Continuous Wavelet Transform for Time-Varying Motion Extraction

Alexia Briassouli, Dimitra Matsiki, Ioannis Kompatsiaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas
abria@iti.gr,matsik@iti.gr,ikom@iti.gr

Abstract

The widespread use of digital multimedia data has made the development of advanced processing techniques necessary, to enable its more efficient analysis. For video content, the estimation of motion is a fundamental step in the extraction of activity, for tracking, motion segmentation, video classification and other applications. The numerous methods that have been proposed over the years for the problem of motion estimation can be divided into two categories. The first group processes data in the spatial domain, and the other in the frequency domain. In this work, an original approach for the estimation of motion in the frequency domain is presented. The proposed method avoids limitations of illumination-based methods, such as sensitivity to local illumination variations and noise by employing the Continuous Wavelet Transform (CWT). All video frames are processed simultaneously, so as to create a frequency modulated (FM) signal, which contains the motion information in its frequency. The resulting FM signal is then processed using the CWT, which extracts its time-varying frequency and consequently its motion. This 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.

I. Introduction

The problem of motion estimation has been extensively studied in the computer vision and image/video processing communities. A vast range of approaches to the problem have been developed, with their respective advantages and disadvantages. The first large group of methods estimates motion by processing the data in the spatial domain. Motion can be estimated by attributing changes in illumination to object motion, as expressed in the differential optical flow

equation [1], or by matching features or video frame blocks [2]. Such methods are spatiotemporally local, and can consequently detect small local motions, but also have the disadvantage of being sensitive to spatially and temporally local measurement noise, local changes in feature appearance, and occlusions. Robust flow estimation techniques have been developed to overcome these inaccuracies and limitations by enforcing smoothness constraints [3] or by eliminating outlier flow values [4], but often have a high computational cost. A second category of approaches is based on global spatiotemporal or frequency domain processing of the available data [5], [6], [7]. A large number of Fourier transform (FT)-based global methods consider that constant motions form energy planes in the 3D spatiotemporal spectrum [8], [9]. The plane normal is determined by the constant velocity, so the detection of the planes where energy is concentrated leads to the estimation of motion. Non-constant motions are addressed in the Fourier domain by pairwise frame processing [6], [10], or by the application of time-frequency distributions [7]. Processing of the entire video sequence has been shown to provide more accurate motion estimates than using pairs of frames [1], which is expected since all information available in the data is used at once. Global methods are robust to local occlusions, local illumination variations and measurement noise, as the errors introduced over a few frames or a few pixels can be corrected from the information available in the rest of the sequence.

A. Motivation, Contributions

This paper focuses on the extraction of motion by processing video sequences in transform rather than spatial domains by following, at the same time, a global approach. Such methods have not been investigated as much as spatial based approaches, although they have several significant advantages. Processing in the transform domain is more robust to illumination variations than spatial approaches. Motion estimation in the spatial domain, e.g. optical flow, is explicitly based on the assumption of constant illumination, which is often not the case in practice, leading to errors in the motion estimation. Such inaccuracies can be overcome by the enforcement of smoothness or continuity constraints, but these often entail a high computational cost. On the other hand, transform based methods are inherently robust to such variations, as illumination variations affect the transform's amplitude [11] and not its phase, which contains the motion information [6]. Additionally, many algorithms have been developed for the efficient computation of transforms [12], making these methods computationally efficient. Neurophysiological evidence

also suggests [13] that frequency domain motion estimation is in accordance with the functioning of the human visual system. Finally, transform based methods usually process the entire frame, rather than focusing on local pixel regions, like spatial methods. The approach examined here is global both in space and in time, processing the entire video sequence at once. This leads to the estimation of the time-varying velocity over time, in a manner that is robust to local noise, occlusions and illumination changes.

Most existing work on frequency-domain processing of video for motion estimation either assumes that inter-frame displacement in the video is constant [8], or handles time-varying motions by processing pairs of frames, instead of the entire video [6]. This is because the Fourier Transform (FT) cannot handle time-varying spectral content. Specifically, the FT phase contains the average phase shift, and consequently the average displacement, between two frames. In order to extract time-varying motion, FT-based motion estimation methods need to be applied successively, to pairs of frames. However, the signal processing literature is rich in methods other than the FT, such as time-frequence distributions or the Continuous Wavelet Transform (CWT) for the estimation of time-varying spectral content, which is often expressed as the time-varying frequency of a signal, also known as the instantaneous frequency [14], [15].

This work examines the application of the Continuous Wavelet Transform (CWT) on video for the estimation of motion. This is motivated by the fact that the CWT can process non-stationary spectra, i.e. extract time-varying frequencies [16], and has been successfully used in a wide range of applications. It has been employed extensively for the estimation of instantaneous frequencies in signals with non-stationary spectral content, for example in acoustics [17], bio-medical applications [18], [19], and the processing of trajectories extracted from gestures in [20].

Since the CWT estimates time-varying frequencies, it can be used for motion estimation if the motion information is incorporated in a signal with time-varying frequency. In order to achieve this, an FM signal is created, whose frequency is modulated by the time-varying displacement. The FM signal is created via a method known as "μ-propagation" (Sec. III). The CWT is then applied to this FM signal to extract its time-varying frequency, and thus its time-varying displacement. The proposed method is inspired by the μ-propagation method applied in Subspace Line Detection (SLIDE) techniques for line and also for motion estimation [21], [22]. However, in those works, only linear or close to linear motions are found, whereas our approach deals with any kind of time-varying velocity. The resulting system has several advantages:

1) Since all video frames are used simultaneously, both in space and time, it is robust to spatially and temporally local noise.
2) Unlike spatiotemporal filtering based methods [5], no prior knowledge is required about the motions taking place.
3) The CWT can be implemented efficiently using state-of-the-art transform estimation algorithms, including the Fast Fourier Transform (FFT), thus lowering its computational cost [23].
4) The CWT transform has the additional advantage of providing the 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 time-varying velocity values are obtained by processing the entire video once.

In brief, the proposed approach forms a frequency modulated (FM) signal from the frames, whose frequency varies in time in proportion to the object motion, with a method called μ-propagation, which is described analytically in Sec. III. We then apply the CWT to the FM signal, to extract its time-varying frequency and, consequently, the time-varying velocity. The resulting velocity estimate can be used in a variety of applications, such as parametric modeling of the motions in a video, and their subsequent classification. It should be noted that the proposed method is appropriate for videos with a static background, or a background that does not change significantly, as explained in Sec. III.

