Continuous Wavelet Transform for Time-Varying Motion Extraction

Alexia Briassouli [1], Dimitra Matsiki [1], Ioannis Kompatsiaris [1]

[1] Informatics and Telematics Institute, Centre for Research and Technology Hellas

Abstract

Widespread use of digital multimedia data has led to the development of advanced processing techniques for them. Motion estimation is a fundamental step in the extraction of activity in videos, for tracking, motion segmentation, video classification and other applications. One large category of methods processes data in the spatial domain, and another in the frequency domain. This paper presents a novel approach for frequency-domain motion estimation, which avoids drawbacks of illumination-based methods, based on the the Continuous Wavelet Transform (CWT). The proposed algorithm processes all video frames at the same time to create a frequency modulated (FM) signal, which is then processed using the CWT. The proposed system is shown to be robust to local measurement noise and occlusions, as it processes the available data in a global, integrated manner. Experiments take place with both synthetic and real video sequences to demonstrate the capabilities of the proposed approach.

1 Introduction

The problem of motion estimation is fundamental for processing of videos, and has consequently received much attention in the literature. A vast range of approaches has been developed for the extraction of motion from moving image sequences, with respective advantages and drawbacks. One large category of motion estimation methods processes data in the spatial domain, based on the differential optical flow equation, [3], on the matching of features or video frame blocks [8]. The motion field they produce is based on the principle that changes in illumination are introduced by motion. Such methods are spatiotemporally local, so they are sensitive to spatially and temporally local measurement noise, local changes in feature appearance, and occlusions. Robust flow estimation techniques overcome such limitations, essentially by eliminating outlier flow values [4], but often have a high computational cost. A second category of approaches for extraction of motion from video is based on spatiotemporally global or frequency domain processing of the available data [5], [6], [9]. Spatiotemporal methods consider that constant motions form energy planes in the the 3D spatiotemporal spectrum [10] and focus on the detection of the planes. These methods address motion estimation in a way that is in accordance with the functioning of the human visual system [12]. Also, processing of the entire video provides more accurate motion estimates than using pairs of frames [3]. For example, the occlusion of an object over two frames would lead to completely erroneous motion estimates, if those two frames were used to obtain the inter-frame displacement. A global method easily overcomes this, as it extracts the motion over the entire time sequence, with a few local inaccuracies, that can be accounted for by imposing a smoothness constraint. Frequency domain global processing has the additional advantage of being robust to illumination variations, noise, and being computationally efficient [5], [6].

Motivation, Contributions

Existing work on frequency-domain processing of video for motion estimation either assumes that inter-frame displacements are constant [10], or time-varying motions are handled by processing pairs of frames [6], as the Fourier Transform (FT) cannot handle time-varying spectral content of the entire video. This motivates us to use the Continuous Wavelet Transform (CWT), which detects time-varying frequencies, i.e. non-stationary spectra. The CWT has been employed in [13] for the processing of trajectories extracted from gestures. In this work, the trajectory itself will be extracted from video by applying the CWT to an appropriately formed FM-signal via the "μ-propagation" technique (Sec. 3). The proposed method has a number of advantages:

(1) Since all video frames are used simultaneously, in space/time, it is robust to spatially/temporally local noise.

(2) Unlike spatiotemporal filtering based methods [9], no prior knowledge is required about the motions taking place.

(3) The CWT can be implemented based on the Fast Fourier Transform (FFT), thus lowering its computational cost [1].

(4) The CWT also provides a visualization of the magnitude of wavelet coefficients, which allows us to observe when and which frequencies are stimulated, their duration, time evolution and their density.

(5) The entire motion trajectory is obtained by processing the entire video only once.

