity over the entire sequence is treated as the intensity of the (unoccluded) still background image at that pixel.

For μ-propagation, the μ value that gives the best resolution depends on the displacement magnitudes [42], [36]. However, the displacements are unknown, so the optimal value of μ is not known $\grave{a}$ priori. Therefore, for each sequence we test all values of μ, from 0.01 to 1, with a step size of 0.01 (except for the Ping Pong-sequence, where we use step size 0.001 to refine the results), and choose the one that gives the best resolution in the ridges of the resulting TFD (Sec. IV).

Results are presented for increasingly complex sequences. We begin with a simple synthetic sequence to demonstrate the various steps in the algorithm. We then use a simple but real sequence, which contains image noise and small deviations from the model of perfect periodicity of motion. This is followed by sequences which contain more image noise, greater deviations from periodicities, addition of translational motion components, and where the moving objects are non-rigid. This allows us to evaluate the robustness of our method under such violations of the model on which it is based. Finally, we test our algorithms on video sequences containing periodic motion superposed on translation.

The ground truth for the periods in the real sequences can be obtained by simply observing the video or its projections in the x and y directions. For more accuracy, we fit a curve to the "center of mass" of each projection curve (e.g. Fig. 3, 6(b), (c)), and estimate its period. This leads to reliable estimates of the period ground truth, with standard deviations whose magnitude does not exceed 0.002 in all cases. We refer to these period estimates as the ground truth (for the motion periods) in the experiments.

The real sequences that are examined (Pendulum, Dribbling, Ping Pong, Swings) are all color, except the Walking sequence. The periodic motion analysis is conducted using the sequence of intensity frames only, since they contain all the motion information of the original color sequence. Alternatively, the R, G, B color components of the video could each be analyzed separately, with the proposed method, to extract the periodicities. However, this would demand three times as many computations, and would not add any information about the motions. Consequently, we only use the frame intensities for the motion estimation, even though our method is actually being applied to color sequences.

The method we use is computationally efficient: all steps of the algorithm take anywhere from fractions of a second to less than 5 seconds. The TFD estimates may require up to two seconds on a Pentium 4 computer, when the video frames have dimensions above 240×320. The correlations for the segmentation require at most 5 seconds, when the frames are large and the block sizes are small.

A. Synthetic Sequence

Our first synthetic sequence contains two planar periodic motions. Two 2D objects (Fig. 2) move diagonally, with their x and y velocities shown in Fig. 2(b). The background was deliber-

(a) (b)

Fig. 2. (a) First frame of a synthetic sequence with two objects, moving periodically in the x and y directions. (b) The solid (dotted) curve shows the velocity of the left (right) box which is in both x and y directions, with a period of 100 (50) frames.

(a) (b)

Fig. 3. (a) Projections of the two object sequence in the x-direction. (b) Projections of the sequence in the y-direction. The horizontal axis denotes time, and the vertical axis denotes the summed up image luminance values in both cases.

ately chosen to be black, to avoid background subtraction in this synthetic sequence. The x and y projections of the sequence are shown in Figs. 3(a), (b): the horizontal axis indicates time, and the vertical axis gives the profile of the summed-up y (or x) values over time. For clarity of presentation, we show a binarized version of the projections (rather than a grayscale plot with the actual values of these sums), to emphasize the periodic patterns in them, caused by the object motions.

Using μ-propagation (Sec. IV), with $\mu = 0.91$ found to give the best TFD resolution, after

(a) (b) (c)

Fig. 4. (a) STFT for x-projections. (b) Peak values of the 2D STFT, showing the variation of the signal's frequency with time. (c) Power spectrum of the sum of the STFT peaks estimated with the MUSIC algorithm, shown on a log scale; the periods of the two moving objects appear as the two dominant peaks with the rest of the signal being ≥ 4 dB weaker. The dominant periods are $T = 2$ and $T = 4$.

searching in the interval 0.01 to 1, we obtain the STFT in Fig. 4(a). The STFT energy is highest at the frequencies that are present in the signal at each time instant, which are proportional to the varying object velocities.