The paper is organized as follows. In Sec. II the basic principles of the CWT are presented. Sec. III 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. III-B and the choice of the μ-parameter is discussed in Sec. III-C. A method for comparing CWT and optical flow results is described in Sec. IV Experimental results with synthetic and real video sequences are presented in Sec. V, and conclusions are drawn in Sec. VI.

II. Continuous Wavelet Transform

The wavelet transform has received considerable attention and is used in many practical applications, as it is able to analyze signals with non-stationary spectra [23], [24], [25]. The CWT transforms the signal under investigation by convolving it with the so-called “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:

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

where * represents complex conjugation, and $\psi_{a,b}(t)$ is the mother wavelet, scaled by a factor a and dilated by b, as follows:

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

Thus, (1) becomes:

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

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

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

where $C_g = \int_0^{+\infty} \frac{|\hat{\psi}(f)|^2}{f} df < \infty$ is the admissibility constraint for the transform [23]. The peaks in $E(a)$ indicate which scales are dominant in the signal under examination.

Since in practice we are interested in extracting the frequency dependent energy spectrum of the signal $s(t)$, we extract the time-varying frequency $f(t)$ 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 [23]. The wavelet used in this work is the Morlet wavelet, defined as:

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

which is essentially a complex wave $e^{i2\pi f_c t}$ with a Gaussian envelope (e^{-t^2/f_b}). The Morlet wavelet is used because its tails taper off smoothly, ensuring a balanced tradeoff between frequency and time resolution. The central frequency of the wavelet is f_c, and f_b is the bandwidth parameter, which controls how much the wavelet "confines" the signal $s(t)$ being analyzed.

III. μ-Propagation for FM Signal Construction and Motion Estimation

The Morlet wavelet, presented in the previous section, is used to extract the time-varying velocities in a video. The method used to extract these trajectories is inspired by the subspace-based line detection algorithm (SLIDE) of [21], [22] and in particular the μ-propagation scheme employed in those works. Unlike SLIDE, however, the proposed approach is not limited to

detecting linear frequency variations and velocities, but can extract frequencies that vary with time in a general manner. In the section that follows we analytically describe how the video signal is processed in order to extract its velocities using the CWT.

A. 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. We first consider the case that each frame consists of a static background $s_b(\bar{r})$ and one moving object $s_o(\bar{r},t)$, with measurement noise $v(\bar{r},t)$. Frame 1 can then be written as:

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

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

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

where $\bar{d}(t)=[d_x(t),d_y(t)]$ is the object displacement and $v(\bar{r},t)$ represents the measurement noise and modeling errors, from frame 1 to t.

It should be noted that, as the object moves across the frame background, a different region of $s_b(\bar{r})$ is occluded and a different region is revealed by it in each frame $t,\ 1\leq t\leq N$. However, since neither the precise background area, nor the precise location of the object pixels are known, this occlusion cannot be accurately modeled a priori. Thus, it is incorporated in the noise term $v(\bar{r},t)$, which includes measurement noise and the modeling inaccuracy [6]. In practice this modeling error does not significantly affect the estimation results. If the size of the object is small in relation to the video frame and background, it only reveals or occludes a few pixels of the background as it moves. If the object covers a large area of the frame, the motion information provided by it dominates the occlusion or dis-occlusion of background pixels, which are much fewer than the object pixels.

Nevertheless, in the method developed in this work, further processing of the video takes place, which benefits from a more accurate frame model. In particular, the video frames are projected in the x and y directions. In that case, the projected background and object pixels will occlude each other, increasing the noise content of the projection signal. By removing the background using any of the well-known background removal methods [26], the projection of the

video frames in the x and y directions provides a good indication of the object motion, without interference from the background luminance values. The projection of the video sequence is necessary because it allows us to use the information present in all video frames at once, without introducing significant computational complexity. If the information from all video frames is used simultaneously, temporally local errors do not degrade the system's performance, since information loss in some frames is compensated for by data from the rest of the sequence.

An example of this procedure is shown for a video with a small object (Fig. 1(a)) with the time-varying displacement of Fig. 1(b). The illumination values of each frame are summed in the horizontal and vertical directions and the resulting 1-D signals are concatenated over time. This results in a representation of the pixels over which the object moves in the horizontal and vertical directions (Fig. 1(c), (d)), which inherently contains the signature of the object's displacement. In this example there is no motion in the vertical direction, so that projection shows the object in the same location over time.

The x and y projections are then written as follows:

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

where $v_x(x,t) = \sum_y v(x,y,t),\ v_y(y,t) = \sum_x v(x,y,t)$. In order to extract the time-varying object displacement present in Eq. (8), we employ the method of "μ-propagation" [21], [22]. This method constructs a frequency modulated (FM) signal from the original, one-dimensional one, as follows:

$$\begin{aligned} F_x(\mu,t) &= \sum_x f_x(x,t) e^{j\mu x} = S_{x,o}(\mu) e^{j\mu d_x(t)} + V_x(\mu,t), \\ F_y(\mu,t) &= \sum_y f_y(y,t) e^{j\mu y} = S_{y,o}(\mu) e^{j\mu d_y(t)} + V_y(\mu,t), \end{aligned} \tag{9}$$

where:

$$\begin{aligned} S_{x,o}(\mu) &= \sum_x s_{x,o}(x,1) e^{j\mu x}, \quad V_x(\mu,t) = \sum_x v_x(x,t) e^{j\mu x}, \\ S_{y,o}(\mu) &= \sum_y s_{y,o}(y,1) e^{j\mu y}, \quad V_y(\mu,t) = \sum_y v_y(y,t) e^{j\mu y}. \end{aligned} \tag{10}$$

The quantities $F_x(\mu,t)$ and $F_y(\mu,t)$ of Eq. (9) are FM signals, as their frequency is modulated by the time-varying motion. The phase of the FM signals is $\mu d_x(t),\ \mu d_y(t)$ respectively, so the corresponding frequencies are $\mu u_x(t),\ \mu u_y(t)$, where $u_x(t) = \frac{d(d_x(t))}{dt},\ u_y(t) = \frac{d(d_y(t))}{dt}$ are the object velocities in the x and y directions.