The proposed approach forms a frequency modulated (FM) signal from the frames, whose frequency varies in time in proportion to the object motion, with the method of μ-propagation. We then apply the CWT to the FM signal, to extract its time-varying frequency and, consequently, the time-varying trajectory. The paper is organized as follows. In Sec. 2 the basic principles of the CWT are presented. Sec. 3 describes the algorithm used to construct the FM signal from which the trajectories will be extracted. The extension of this method for multiple moving objects and trajectories is provided in Sec. 3 and discussion on the choice of the μ-parameter is included in Sec. 3. Experimental results with synthetic and real video sequences are presented in Sec. 4, and conclusions are drawn in Sec. 5.

2 Continuous Wavelet Transform

The CWT has received considerable attention, as it is able to analyze signals with non-stationary spectra [1], by convolving the signal under investigation with the "wavelet signal". The latter can be shifted in space and scaled, leading to large values of the wavelet transform when its spatial location and amount of scaling matches the signal. For a time-varying signal $s(t)$, its CWT is defined as:

$$W_s(a,b) = \int_{-\infty}^{+\infty} s(t)\psi^*_{a,b}(t)dt, \quad (1)$$

where * represents complex conjugation, $\psi^*_{a,b}(t)$ is the mother wavelet, scaled by a and dilated by b:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right). \quad (2)$$

Thus, (4) becomes:

$$W_s(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} s(t)\psi^*\left(\frac{t-b}{a}\right)dt. \quad (3)$$

For each scale a, the total amount of signal energy is:

$$E(a) = \frac{1}{C_g}\int_{-\infty}^{+\infty} |W_s(a,b)|^2 db, \quad (4)$$

where $C_g = \int_0^{+\infty} \frac{|\hat{\psi}(f)|^2}{f}df < \infty$ is the admissibility constraint for the transform [1]. The peaks in $E(a)$ indicate which scales are dominant in $s(t)$. In practice we are interested in the energy spectrum of $s(t)$, so we estimate $f(t)$ based on its relation to the wavelet scale a. Is is shown that the frequency associated with a wavelet of scale a is given by $f = f_c/a$, where f_c is the pass-band center of the mother wavelet [1]. The wavelet used in this work is the Morlet wavelet, defined as:

$$\psi(t) = \frac{1}{\sqrt{\pi f_b}}e^{i2\pi f_c t}e^{-t^2/f_b}, \quad (5)$$

which is essentially a complex wave $e^{i2\pi f_c t}$ with a Gaussian envelope (e^{-t^2/f_b}). Here, f_c is the central frequency of the wavelet, and f_b is the bandwidth parameter, which controls how much the wavelet "confines" $s(t)$.

3 μ-Propagation for FM Signal Construction

The Morlet wavelet, presented in the previous section, is to be applied to a video signal, in order to extract the time-varying trajectories in it. The method used to extract these trajectories is inspired by the subspace-based line detection algorithm (SLIDE) of [2], and in particular the μ-propagation scheme employed in those works.

One moving object

A video consists of a time-series of two-dimensional frames, with luminance values $f(\bar{r},t)$ at each pixel location $\bar{r} = [x,y]$, and at frame t. For simplicity, we first consider the case of a static background $s_b(\bar{r},t)$ and one object $s_o(\bar{r},t)$, so frame 1 is:

$$f(\bar{r},1) = s_b(\bar{r}) + s_o(\bar{r},1) + v(\bar{r},1), \quad (6)$$

For a video with N frames, frame t, $1 \le t \le N$ is:

$$f(\bar{r},t) = s_b(\bar{r}) + s_o(\bar{r} - \bar{d}(t),1) + v(\bar{r},t), \quad (7)$$

where $\bar{d}(t) = [d_x(t), d_y(t)]$ is object displacement, $v(\bar{r},t)$ is measurement noise and modeling errors, from frame 1 to t. In practice, a different part of the background $s_b(\bar{r})$ is occluded by the object in each frame t. However, neither the precise background area, nor the precise location of the object are known, so this occlusion cannot be accurately modeled a priori and it is incorporated in the noise term $v(\bar{r},t)$ (measurement noise and the modeling inaccuracy [6]). In practice only small parts of the background are covered and uncovered by the moving objects, throughout the video, so this modeling error is often insignificant and does not introduce large inaccuracies to the frequency estimation. However, in order to increase the accuracy and reliability of the system, the background can be removed by any of the well-known background removal methods [11]. This reduces the effect of $v(\bar{r},t)$ (which now only represents measurement noise), as there is no inaccuracy in the mathematical model of Eqs. (6), (7), and allows us to project the video frames in the x and y directions without interference from background luminance values, as follows:

$$\begin{aligned} f_x(x,t) &= \textstyle\sum_y \Big(s_o(x-d_x(t), y-d_y(t), 1) + v(x,y,t) \Big) \\ &= s_{x,o}(x-d_x(t),1) + v_x(x,t). \\ f_y(y,t) &= \textstyle\sum_x \Big(s_o(x-d_x(t), y-d_y(t), 1) + v(x,y,t) \Big) \\ &= s_{y,o}(y-d_y(t),1) + v_y(y,t) \end{aligned} \quad (8)$$

The method of μ-propagation [2], [7] is employed in order to extract the time-varying object displacement. This method essentially constructs a frequency modulated (FM) signal from the original, one-dimensional signal, as follows:

$$F_x(\mu,t) = \sum_x f_x(x,t)e^{j\mu x} = S_{x,o}(\mu)e^{j\mu d_x(t)} + V_x(\mu,t), \quad (9)$$

where:

$$S_{x,o}(\mu) = \sum_x s_{x,o}(x,1)e^{j\mu x}, \quad V_x(\mu,t) = \sum_x v_x(x,t)e^{j\mu x}.$$

$F_x(\mu,t)$ of Eq. (9) is an FM signal, as it contains the time-varying displacement $d_x(t)$ in its phase $\phi(t) = \mu d_x(t)$.

The CWT of $F_x(\mu,t)$ has the most energy along the time-varying frequency $\mu d_x(t)$. Consequently, in order to extract time-varying motions, we find at which frequencies the CWT's energy is maximized, for each time instant t. This results in:

$$f_{est}(t) = \frac{d\phi(t)}{dt} = \frac{d(\mu d_x(t))}{dt} = \mu\frac{d(d_x(t))}{dt} = \mu u_x(t), \quad (10)$$

which gives the object velocity, since μ is a known constant. It should be noted that we present the case of μ-propagation only for the x-projection, as the same analysis applies to the y-projection $f_y(y,t)$.

Multiple moving objects

The model of the previous section is extended to the case of multiple moving objects in a video. The simultaneous processing of the entire video helps overcome problems created by local occlusions, e.g. if one object hides another, as our method uses information from all frames at once, and can thus overcome local errors. Such a case is shown in Sec. 4, where the method successfully extracts the trajectories of two boxes which occlude each other over a few frames. The model of (7), for M objects is:

$$f(\bar{r},t) = s_b(\bar{r}) + \sum_{i=1}^{M} s_i(\bar{r} - \bar{d}_i(t), 1) + v(\bar{r},t). \qquad (11)$$

After background removal and μ-propagation on the horizontal and vertical projections of Eq. (11), we obtain:

$$\begin{aligned} F_x(\mu,t) &= \sum_{i=1}^{M} S_{x,i}(\mu) e^{j\mu d_{x,i}(t)} + V_x(\mu,t), \qquad (12) \\ F_y(\mu,t) &= \sum_{i=1}^{M} S_{y,i}(\mu) e^{j\mu d_{y,i}(t)} + V_y(\mu,t), \end{aligned}$$

where:

$$S_{x,i}(\mu) = \sum_x s_{x,i}(x,1) e^{j\mu x},\ S_{y,i}(\mu) = \sum_y s_{y,i}(y,1) e^{j\mu y}. \qquad (13)$$

The CWT is then applied to $F_x(\mu,t)$, and the resulting signal has maxima at $f_i(t) = \mu d_{x,i}(t),\ i = 1,...,M$, from which the object velocities $u_{x,i}(t) = \frac{d(d_{x,i}(t))}{dt}$ can be found as in Eq. (10).