In order to separate the two object trajectories, we sum the STFT peak values (Fig. 4(b)) at each time instant, and apply standard spectral analysis techniques to the sum signal (MUSIC algorithm), to get the dominant frequencies that correspond to the component periods. Fig. 4(c) shows that the signals' periods, $T_1 = 2, T_2 = 4$, are successfully extracted. The two objects are then extracted by correlating frames that are separated by an integer number of periods (Fig. 5). The results are very accurate, as the sequence is synthetic, with no measurement noise, or background present.

B. Pendulum

Our next sequence is a real but simple sequence containing a pendulum undergoing periodic motion, amid sensor noise and the noisy effect of its moving shadow. Fig. 6(a) shows the first frame of this sequence. The background is removed and the magnitudes of the projections on the x and y axes are shown in Figs. 6 (b) and (c) respectively, as a function of time (i.e. frame number, shown on the horizontal axis). As before, μ-propagation is applied to the horizontal and

(a) (b)

Fig. 5. Extracted objects. (a) Top left. (b) Bottom right.

(a) (b) (c)

Fig. 6. Pendulum sequence: (a) First frame. Projections (with the background removed): (b) in the x-direction, (c) in the y-direction. The horizontal axis denotes time, and the vertical axis denotes the summed up image values.

vertical projections, this time with the values $\mu = 0.04$ and $\mu = 0.1$ respectively, found to be the best after the search over the range 0.01 to 1.

The STFTs for the x and y projections are shown in Fig. 7, and spectral analysis of their maxima leads to the period estimate $T = 3$, in the x direction (Fig. 8(a)). From Fig. 6(b) we see that the ground truth is $T = 2.85$, so the estimate is close to the true value of the horizontal period. Similarly, in Fig. 8(b) we estimate the y-period as $T = 5.23$, compared to the ground truth 5.7. The pendulum shadow does not affect the estimation significantly, because its intensity is very low, and close to that of the background, whereas the pendulum is much brighter. Finally, the pendulum itself is extracted by correlating frames that are separated by an integer number of periods in both x and y directions.

(a) (b)

Fig. 7. Pendulum sequence: STFT for the projections of the sequence in the (a) x-direction, and (b) y-direction.

(a) (b) (c)

Fig. 8. Pendulum sequence: (a) Power spectrum in the x direction. The first peak ($T = 3$) is about 17 dB higher than the next weaker one. (b) Power spectrum in the y direction. (c) Extracted object (pendulum).

C. Dribbling

This sequence shows a girl dribbling a basketball in a "V-shaped" motion (Fig. 9(a)). This is a slightly more difficult sequence than the pendulum, since the girl's hands do not perform the exact same motion each time and the illumination is not constant. Fig. 9(b), (c) show the foreground mask in two frames of the sequence (i.e. the result of background removal) and Fig. 10 shows the resulting projections. We see that the basketball's resulting motion displays a strong, although not strict, periodicity. In Fig. 10 we see that the period in the x-projection of

(a) (b) (c)

Fig. 9. Dribbling sequence: (a) Frame 10. (b) Frame 30 foreground mask. (c) Frame 57 foreground mask.

(a) (b)

Fig. 10. Dribbling sequence: (a) x-projections. (b) y-projections.

the sequence, for $\mu = 0.06$, is $T = 2.925$, (ground truth $T = 3.25$), and in the y-projection, for $\mu = 0.28$, $T = 6.83$, (ground truth $T = 6.5$). The segmented basketball is shown in Fig. 12(c), against a black background. Note that only the basketball is moving periodically, so it is the only object extracted.

D. Ping Pong

This sequence, of two people playing ping-pong, is quite challenging as it contains many deviations from the model we have used: the trajectory followed by the ball does not have the

(a) (b)

Fig. 11. Dribbling sequence: (a) STFT, for the projections of the sequence in the x-direction. (b) STFT, for the projections of the sequence in the y-direction. The horizontal axis denotes time in both cases.

(a) (b) (c)

Fig. 12. Dribbling sequence: Power spectrum of the STFT maxima gives (a) $T = 2.925$ in the x-direction, (b) $T = 6.83$ in the y-direction. (c) Recovered basketball.

same shape each time it is hit, since it depends on the way the ball is hit. The players themselves move in a repetitive but not strictly periodic manner, as their body appears in slightly different positions in each frame and their arm is not always in exactly the same position. The video also has additive measurement noise, varying illumination, and is not stable.