The CWTs of $F_x(\mu,t)$, $F_y(\mu,t)$ have the most energy along their time-varying frequencies, which is why they are often used for instantaneous frequency estimation [27], [28]. However, the time-varying frequencies are also proportional to the object velocities, so they can be used here for the extraction of motion. Consequently, in order to extract time-varying motions, we find at which frequencies the CWT energy is maximized, for each time instant t. This results in:

$$\begin{aligned} f_{x,est}(t) &= \frac{d\phi_x(t)}{dt} = \frac{d(\mu d_x(t))}{dt} = \mu\frac{d(d_x(t))}{dt} = \mu u_x(t), \\ f_{y,est}(t) &= \frac{d\phi_y(t)}{dt} = \frac{d(\mu d_y(t))}{dt} = \mu\frac{d(d_y(t))}{dt} = \mu u_y(t), \end{aligned} \tag{11}$$

from which the object velocities $u_x(t)$, $u_y(t)$ can be immediately derived, since μ is a known constant. As noted before, unlike the SLIDE algorithm [22], [21], the method proposed here is not limited to the study of linear or nearly linear motions. The frequencies extracted in Eq. (11) allow the extraction of velocities $u_x(t),\ u_y(t)$ that can vary in any manner over time.

B. Multiple moving objects

The model of the previous section is extended to the case of multiple moving objects in a video. When multiple objects are moving throughout the video, it is possible that they occlude each other over several frames, which may introduce errors in motion estimation using local methods. The proposed approach is robust to such errors, due to the simultaneous processing of the entire video. Specifically, information that may be lost in some video frames can be recovered from other frames of the sequence. Such a case is shown in the experiments of Sec. V, where the method successfully extracts the trajectories of objects that occlude each other over a few frames. The model of (7), for M objects is:

$$f(\bar{r},t) = s_b(\bar{r}) + \sum_{i=1}^{M} s_i(\bar{r} - \bar{d}_i(t), 1) + v(\bar{r},t). \tag{12}$$

The term $v(\bar{r},t)$ incorporates the effects of occlusion and dis-occlusion of background pixels, but also object pixels from the moving objects over the N frames. The method used to extract the

multiple motions is similar to that of the one object case, i.e. it is also based on the projection of the video frames in the x and y directions. Thus, the background needs to be removed to increase the accuracy of the model and avoid the occlusion of object projections by background projections. In the case of multiple moving objects there may also be object occlusion over several frames in the projection. Nevertheless, the object motions can still be extracted even in that case, since all video frames are used at once. Thus, after background removal and μ-propagation on the horizontal and vertical projections of Eq. (12), we obtain:

$$F_x(\mu,t)=\sum_{i=1}^{M} S_{x,i}(\mu)e^{j\mu d_{x,i}(t)}+V_x(\mu,t),\;\; F_y(\mu,t)=\sum_{i=1}^{M} S_{y,i}(\mu)e^{j\mu d_{y,i}(t)}+V_y(\mu,t), \tag{13}$$

where:

$$S_{x,i}(\mu)=\sum_{x} s_{x,i}(x,1)e^{j\mu x},\;\; S_{y,i}(\mu)=\sum_{y} s_{y,i}(y,1)e^{j\mu y}. \tag{14}$$

The CWT is then applied to $F_x(\mu,t), F_y(\mu,t)$, and its energy is maximized around the dominant frequencies, which are $f_{x,i}(t)=\mu u_{x,i}(t), f_{y,i}(t)=\mu u_{y,i}(t),\;\; i=1,...,M$. As for the one object case, the object velocities $u_{x,i}(t)=\frac{d(d_{x,i}(t))}{dt}, u_{y,i}(t)=\frac{d(d_{y,i}(t))}{dt}$ can be immediately extracted, since μ is a known constant.

C. 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 [29], higher values of μ lead to a higher resolution in the velocity estimation, but limit the estimation to lower frequencies and velocities. The scales $a(t)$ for the CWT are related to the actual frequencies [23] as follows:

$$f(t)=\frac{f_c}{a(t)\Delta}, \tag{15}$$

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, on which the correspondence between the estimated scales/frequencies and object velocities depends. In this work we use the Morlet wavelet, with $f_c=0.8125$ and $\Delta=1$. Then, combining Eqs. (11) and (15), we get:

$$\mu u_x(t)=\frac{0.8125}{a(t)}. \tag{16}$$

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.

IV. Comparison with Optical Flow

In order to obtain a more clear picture of the advantages of the proposed method over traditionally used spatial techniques, we compare the motion estimates extracted via the CWT with those derived using an iterative, multi-stage version of the Lukas-Kanade algorithm [30]. We choose to implement this flow algorithm, as its multi-stage approach allows the reliable estimation of both small and large displacements, at a lower computational cost [31].

Certain pre-processing steps of the flow estimation results are necessary, so as to obtain comparable results with the proposed CWT method. The CWT technique directly finds the entire vector of velocities as they vary with time, in each direction, i.e. all $u_x(t)$, $u_y(t)$, for $1 \leq t \leq N$. The optical flow method finds the flow values corresponding to each pixel $\bar{r}$ between each pair of frames. In order to find how the flow estimates vary over time, and thus compare them with the CWT results, we create the "optical flow profiles". Specifically, assuming the background has been removed, we sum the flow estimates in the x and y directions, over all frames. Since there will be non-zero flow estimates for all moving object pixels (under the assumption of rigid body motion), we examine the displacement of the center of mass of the moving pixels at each time instant. This results in "optical flow profiles" (in both directions) $of_x(t)$, $of_y(t)$, which correspond to "velocity profiles" $u_x(t)$, $u_y(t)$ estimated by the CWT.

We compare the extracted velocities from both methods with ground truth, which is available a priori for the experiments with synthetic sequences, and is measured manually for the real videos. The error metric we use to compare the accuracy of the two approaches is:

$$error_x = \left| \frac{u_x(t) - u_x^{GT}(t)}{u_x^{GT}(t)} \right|, \; error_y = \left| \frac{u_y(t) - u_y^{GT}(t)}{u_y^{GT}(t)} \right|, \tag{17}$$

where $u_x^{GT}(t), \; u_y^{GT}(t)$ are the ground truth values. The results for the experiments of Sec. V are tabulated in Table I. As seen in this table and the experiments below, the sensitivity of optical flow to local noise and local illumination changes render the CWT approach more robust.