Selection of the μ parameter

In order to obtain an accurate representation of the time-varying frequency of the signals $F_x(\mu,t)$ and $F_y(\mu,t)$, we would ideally like to have an optimal value for the parameter μ. As explained in the literature [7], higher values of μ lead to a higher resolution in the velocity estimation, but allow the estimation of lower frequencies and velocities. The scales $a(t)$ for the CWT are related to the actual frequencies [1] as follows:

$$f(t) = \frac{f_c}{a(t)\Delta}, \qquad (14)$$

where f_c is the central frequency for the wavelet used, $a(t)$ is the scale extracted by the CWT, and Δ is the sampling period. The maximal frequency obtained is given by $f_{max} = \mu u_{max}$. In the experiments, a range of values for the scales a is pre-selected, and the correspondence between the estimated scales/frequencies and actual object velocities depends on this range. In this work we are using the Morlet wavelet, with $f_c = 0.8125$, and $\Delta = 1$, so Eqs. (14) and (10) become:

$$\mu u_x(t) = \frac{0.8125}{a(t)}. \qquad (15)$$

Since both $u_x(t)$ and $a(t)$ are unknown, we cannot predetermine the optimal value of μ with no prior knowledge of the motions present and/or the range of $a(t)$ which is optimal for our application. Consequently, we currently determine μ experimentally, by examining a range of values for $a(t)$, μ, and the resolution of the resulting velocity/displacement estimates $u(t)$. Future research is currently underway for optimally determining the value of μ in a non-empirical manner.

4 Experiments

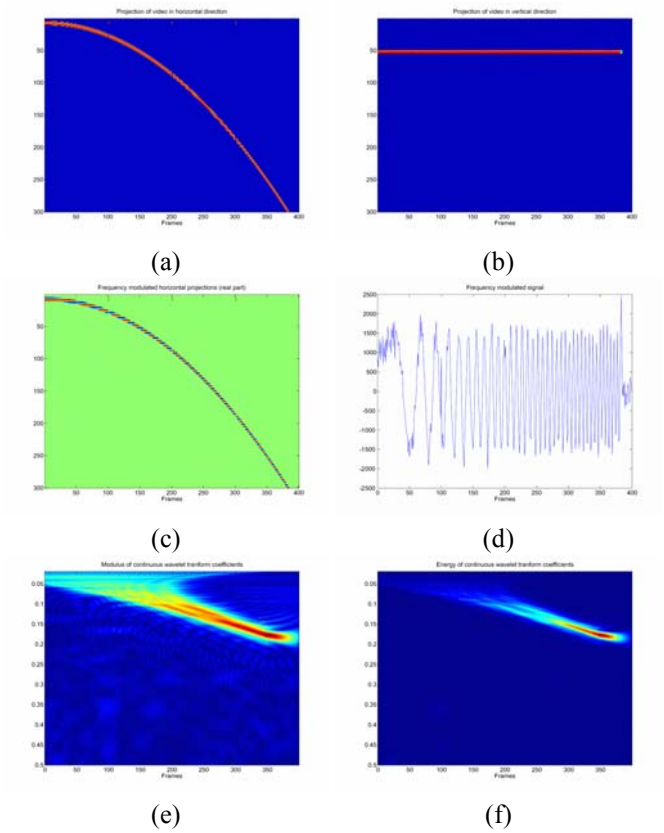
In order to test the proposed approach, we performed a number of experiments, with synthetic and real videos. The purpose of using synthetic videos is to control the precise parameters of the object displacements, and thus have ground truth available, for the verification of the trajectory estimation results.