Fig. 13 shows two frames of this sequence, containing one motion cycle for the left player. The ball does not appear in the same position in these frames, as its motion has a phase difference from the left player's motion. Due to the different ways the ball is hit each time, its phase

(a) (b)

Fig. 13. Ping Pong sequence: (a) Frame 80. (b) Frame 100.

difference undergoes small variations as well, but on average its motion has approximately the same period as the player. In Fig. 14 we show the x and y projections over the video sequence and in Figs. 15-16 we show the TFDs and the period estimates respectively. For the x projections, with $\mu = 0.013$, we get $T = 8.1$ with ground truth $T = 7.3$, and for the y projections, and $\mu = 0.19$, we find $T = 7.03$, with ground truth $T = 7.1$. Finally, Fig. 17 shows segmentation results for frames 80 and 100, where the left player's arms are in the same position. Part of the table is also recovered, due to the significant jitteriness of the camera, whereas the ball is not extracted, since it is in different positions in these two frames. Although this sequence is particularly challenging, our method has led to accurate period estimates and good segmentation results, that spatial methods would be unable to achieve.

E. Swings

This real sequence consists of two children on swings. In this video, the moving entities themselves (the children) are not rigid. Fig. 18(a), (b) show two frames of this sequence. The camera is not very stable, so the video is slightly jittery, and the illumination changes from frame to frame. As the children move, they occlude each other in some frames. Also, their trajectory is not precisely the same in each repetition. These features make this sequence particularly interesting and more challenging than the previous ones. The median of the video is shown in Fig. 18(c), and the x and y projections in Fig. 19. The illumination changes significantly

(a) (b)

Fig. 14. Ping Pong sequence projections: (a) in the x-direction, (b) in the y-direction.

(a) (b)

Fig. 15. Ping Pong sequence STFT: (a) for the projections in the x-direction. (b) for the projections in the y-direction.

in this video, so the median alone is not sufficient for background removal. Thus, we have complemented the initial median background removal with local background estimates from the neighborhood of each frame, since the intensity is more stable locally.

There are about 2.7 periods for each child in the x and y projections, and their trajectories have a "phase difference" (Fig. 19). μ-propagation is applied to both horizontal and vertical projections with $\mu = 0.04$, giving the STFT in Fig. 20. The power spectrum for the projections in each direction is shown in Fig. 21, and gives period estimates $T_x = 2.937$ and $T_y = 2.925$,

(a) (b)

Fig. 16. Ping Pong sequence: Power spectrum of the STFT maxima gives (a) $T = 8.01$ in the x-direction, (b) $T = 7.03$ in the y-direction.

(a) (b)

Fig. 17. Left player's arm and hand segmentation: (a) Frame 80. (b) Frame 100.

which are close to our observation ($T = 2.7$).

For 125 frames, we expect the objects to re-appear in the same position at every 46 frames. Their non-rigid nature makes the segmentation very challenging, for example, their legs are not always in the same position, even if the frames are separated by an integer number of periods and the rest of their body is in the same position. In Fig. 22 we show the segmentation of each child, from pairs of frames where the child's body had the same shape. The artifacts in the segmentation are due to the poor video quality, especially the significantly varying illumination between frames. This made the background removal particularly difficult (some background is still visible in the final result, e.g. beneath the boy) and introduced some noise on the segmented

(a) (b) (c)

Fig. 18. Swings sequence: (a) Frame 10. (b) Frame 60. (c) Median filtered sequence.

(a) (b)

Fig. 19. Swings sequence: (a) x projection. (b) y-projection.

objects (black spots). The boy and girl are considered to be two different moving objects, since their motions are different. In particular, they have the same period but not the same phase, so they do not appear simultaneously in the same position in frames that are separated by an integer number of periods.

The correct period estimates and segmentation results show the robustness of the proposed method. Purely spatial based methods would encounter many difficulties in such a case, as the children are non-rigid, at times they occlude each other, the periodic motion is not perfect, and the video is quite noisy, since the illumination changes significantly from frame to frame.