V. EXPERIMENTS

In order to test the proposed approach, we perform 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 velocity estimation results. The videos used here can be seen on the website http://mklab.iti.gr/content/videos-motion-estimation-cwt, as referenced in [32]. Experiments are also conducted using optical flow in order to compare the CWT-based method with the spatial domain results. The CWT has a lower computational cost than the flow method in all cases, for a 3.4GHz Pentium 4 PC with 2G RAM. Although the CWT is implemented in Matlab, and also involves the construction of the FM signal, it runs in approximately the same time (or less) as the optical flow, which is implemented in C++ and applied directly to the video (i.e. there is no pre-processing). Specifically, short video sequences of about 15 seconds, such as the synthetic video sequences with 200×200 pixel frames, are processed in about 5 seconds by our method, while the optical flow approach takes 10 seconds. The longer sequences, lasting 60 sec can be processed in 20 seconds for frames of size 384×288 pixels by our approach and in 1-2 minutes by the optical flow method.

A. Parabolic displacement for two objects

In this experiment, two squares are translating towards each other with a parabolically increasing displacement [32], so their respective velocities vary linearly with time. Their displacements are $d_{x,1}(t) = 0.002t^2,\ d_{x,2}(t) = -d_{x,1}(t)$, so the velocities are $u_{x,1}(t) = 0.004t,\ u_{x,2}(t) = -0.004t$. For satisfactory resolution, we set $\mu = 2$, and scales in $[0.02, 1]$, with a step size of 0.001. The horizontal projection over time is shown in Fig. 2(a), while the vertical is not, since there is no motion in that direction. The accumulation of the optical flow estimates in the horizontal direction results in Fig. 2(b): comparing it with Fig. 2(a), we see that they correspond to the object displacements, which is expected since this is a synthetic video with no noise. 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. Fig. 2(d) shows the CWT, and Fig. 2(e) 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 succeeds to extract both linear velocities for the two objects. Both trajectories are parabolic and vary 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(d), (e) show, the CWT energy is highest along slopes with opposite signs, corresponding to linear velocities in opposite directions, which agrees with the ground truth. Fig. 2(f) shows the velocity profiles extracted from the maxima of the CWT, where, again, it is evident that both object motions have been successfully retrieved. It should be noted that after frame 250, the motion is very low and the CWT energy spreads out, leading to unreliable estimates at those frames. For this reason, the motion estimates after frame 250 are discarded, both from the CWT estimates and the optical flow profile.

B. Real sequence of a pendulum

In this experiment, the motion of a pendulum [32] is examined. Here, $\mu = 0.5$ gives good frequency tracking and resolution, and the range of scales used is $[0.02, 0.4]$, with a step size of 0.001. The optical flow profile is shown in Fig. 3(c), where it can be seen that it captures the periodic variation of the pendulum's motion, as expected, since this is a video with very constant illumination, with small inter-frame displacements, and of good quality. The horizontal projections are shown in Fig. 3(d) and the result of their FM-modulation in Fig. 3(e), where sinusoidal variations can be seen inside the projection. The CWT energy (Fig. 3(f)) varies periodically with time, and correctly captures the repetitions in the pendulum's motion. Thus, the maxima of the CWT energy are the periodic pendulum velocity is extracted from them, as explained in Sec. III. The error metric for both methods is comparable in this case (Table I). This can be explained by the fact that the video has stable luminance values, which lead to reliable optical flow estimates.

C. Real surveillance video of person walking with a parabolic displacement

A real indoors surveillance video of a person walking across a room is examined [32]. Here, we use $\mu = 0.2$ to obtain satisfactory frequency tracking and resolution, and the range of scales used is $[0.02, 0.4]$, 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. Fig. 4(c) shows the results of μ-propagation,

i.e. 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 optical flow profile for this video, in Fig. 4(d), contains a faint signature of the person's path but also very significant amounts of noise. This is due to the high illumination variations present in this video, which lead to large errors in the flow estimates. The CWT and its energy are shown in Fig. 4(e), (f): the CWT energy increases linearly with time, despite the quality of the video, which introduces large errors in the optical flow approach. The increase in CWT energy is very small because the frames are sampled densely, leading to very small velocity increases from frame to frame.

D. Real sequence "Running"

In this experiment we examine a video with a person running to the left and right [32], with intervals of no activity, as the person exits the scene for some frames. The FM modulated projection, shown in Fig. 5(b), contains a sinusoidal pattern introduced by the motion. The real part of the resulting FM signal of Fig. 5(c) is equal to zero when there is no motion. The same pattern repeats at every second non-zero interval, corresponding to motion to the left or to the right. The optical flow profile in the horizontal direction (Fig. 5(c)) captures the running trajectories as well, but with some background noise. The CWT and its energy (Fig. 5(d), (e)) successfully capture the variations of the velocity with time, as their energy is high in the frames where the person is running, and zero when there is no motion. As before, the object motion can be directly extracted from the CWT maxima. In this case, the resulting error of the CWT result in Table I is slightly less than that of the optical flow.

E. Real sequence "Jump-Run"

In this experiment we examine a video with a person jumping towards the right, followed by a video of another person running to the left [32]. Thus, there are two different kinds of motion taking place in opposite directions. The horizontal projections are shown in Fig. 6(a), where it is evident that the motions change direction. The real part of the resulting FM signal is plotted in Fig. 6(b), where a change in phase can be discerned at frame 45. The trajectories extracted by accumulating the corresponding optical flow estimates can be seen in Fig. 6(c): there is not significant background noise, but there are large errors in the flow estimates at the frame where the videos change from jumping to running. This is indicative of the sensitivity of flow-based

techniques to local variations and noise: there is a significant change in the illumination at the point where the two different videos are joined, which introduces large inaccuracies to the flow-based approach. The CWT and its energy, on the other hand, prove to be much more robust to this change (Fig. 6(d), (e)), and successfully capture the variations of the velocity with time. Their energy forms two curves, around frame 45, corresponding to the jumping and running motions, in opposite directions. Fig. 6(f) shows the maxima of the CWT energy, corresponding to the gradient of the object trajectories, which are close to the ground truth (Table I).