Parabolic displacement for one object

In this experiment, we examined the performance of the proposed algorithm for a square that translates over a black background with a parabolically increasing displacement (see http://inf-server.inf.uth.gr/ briassou/box1.zip). The displacement is set to $r_1(t) = 0.002t^2$, so the velocity is $u_1(t) = 0.004t$. The scales $a(t)$ were set in the range $[0.02, 0.5]$, with a step size of 0.001, and we found that $\mu = 0.8$ led to the good results of this experiment. Figs. 1(a), (b) show the horizontal and vertical projections of this video, over time, corresponding to $f_x(t)$, $f_y(t)$ of Eq.(8). The horizontal motion is parabolic and there is no motion in the vertical direction. Fig. 1(c) shows the real part of $f_x(t)e^{j\mu x}$ over time, where the frequency modulation from the μ-propagation forms sinusoidal "stripes" in the curves. The real part of the resulting time-varying FM signal $F_x(\mu,t) = S_{x,o}(\mu)e^{j\mu d_x(t)}$ (Eq. (9)) is plotted in Fig. 1(d). Here, it is already obvious that its frequency increases with time, as expected, since it is proportional to $u_1(t) = 0.004t$. The wavelet transform is then applied to $F_x(\mu,t)$, for a range of scales $[0.02, 0.05]$, and its results are shown in Fig. 1(e), (f): the first figure shows the CWT, and the second its squared magnitude. The moving object has a parabolic displacement, so its velocity increases linearly. Indeed, Fig. 1(e), (f) shows that the proposed method succeessfully extracted the linear velocity of the object motion. Note that this figure shows the extracted scales $a(t)$ as they vary with time, which *decrease*, since they are inversely proportional to the linearly increasing velocity and frequency.

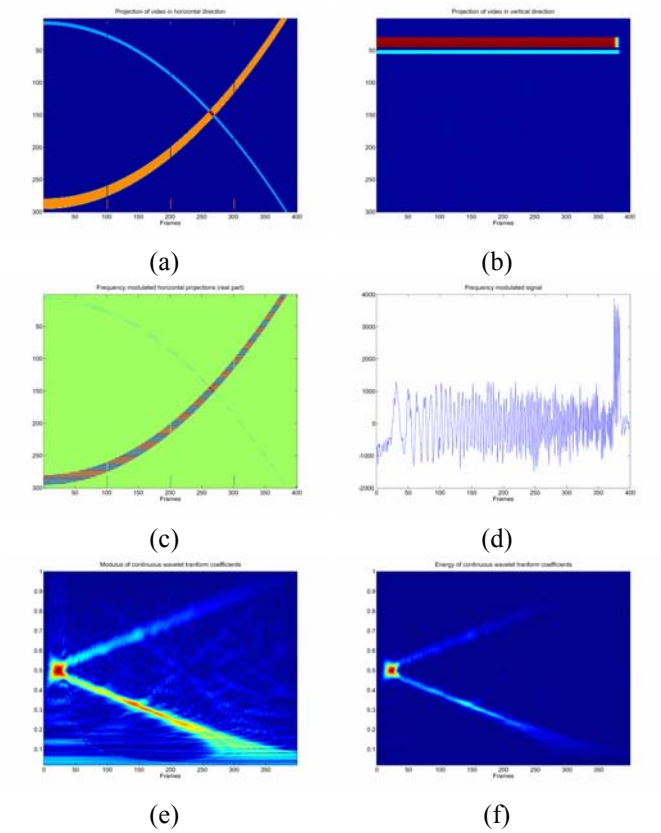
Parabolic displacement for two objects

In this experiment, two squares are translating with a parabolically increasing displacement, towards each other (see http://inf-server.inf.uth.gr/ briassou/box2.zip), so their



(a) (b) (c) (d) (e) (f)

Figure 1: (a), (b): Projections of the synthetic sequence in the horizontal and vertical directions. The parabolic nature of the object displacement is evident in the horizontal projection. There is no motion in the vertical direction. (c): Result of μ-propagation. The frequency modulation creates sinusoidal patterns in the projections. (d): Real part of frequency modulated horizontal projections of the video. (e), (f): CWT and its energy for a parabolic trajectory, i.e. linear velocity. The energy corresponds to the scales $a(t)$, which decrease linearly with time, since $f(t)$ increases linearly and $a(t) \sim 1/f(t)$.