(a) (b)

Fig. 20. (a) STFT for the projections of the sequence in the x-direction. (b) STFT for the projections of the sequence in the y-direction.

(a) (b)

Fig. 21. Power spectrum of the STFT maxima estimated via the MUSIC algorithm. (a) x projections. (b)y projections.

(a) (b)

Fig. 22. Object extracted from (a) frame 10, (b) frame 65.

(a) (b) (c)

Fig. 23. (a) Frame 12 of the Walking sequence. (b) Frame 60 of the Walking sequence. (c) Frame 60 of the Walking sequence after compensation for the overall displacement.

F. Walking

This real sequence consists of a person walking and swaying his arms up and down as he walks. Thus, it contains not only "pure" periodic motion (from arms and legs with respect to the body), but also a translational component (body). The 80-frame sequence contains about 5 periods of leg motion and arm motion. Figure 23(a) and (b) show two frames of this sequence. The x projection in Fig. 24(a) shows the periodic component of leg and arm motion superposed on the diagonal line, which corresponds to the body translation. The person moves his arms in synchronization with his legs, so the arm and leg motions in the x direction are considered as one motion. The vertical projections of Fig. 24(b) are due to the motion of the person's arms, which are oscillatory around the same y coordinate; i.e., they do not have a translatory component. In this case there is no motion from the legs, as their y motion is insignificant in this sequence.

The maxima of the STFT (Fig. 25) are extracted in each direction, and their power spectra are shown in Fig. 26: we find that the period present in the x projection of the sequence is $T = 4.32$ with $\mu = 0.11$ (ground truth $T = 5$), and in the y projection, $T = 4.5$ with $\mu = 0.01$ (ground truth $T = 5$).

After very careful inspection of the sequence, we observed that the arms and legs always appear together in frames that are separated by an integer number of periods, so they are extracted together, as one object. Unlike the other sequences, the person's body is also extracted, because it translates between frames, i.e. it cannot be considered as a part of the static background. In

(a) (b)

Fig. 24. (a) Projections of the Walking sequence in the x direction without background effect. (b) Projections of the Walking sequence in the y direction without background effect. The horizontal axis denotes time, and the vertical axis denotes the summed up image values in both cases.

the other sequences, the non-periodically moving parts of the body (e.g. Dribbling sequence) are not extracted, since they do not move at all between frames, and are therefore treated like background.

In Fig. 23(a) and (b), the person's legs and arms in frames $n = 12$ and $n' = n + 3 \cdot n_T = 60$ are in the same position,[1] but the person's body has translated to the left. In this case, the segmentation of the periodically moving objects cannot be performed via simple frame-wise correlation. To overcome this difficulty, we first estimate the "mean" translation of the entire person between frames n and n'. This translation can be satisfactorily estimated from the phase of the FTs of the video sequence, as explained in Section V-D. We then compensate for the translation, by shifting the person back from frame n' to frame n, as shown in Fig. 23(c). The first frame and the motion compensated frame (Fig. 23(a) and (c)) can then be correlated, to give the part of the moving body that was moving periodically.

We first choose a large block size to find a large area containing the moving object, and then use smaller block sizes to refine the segmentation. Finally, the periodically moving object, which is in the same position in the initial frame and the motion-compensated frame, is successfully extracted, as shown in Fig. 26(c) by combining the correlation result and the foreground frame.

[1]For this 80 frame sequence, there are five leg periods, so the leg motion repeats every $n_T = 16$ frames.

(a) (b)

Fig. 25. (a) STFT for the projections of the sequence in the x direction. (b) STFT for the projections of the sequence in the y direction.

(a) (b) (c)

Fig. 26. Power spectrum of the STFT maxima: (a) x direction. The peak shows that the period for the legs is $T = 4.32$. (b) y direction. The peak shows that the period for the arms is $T = 4.5$. (c) Object extracted from frame 12.

The person's color is almost black, so we added a white border to distinguish him from the black background.