F. Real sequence "Jump-Run-Walk"

In this experiment we examine a sequence similar to the previous one, but with an additional segment of a person walking added to the end, i.e. a video of a person first jumping, then running and finally walking [32]. There are three different kinds of motion that take place in successively changing directions. The horizontal projections are shown in Fig. 7(a), and the real part of the resulting FM signal is plotted in Fig. 7(b), where again, changes in phase can be discerned at frames 39 and 88. The optical flow profiles extracted by accumulating the corresponding optical flow estimates can be seen in Fig. 7(c). As before, there are large errors in the flow estimates at the points where the videos change from jumping to running and from running to walking. This demonstrates, again, that flow-based techniques become very unreliable in the presence of local illumination changes. The CWT and its energy, on the other hand, prove to be much more robust to this change, as shown in Fig. 7(d), (e), and their maxima can provide the time-varying velocities with accuracy. Indeed, Fig. 7(f) shows the maxima of the CWT energy, which correspond to the gradient of the object trajectories, and give low relative errors, as seen in Table I. The last part of the motion corresponds to walking so its magnitude is lower than that of the other motions.

VI. 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 and object occlusions, by processing all video frames simultaneously. This makes it more reliable than most local spatial methods, which require high computational complexity in order to provide robust results in such cases. Contrary

TABLE I

ERROR METRIC FOR MOTION IN x DIRECTION, WITH CWT AND OPTICAL FLOW METHODS

Error/Video	Two Boxes	Pendulum	Walk	Run	Jump-Run	Jump-Run-Walk
CWT Error	0.05	0.09	0.2	0.12	0.2	0.21
Optical Flow Error	0.07	0.11	0.8	0.14	0.53	0.71

to existing, global spectral methods (e.g. FT-based), the CWT can process non-stationary spectra, and thus allows us to track time-varying object motions. Experiments with both synthetic and real video sequences lead to accurate velocity estimation, even in the presence of local occlusion. Future work involves the use of the extracted velocities in classification/recognition systems, as well as more refined processing of the video sequence, to deal with data of poor quality.

VII. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Communitys Seventh Framework Pro- gramme FP7/2007-2013 under grant agreement FP7-214306 - JUMAS, from FP6 under contract number 027685-MESH and 045547- VidiVideo.

REFERENCES

[1] J. Barron and R. Eagleson, "Recursive estimation of time-varying motion and structure parameters," *Pattern Recognition*, vol. 29, no. 5, pp. 797–818, Dec. 1996.

[2] 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.

[3] H. H. Nagel and E. Enklemann, "An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 9, p. 565593, Sep. 1986.

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

[5] D. J. Heeger, "Optical flow from spatiotemporal filters," in *Proc. IEEE 1st Int. Conf. Computer Vision*, June 1987, pp. 181–190.

[6] 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.

[7] A. Briassouli and N. Ahuja, "Extraction and analysis of multiple periodic motions in video sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 7, pp. 1244–1261, July 2007.

[8] 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*, vol. 5, April 1993, pp. 213–216.

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

[10] A. Briassouli and N. Ahuja, "Integration of frequency and space for multiple motion estimation and shape-independent object segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 5, p. 657, May.

[11] R. N. Bracewell, *The Fourier Transform and its Applications*. New York: McGraw-Hill, 1986.

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

[13] B. A. Watson and A. J. Ahumada, "Model of human visual-motion sensing," *J. Opt Soc. Amer. A*, vol. 2, p. 322342, 1985.

[14] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. i. fundamentals," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520 – 538, 1992.

[15] ——, "Estimating and interpreting the instantaneous frequency of a signal. ii. algorithms and applications," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 540 – 568, 1992.

[16] Y. Larsen and A. Hanssen, "Wavelet-polyspectra: analysis of non-stationary and non-gaussian/non-linear signals," in *Statistical Signal and Array Processing, 2000. Proceedings of the Tenth IEEE Workshop on*, Aug. 2000, pp. 539–543.

[17] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1, no. 9-12, May 1995, pp. 756 – 759.

[18] B. Angelsen, "Instantaneous frequency, mean frequency, and variance of mean frequency estimators for ultrasonic blood velocity doppler signals," *IEEE Transactions on Bio-Medical Engineering*, vol. 28, no. 11, pp. 733 – 741, Nov. 1981.

[19] S. G. Kuklin, A. Dzizinskii1, Y. Titov, and A. Temnikov, "Continuous wavelet analysis: A new method for studying nonstationary oscillations in the cardiac rhythm," *Journal Human Physiology*, vol. 32, no. 1, pp. 116–121, Jan. 2006.

[20] 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*, 2003, pp. 132 – 139.

[21] H. Aghajan and T. Kailath, "SLIDE: Subspace-based line detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 11, pp. 1057–1073, Nov. 1994.

[22] H. Aghajan, B. H. Khalaj, and T. Kailath, "Estimation of multiple 2d uniform motions by SLIDE: Subspace-based line detection," *IEEE Transactions on Image Processing*, vol. 8, no. 4, pp. 517–526, Oct. 1999.

[23] P. S. Addison, *The Illustrated Wavelet Transform Handbook*. Bristol, UK: Institute of Physics Publishing, 2002.

[24] M. Vetterli and C. Herley., *Wavelets And Filter Banks: Theory And Design*, 1992, vol. 40, no. 9.

[25] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674– 693, 1989.

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

[27] L. Angrisani, M. D'Arco, R. S. L. Moriello, and M. Vadursi, "On the use of the warblet transform for instantaneous frequency estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 54, no. 4, pp. 1374 – 1380, Aug. 2005.

[28] K. Qian, H. Seah, and A. Asundi, "Retrieving the instantaneous frequency from images by wavelet ridges," in *Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia*, vol. 2, no. 15-18, Dec 2003, pp. 835 – 839.

[29] 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.

[30] J. Bouguet, “Image sequence enhancement using multiple motions analysis,” in *Intel Corporation. Microprocessor Research Labs*.

[31] J. L.Barron, D. J. Fleet, and S. S. Beauchemin, “Systems and experiment: Performance of optical flow techniques,” *Int. J. Comput. Vis.*, vol. 12, no. 1, p. 4347, 1994.