(a) (b) (c) (d) (e) (f)

Figure 2: (a), (b): Projections of the synthetic sequence in the horizontal and vertical directions. The parabolic nature of the two object displacements is evident in the horizontal projection, as well as the point where their paths cross. (c) Effect of FM modulation. (d) Real part of frequency modulated horizontal projections of the video. (e), (f): CWT and its energy for two parabolic trajectories, i.e. linear velocities. The energy corresponds to the scales $a(t)$, which decrease linearly with time, since $f(t)$ increases linearly and $a(t) \sim 1/f(t)$.

respective velocities vary linearly with time. Their projections are shown in Fig. 2(a), (b). We used the same velocity model as in Sec. 4, but the second square moves in the opposite direction from the first. Fig. 2(c) shows the real part of $f_x(t)e^{j\mu x}$ over time, where sinusoidal patterns appear in both curves, from the frequency modulation of the μ-propagation step. The real part of $F_x(\mu, t)$, which is a superposition of two sinusoidally modulated signals, is plotted in Fig. 2(d). For satisfactory resolution, we set $\mu = 2$, and scales in $[0.02, 1]$, with a step size of 0.001. Fig. 2(e) shows the CWT, and Fig. 2(f) shows its squared magnitude. The energy is higher at scales, and hence frequencies, that are symmetric, because they correspond to the two object motions that are in the opposite direction. Thus, the proposed method succeeded to extract both linear velocities for the two objects. Both trajectories were parabolic, and varied in the same way with time, but with opposite directions, so the extracted velocities (their time derivatives) are linear, with opposite slopes, as expected. This experiment demonstrates the strength of the proposed method, as it is able to extract time-varying trajectories even for multiple moving objects in a video, which occlude each other over some frames. As Figs. 2(e), (f) show, the two trajectories begin at the same point at the first frame, which indeed is correct, as it corresponds to $u_1(t) = 0.004t = 0$, $u_2(t) = -0.004t = 0$ for $t = 0$. For the frames that follow, we also note that the slopes of the resulting linear velocities are opposite, which, again, agrees with the ground truth, namely that the objects are moving in opposite directions.

Real surveillance video of person walking with a parabolic displacement

In this experiment, a real indoors surveillance video of a person walking across a room was examined (see http://inf-server.inf.uth.gr/ briassou/walk.zip). Fig. 3 shows two frames, where the person is entering and almost leaving the area under

Figure 3: Real sequence of person walking. Frames 70, 145.

surveillance, in a diagonal/parabolic path. Here, we used $\mu = 0.2$ for obtain satisfactory frequency tracking and resolution, and the range of scales used was $[0.02, 0.4]$, with a step size of 0.001. The horizontal and vertical projections of this video, over time, are shown in Figs. 4(a), (b), where we see that paths followed by the person in both the horizontal and vertical directions are parabolic. Figs. 4(c), (d) show the real part of $f_x(t)e^{j\mu x}$, $f_y(t)e^{j\mu y}$ over time, where sinusoidal patterns appear in both curves, from μ-propagation. The CWT and its energy are shown in Fig. 4(e), (f): the CWT energy increases linearly with time, as expected. In this case, the increase is very small because of the frames were sampled densely, so the velocity increases were very small from frame to frame.

Real sequence of a pendulum

In this experiment, the motion of a pendulum was examined (see http://inf-server.inf.uth.gr/ briassou/pendulum.avi). Here, we used $\mu = 0.5$ for good frequency tracking and resolution, and the range of scales used was $[0.02, 0.4]$, with a step size of 0.001. The horizontal and vertical projections of this video, over time, are shown in Figs. 6(a), (b), where the periodicity in the pendulum's motion are evident. Figs. 4(c), (d) show the real part of $f_x(t)e^{j\mu x}$, $f_y(t)e^{j\mu y}$ over time, where sinusoidal patterns appear in both curves, from μ-propagation. The CWT and its energy are shown in Fig. 4(e), (f): the CWT energy varies periodically with time, as expected, so it correctly captures the pendulum's motion.