VIII. Evaluation Results, Comparison

In order to clearly evaluate the performance of our method, we systematically estimate the errors in the period estimates and the segmentation results. The error e_T in the period estimates T_{est} is given by the absolute difference of T_{est} and the ground truth T, i.e. $e_T = |T_{est} - T|$. The

ground truth for the object segmentation is obtained by manually segmenting out each moving object $S_i(x, y)$. The segmentation error e_S is given by the number of pixels where the extracted and actual objects differ, divided by the number of pixels in the real object area.

For our experiments, we then obtain Table I, giving the errors in the period and segmentation estimates for each sequence. We see that the errors in the period estimates are quite low, and always under 1, so the error in the estimated number of repetitions is never of the order of a period. The segmentation errors are also low, relatively to the object's real size, and usually originate from imperfect background removal, as the videos suffered from varying illumination (especially Swings, Ping Pong). Segmentation errors may also be caused by blocking artifacts, introduced by the correlation procedure. Note that in some experiments there is only one object, so there are blanks ("$-$") in the table.

TABLE I

ERRORS IN THE PERIOD ESTIMATES AND OBJECT SEGMENTATION

Video	e_T (x dir)	e_T (y dir)	e_S for object 1	e_S for object 2
Synthetic	0	0	0.1635	0.1161
Pendulum	0.15	0.47	0.22	-
Dribbling	0.325	0.33	0.097	-
Ping Pong	0.8	0.07	0.23	-
Walking	0.237	0.225	0.14	-
Swings	0.68	0.5	0.21	0.22

A. Comparison with Spatial Correlation Method

We compare the proposed method to a classic spatial correlation based method [4]. That work focuses on walking sequences, so we compare it with our method for the Walking video.

The spatial method first requires a preprocessing step to find the foreground objects (person walking) in the video. This is equivalent to the background removal step of our approach. The foreground objects then need to be aligned, so that they can then be spatially correlated for the periodicity detection. This alignment is needed only at the segmentation stage of our algorithm and has complexity $\mathcal{O}(N)$.

In the spatial method, correlations have to be performed with all the frames in the sequence, i.e. $\sum_{i=2}^{N} i = N(N+1)/2$ frame correlations are needed to acquire the similarity plot for each video, making its complexity $\mathcal{O}(N^2)$. In our method there are only 2 frame correlations, and these are needed only at the segmentation stage.

To detect repetitions in the spatial-only approach, the similarity plot is compared against a lattice, whose points need to be dense enough to ensure that the repetitions will be detected. Prior knowledge or a good initial guess (i.e. heuristics) are needed to obtain reliable period estimates. The periodicity detection process itself is also ad-hoc, as it involves comparing the location of high similarity values with that of the lattice points. False alarms may be introduced by the repetition of high correlation (similarity) values that are not caused by periodic motions. For the Walking sequence this method gave $T_x = T_y = 5$, but only after prior knowledge was used to create a dense enough lattice to detect repeating maxima in the similarity plot.

In our method, no prior knowledge of the period is needed: it is directly estimated from the spectral analysis of the TFD maxima, so false alarms are avoided. The complexity of the period estimation is of the order $\mathcal{O}(N \log(N))$, because of the TFD estimations.

Thus, correlation based methods are more likely to make erroneous estimates of the object period, unless they have good prior knowledge. They are based on heuristics for the estimation of the periods, whereas our approach is based on traditional signal processing techniques. Finally, the computational cost of our method is lower, of the order $\mathcal{O}(N)\log(N)$, as opposed to $\mathcal{O}(N^2)$ of the spatial method.

IX. Discussion

As described in Sec. I, Sec. I-B Sec. I-C, the proposed method for multiple periodic motion estimation provides an efficient and reliable way to estimate many different periods in a video sequence. It overcomes many difficulties and shortcomings of spatial-only methods, as it takes advantage of frequency data as well, and combines that information in an effective manner.

1) In Sec. V-A, we describe how the current approach can detect and estimate more than one period present in a video sequence simultaneously (Sec. VII-A, VII-E). This is in contrast to the existing literature, where the motion of every object is analyzed separately, often with the help of human intervention.