[32]

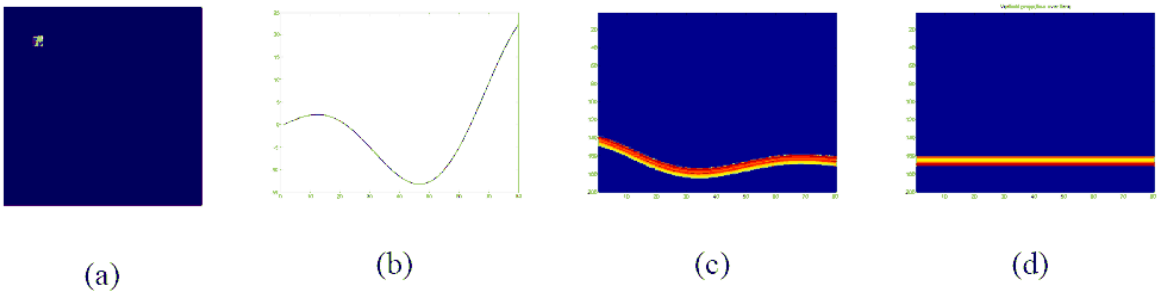

Fig. 1. (a) First frame of synthetic sequence. (b) Time-varying displacement. (c), (d) Horizontal and vertical projections of video frames over time. They contain the signature of the time-varying displacement in the horizontal direction and zero displacement for the vertical motion.

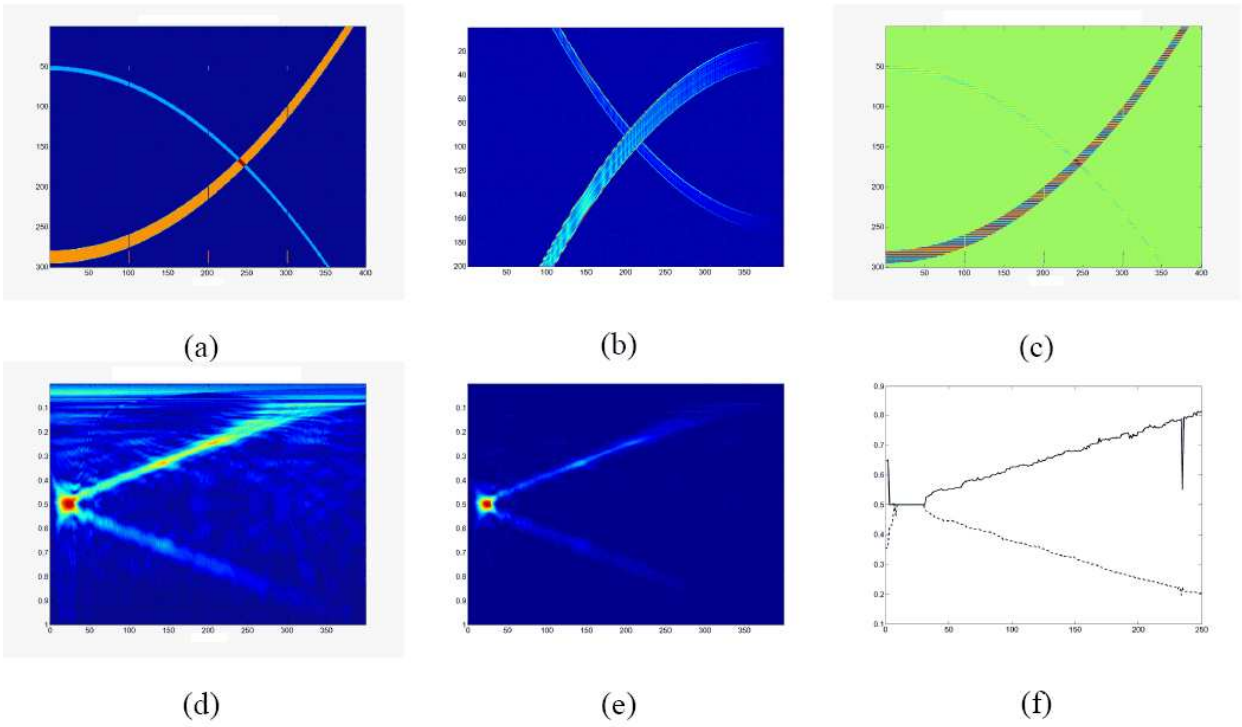

Fig. 2. (a) Projections of the synthetic sequence in the horizontal direction. The parabolic nature of the two object displacements is evident in the horizontal projection. (b) Optical flow profile: it captures the two motions but with a lot of noise. (c) Effect of FM modulation. (d), (e): CWT and its 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 $a(t) \sim 1/f(t)$. (f) CWT maxima, with the instantaneous frequencies corresponding to the object velocities.

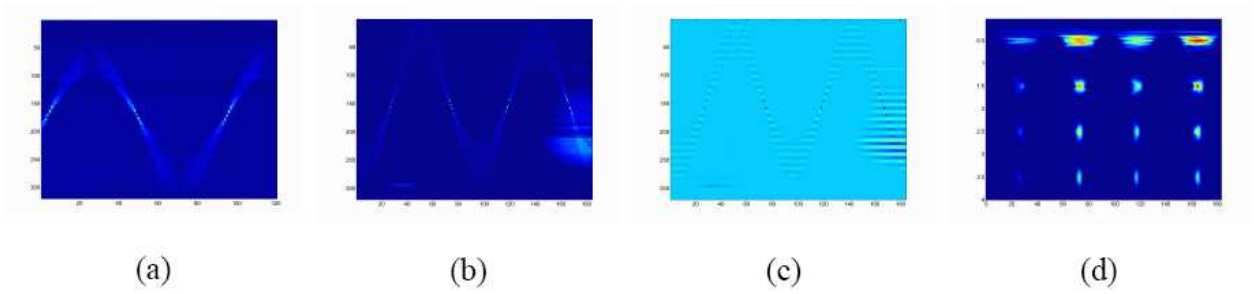

Fig. 3. (a) Optical flow profile: the periodic nature of the motion is visible but the motion information is "spread out" over time. (b) Projection of sequence in x direction. (c) Effect of μ-propagation: there is sinusoidal modulation in the projections. (d) CWT squared magnitude maxima are centered around frequencies that correspond to the pendulum's periodic motion.