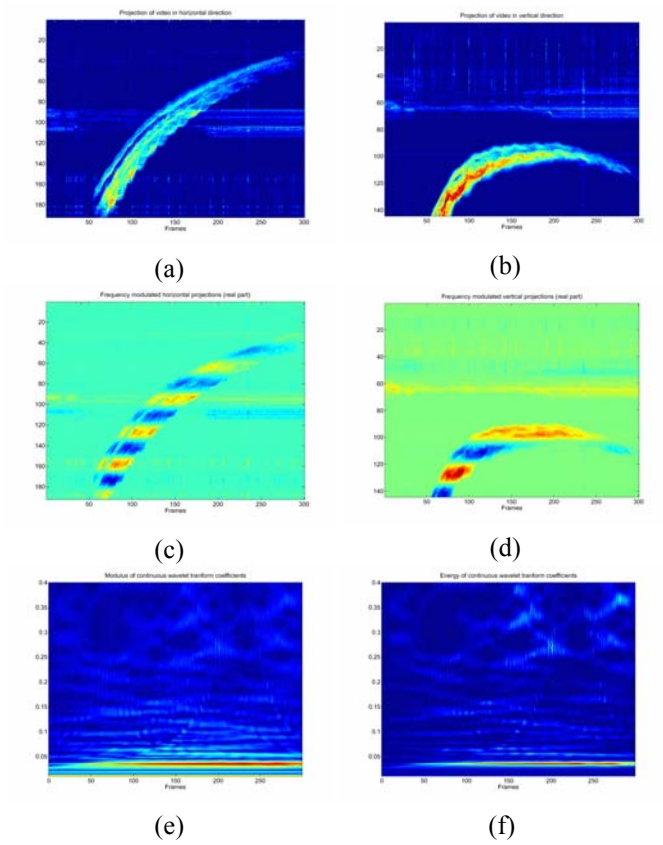


Figure 4: (a), (b): Projections of the real video in the horizontal and vertical directions. The parabolic nature of the path can be seen. (c), (d): Projections of the real sequence in the horizontal and vertical directions after μ-propagation. (e), (f): CWT for parabolic trajectory, i.e. linear velocity. The increase in speed is very small so the slope of the line is also very small.

5 Conclusions

We have presented a novel method for the extraction of trajectories from video sequences using the CWT. The proposed approach has the advantages of being robust to local spatiotemporal illumination variations, local measurement noise, object occlusions in space and time, by processing all video frames simultaneously. This makes it more robust than most local spatial methods, which require high computational complexity in order to provide robust results in such cases. Contrary to existing, global spectral methods (e.g. as FT-based), the CWT allows us to track time-varying object motions. Experiments with both synthetic and real video sequences lead to accurate trajectory extraction, even in the presence of local occlusion. Future work involves the use of

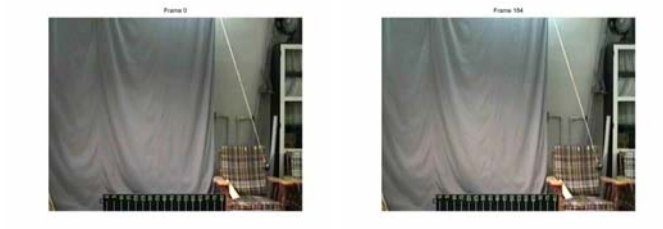


Figure 5: Real sequence of pendulum. Frames 0, 184.