2) In Sec. VI, it is shown that the proposed approach can deal with motions that deviate from strict periodicity. This is also verified in the experiments (Sec. VII-D, VII-E).
3) Current spatial methods work with frame pairs, whereas our approach estimates the periods by processing all the frames simultaneously. Thus, it avoids problems caused by local inaccuracies, such as local illumination variation, or even inter-object occlusison. The current method could also be extended for the case of periodicities superposed on rotational motions, but would require a different approach, and is a topic for future research.
4) As shown in Sec. V-C and Sec. V-D, our approach can deal with periodic motion superposed on translation without pre-processing, as opposed to existing spatial methods.
5) Once the periods in the video are estimated, the periodically moving objects are extracted via spatial correlation methods (Sec. V-B).

X. Conclusions

In this paper we have presented a novel method for the estimation of multiple periodic motions, and the subsequent extraction of the periodically moving objects from a video sequence. The proposed method avoids many problems of existing methods, such as the use of point or feature correspondences, and the placement of markers in the video. Its novelty and a major reason for its success is that it is based on time-frequency distributions, from which multiple periodic motions are extracted. Experiments demonstrate that it can successfully extract the object periods, which agree with our empirical estimates. Finally, the estimated object periods are used to extract the corresponding objects, and thus achieve motion segmentation as well. Further areas of research include the combination of the proposed periodicity detection method with sophisticated motion segmentation algorithms, as well as its extension to more complicated kinds of motion.

References

[1] Y. A. Ivanov and A. F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 852 – 872, Aug. 2000.

[2] M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 844 – 851, Aug. 2000.

[3] C. Lu and N. Ferrier, "Repetitive motion analysis: Segmentation and event classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 258 – 263, Feb. 2004.

[4] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 781–796, Aug. 2000.

[5] D. Cremers, S. Osher, and S. Soatto, "Kernel density estimation and intrinsic alignment for knowledge-driven segmentation: Teaching level sets to walk," *Pattern Recognition*, vol. 3175, pp. 36 – 44, Feb. 2004.

[6] D. Cremers and S. Soatto, "Probabilistic and sequential computation of optical flow using temporal coherence," *International Journal on Computer Vision*, vol. 62, no. 3, pp. 249–265, May 2005.

[7] T.Brox, A.Bruhn, N.Papenberg, and J.Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic, 2004*, vol. 4, May 2004, pp. 25–36.

[8] S. Seitz and C. R. Dyer, "View-invariant analysis of cyclic motion," *International Journal of Computer Vision*, vol. 25, no. 3, pp. 231–251, 1997.

[9] P. Tsai, M. Shah, K. Keiter, and T. Kasparis, "Cyclic motion detection for motion based recognition," *Pattern Recognition*, vol. 27, no. 12, pp. 1591–1603, 1994.

[10] F. Liu and R. W. Picard, "Periodicity, directionality, and randomness: Wold features for image modeling and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 722–733, July 1996.

[11] R. Polana and R. Nelson, "Detection and recognition of periodic, nonrigid motion," *International Journal of Computer Vision*, vol. 23, no. 3, pp. 261–282, 1997.

[12] L. Wang, T. Tan, W. Hu, and H. Ning, "Automatic gait recognition based on statistical shape analysis," vol. 12, no. 9, pp. 1120–1131, Sept. 2003.

[13] A. Briassouli and N. Ahuja, "Fusion of frequency and spatial domain information for motion analysis," in *ICPR 2004, Proceedings of the 17th International Conference on Pattern Recognition*, vol. 2, Aug. 2004, pp. 175–178.

[14] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt, "Performance of optical flow techniques," in *1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June, pp. 236–242.

[15] J. Domingo, G. Ayala, and E. Dias, "A method for multiple rigid-object motion segmentation based on detection and consistent matching of relevant points in image sequences," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, April 1997, pp. 3021 –3024.

[16] P. Milanfar, "Two-dimensional matched filtering for motion estimation," *IEEE Transactions on Image Processing*, vol. 8, no. 3, pp. 438–444, March 1999.

[17] W. Hoge, D. Mitsouras, F. Rybicki, R. Mulkern, and C.-F. Westin, "Registration of multi-dimensional image data via sub-pixel resolution phase correlation," in *Proceedings IEEE Int. Conf. Image Processing 2003*, Sept. 2003, pp. 707–710.