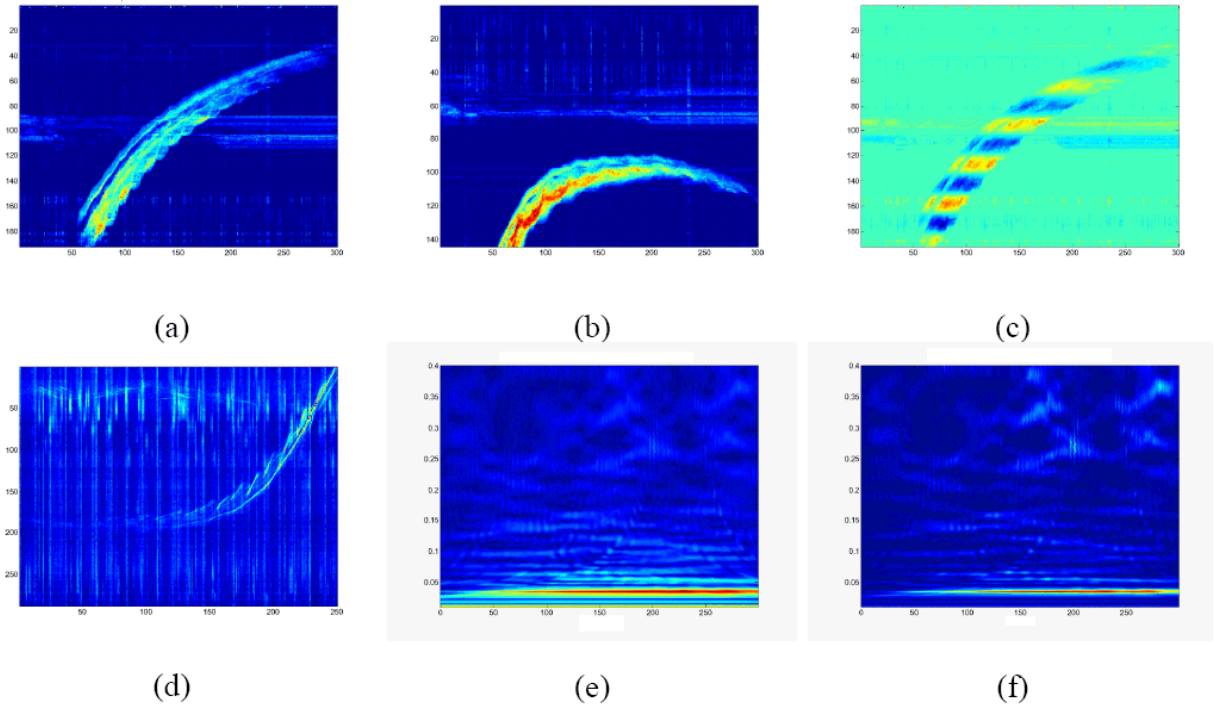

Fig. 4. (a), (b): Projections of the real video in the horizontal and vertical directions. The parabolic nature of the path can be seen. (c) Projections of the real sequence in the horizontal and vertical directions after μ-propagation. The sinusoidal modulation effect is visible. (d) Optical flow profile: it agrees with the motion, but is noisy.(e), (f): CWT for parabolic displacement i.e. linear velocity. The increase in speed is very small so the slope of the line is also very small.

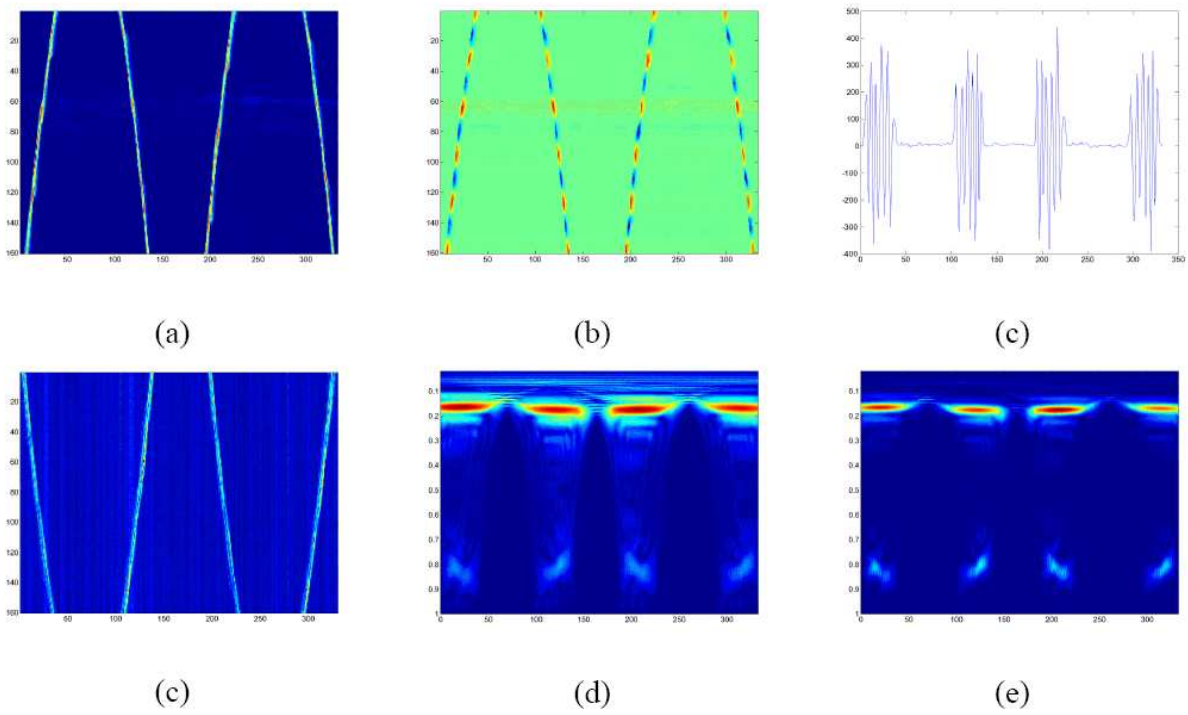

Fig. 5. (a) Projection of the real video in the horizontal direction. The changes in direction of motion and the intervals of no motion can be seen. (b) FM-modulated projection from μ-propagation. (c) Real part of FM signal. (d) Horizontal optical flow profile. (e), (f): CWT and its squared magnitude: its energy is high in the frames where the person is running, and zero elsewhere.