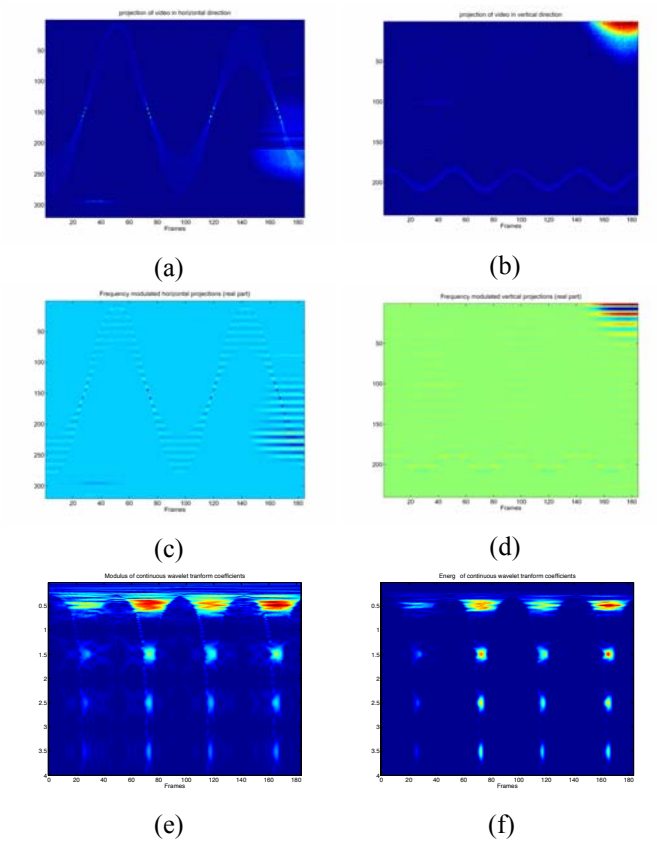


Figure 6: (a), (b): Projections of the pendulum video in the horizontal and vertical directions. The periodic nature of the path can be seen. (c), (d): Effect of μ-propagation: there is sinusoidal modulation in the projections. (e), (f): CWT for parabolic trajectory, i.e. linear velocity. The increase in speed is very small so the slope of the line is also very small.

the extracted trajectories in classification/recognition systems, as well as more refined processing of the video sequence, to deal with data of poor quality.

References

[1] P. S. Addison. *The Illustrated Wavelet Transform Handbook*. Inst. of Physics Publ., Bristol, UK, 2002.

[2] H. Aghajan and T. Kailath. SLIDE: Subspace-based line detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(11):1057–1073, Nov. 1994.

[3] J. Barron and R. Eagleson. Recursive estimation of time-varying motion and structure parameters. *Pattern Recognition*, 29(5):797–818, Dec. 1996.

[4] M. J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Proc. IEEE 4th Int. Conf. Computer Vision*, pages 231–236, May 1993.

[5] A. Briassouli and N. Ahuja. Extraction and analysis of multiple periodic motions in video sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1244–1261, July 2007.

[6] W. Chen, G. B. Giannakis, and N. Nandhakumar. A harmonic retrieval framework for discontinuous motion estimation. *IEEE Transactions on Image Processing*, 7(9):1242–1257, Sept 1998.

[7] I. Djurovic and S. Stankovic. Estimation of time-varying velocities of moving objects by time-frequency representations. *IEEE Transactions on Signal Processing*, 47(2):493–504, Feb. 1999.

[8] 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*, volume 4, pages 3021 – 3024, April 1997.

[9] D. J. Heeger. Optical flow from spatiotemporal filters. In *Proc. IEEE 1st Int. Conf. Computer Vision*, pages 181–190, June 1987.

[10] A. Kojima, N. Sakurai, and J. I. Kishigami. Motion detection using 3D-FFT spectrum. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 213–216, April 1993.

[11] M. Piccardi. Background subtraction techniques: a review. In *Proc. IEEE Conf. on Systems, Man and Cybernetics*, pages 3099–3104, 2004.

[12] B. A. Watson, A. J. Ahumada. Model of human visual-motion sensing. *J. Opt Soc. Amer. A*, 2:322342, 1985.

[13] Y. Xiong, F. Quek and D. McNeill, Hand motion gestural oscillations and multimodal discourse. In *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces, 2003*, pages = 132–139, 2003.