[18] R. E. Blahut, *Fast Algorithms for Digital Signal Processing*. New York: Addison-Wesley, 1984.

[19] P. Duhamel and M. Vetterli, "Fast fourier transforms: A tutorial review," *Signal Processing*, vol. 19, pp. 259–299, 1990.

[20] J. S. Walker, *Fast Fourier Transform*, 2nd ed. Boca Raton, FL: CRC Press, 1996.

[21] R. W. Young and N. G. Kingsbury, "Frequency-domain motion estimation using a complex lapped transform," *IEEE Transactions on Image Processing*, vol. 2, no. 1, pp. 2–17, 1993.

[22] W. Yu, G. Sommer, and K. Daniilidis, "Multiple motion analysis: In spatial or in spectral domain?" *Computer Vision and Image Understanding*, vol. 90, pp. 129–152, 2003.

[23] W. Chen, G. B. Giannakis, and N. Nandhakumar, "A harmonic retrieval framework for discontinuous motion estimation," *IEEE Transactions on Image Processing*, vol. 7, no. 9, pp. 1242–1257, Sept 1998.

[24] M. Piccardi, "Background subtraction techniques: a review," in *Proc. IEEE Conf. on Systems, Man and Cybernetics*, 2004, pp. 3099–3104.

[25] R. Pless, J. Larson, S. Siebers, and B. Westover, "Evaluation of local models of dynamic backgrounds," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,Proc. CVPR 2003*, June, pp. 1063–1069.

[26] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Computer Vision Patt. Recog.*, June 1999, pp. 246–252.

[27] D. H. A. Elgammal and L. Davis, "Nonparametric model for background subtraction," in *Proc. European Conf. Computer Vision*, June 2000, pp. 751–767.

[28] L. Cohen, "Time-frequency distributions-a review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, July 1989.

[29] A. M. Sayeed and D. L. Jones, "Analysis and synthesis of multicomponent signals using positive time-frequency distributions," *IEEE Transactions on Signal Processing*, vol. 47, no. 2, pp. 493–504, Feb. 1999.

[30] R. Czerwinski and D. Jones, "Adaptive short-time Fourier analysis," *IEEE Signal Processing Letters*, vol. 4, no. 2, pp. 42–45, Feb. 1997.

[31] A. M. Sayeed and D. L. Jones, "Optimal quadratic detection and estimation using generalized joint signal representations," *IEEE Transactions on Signal Processing*, vol. 44, no. 12, pp. 3031–3043, Dec. 1996.

[32] ——, "Optimal kernels for nonstationary spectral estimation," *IEEE Transactions on Signal Processing*, vol. 43, no. 2, pp. 478–490, Feb. 1995.

[33] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, pp. 51–83, 1978.

[34] P. Kootsookos, B. Lovell, and B. Boashash, "A unified approach to the STFT, TFDs and instantaneous frequency," *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 1971–1982, 1992.

[35] I. Djurovic and S. Stankovic, "Estimation of time-varying velocities of moving objects by time-frequency representations," *IEEE Transactions on Signal Processing*, vol. 47, no. 2, pp. 493–504, Feb. 1999.

[36] S. Stankovic and I. Djurovic, "Motion parameter estimation by using time-frequency representations," *Electronics Letters*, vol. 37, no. 24, pp. 1446–1448, Nov. 2001.

[37] P. Kornprobst, R. Deriche, and G. Aubert, "Nonlinear operators in image restoration," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. Puerto-Rico: IEEE, June 1997, pp. 325–331.

[38] A. Hirani and T. Totsuka, "Combining frequency and spatial domain information for fast interactive image noise removal," in *ACM SIGGRAPH*, 1996, pp. 269–276.

[39] B. Barkat and K. Abed-Meraim, "A blind components separation procedure for FM signal analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002*, vol. 2, June 2002, pp. 1425 – 1428.

[40] S. M. Kay, *Modern Spectral Estimation, Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[41] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, 1987.

[42] D. Jones and T. Parks, "A resolution comparison of several time-frequency representations," *IEEE Transactions on Signal Processing*, vol. 40, no. 2, pp. 413 – 420, Feb. 1992.