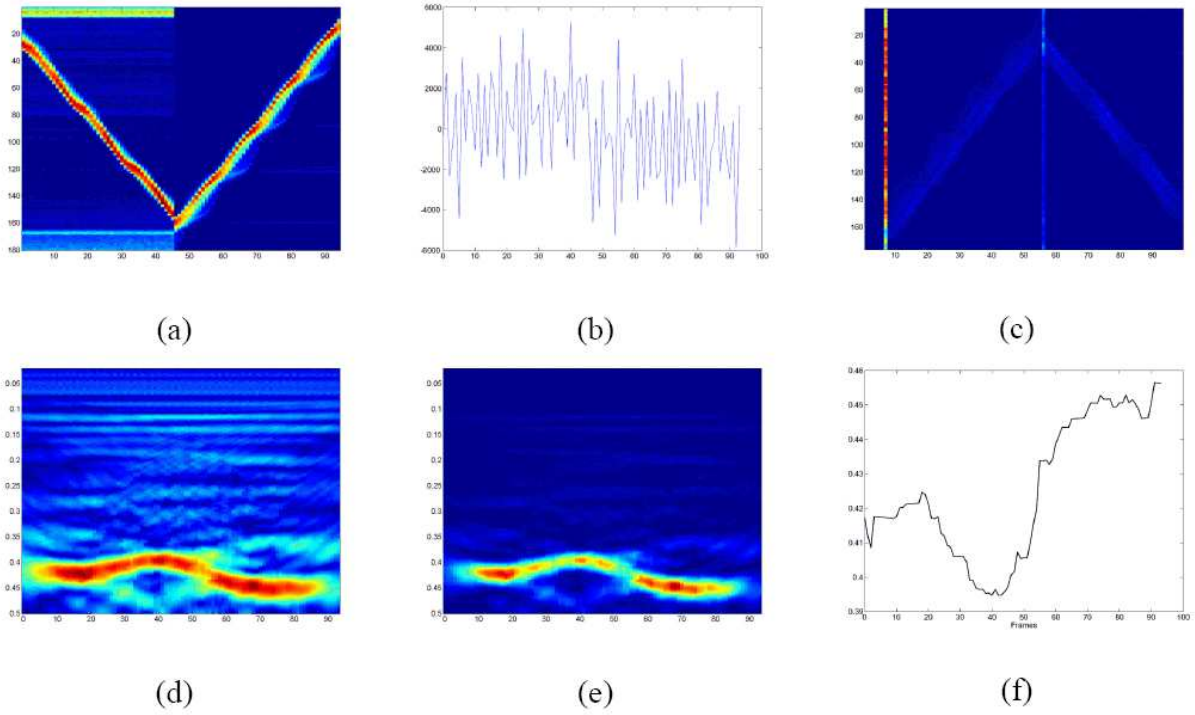

Fig. 6. (a) Projection of the real video in the horizontal direction. The different directions of the motions are evident, along with some periodicity from their repetitions. (b) FM-modulated projection after μ-propagation. (c) Displacement over time from optical flow estimates. (d), (e): CWT and its squared magnitude: its energy increases inversely proportionally with the velocities that are proportional to the instantaneous frequency, as expected, since these figures show the energy of the CWT scales $a(t) \sim 1/f(t)$. (f) Extracted trajectories for real sequence. The change in the direction is captured and small local noise variations can be dealt with at a post-processing, smoothing stage.

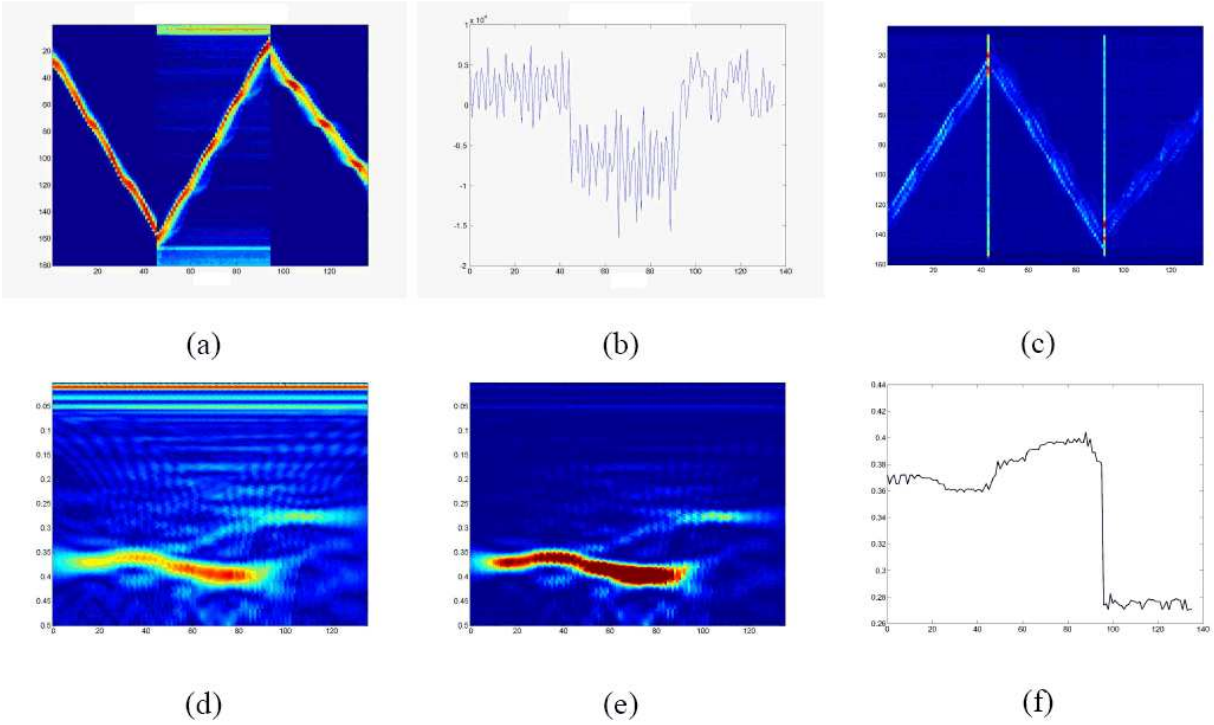

Fig. 7. (a) Projection of the real video in the horizontal direction. The different directions of the motions are evident, along with some periodicity from their repetitions. (b) FM-modulated projection after μ-propagation. (c) Displacement over time from optical flow estimates. (d), (e): CWT and its squared magnitude: its energy varies proportionally with the velocities in each video segment. (f) Extracted trajectories for real sequence. The change in the direction is captured and small local noise variations can be dealt with at a post-processing, smoothing stage.