WAVELET DOMAIN PROCESSING FOR TRAJECTORY EXTRACTION

Alexia Briassouli, Dimitra Matsiki, Ioannis Kompatsiaris

Informatics and Telematics Institute
Centre for Research and Technology Hellas

ABSTRACT

This work presents a novel approach to the extraction of trajectories in video. It has the advantage of simultaneously processing all spatiotemporal information, i.e. all the video frames, and thus overcoming disadvantages of local approaches. A projection of the video frames over time is used to create a frequency modulated signal, which is then processed with the Continuous Wavelet Transform (CWT). The CWT's power is concentrated around the prominent time-varying frequencies, which are proportional to the object trajectory. Consequently it can extract the time-varying motion trajectories, which can be used in future extensions to fully characterize the type motion taking place. This approach is robust to local measurement noise and occlusions, since it processes the available data in a global, integrated manner. Experiments take place with synthetic and real sequences demonstrate the capabilities of this approach.

***Index Terms*—** motion estimation, wavelets, video processing

1. INTRODUCTION

Numerous methods have been developed for the problem of motion estimation and trajectory extraction, with various advantages and disadvantages. Local methods find displacements between pairs of frames, based on the flow equation [1], [2], [3], on feature, or on block matching. They are based on the constant illumination assumption, and are spatiotemporally local, so they are therefore sensitive to spatially and temporally local measurement noise and occlusions. These problems have been addressed by robust flow estimation techniques, which essentially eliminate outlier flow values [4], [5].

Global processing of the data has also been used to address these limitations [6], [7]. Constant motions form energy planes in the the 3D spatiotemporal spectrum [8], which are fitted to parametric models for the extraction of the related motion features. Global methods have the inherent advantage of addressing the problem of motion estimation similarly to the way the human visual system functions, according to neurophysiological evidence [9]. Processing of the entire video provide mores accurate motion estimates than pairwise motion estimation [10], since all the available data is being used at once. However, most current literature on the frequency-based motion estimation 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 [7]. We present a method that extracts time-varying trajectories from a video by applying the wavelet transform to all video frames, without limiting the motion to be piecewise constant. This has several advantages:
(1) All video frames are used at once, so the proposed method is robust to local noise, both in space and time. The occlusion of a moving object over a few frames, for example, will only introduce a small gap in the extracted trajectory, but most of the motion information will be extracted. Illumination variations between successive frames also produce local errors, which are overcome by the use of all video frames.
(2) As opposed to existing spatiotemporal filtering based methods [6], no prior knowledge is required about the motions taking place, nor noise-sensitive and computationally intensive filtering.
(3) State-of-the-art transform estimation algorithms exist for estimating the wavelet transform (Fast Fourier Transform (FFT) based methods), which lower its computational cost [11].
(4) The wavelet transform also provides a visualization of its coefficients' magnitude, thus allowing users to observe when and which frequencies are stimulated, their duration, time evolution and their density.

In our proposed method, a frequency modulated (FM) signal is formed from projections of the video frames in the horizontal and vertical directions. The frequency of this signal varies in time proportionally to the object trajectory, by its construction with a method called "μ-propagation". The wavelet transform is then applied to the FM signal for the extraction of its time-varying frequency and, as a consequence, the time-varying trajectory.

The paper is organized as follows. Sec. 2 presents the basic principles of the CWT. In Sec. 3 the algorithm used to construct the FM signal from which the trajectories will be extracted is described. Discussion on the choice of the μ-parameter is included in Sec. 3.1. 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 wavelet transform is used in many practical applications, as it is able to analyze signals with non-stationary spectra [12], [11], [13]. In order to acquire the wavelet transforms, the input signal is convolved with the so-called "wavelet signal" (or mother wavelet). The wavelet signal can be shifted in space and scaled, and leads to higher values of the wavelet transform when the spatial shift and scaling lead to a signal that matches the input. A time-varying signal $s(t)$ has the following wavelet transform:

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

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

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

Thus, (5) becomes:

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

and for $b=0$ we have:

$$W_s(a,0) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} s(t)\psi^*\left(\frac{t}{a}\right)dt. \tag{4}$$

For each scale a, the total amount of energy contained in the signal is given by by:

$$E(a) = \frac{1}{C_g}\int_{-\infty}^{+\infty} |W_s(a,b)|^2 db, \tag{5}$$

where $C_g = \int_0^{+\infty} \frac{|\hat{\psi}(f)|^2}{f}df < \infty$ is the admissibility constraint for the transform [11], [12]. The scales that are dominant in the input signal $s(t)$ form peaks (or "ridges") in the energy $E(a)$. The frequency dependent energy spectrum of the signal $s(t)$ can then be derived by extracting the time-varying frequency $f(t)$ based on its relation to the wavelet scale a. The frequency associated with a wavelet of scale a is given by $f = f_c/a$, where f_c is the passband center of the mother wavelet [11]. In this work we use the Morlet wavelet, defined as:

$$\psi(t) = \frac{1}{\sqrt{\pi f_b}} e^{i2\pi f_c t} e^{-t^2/f_b}, \tag{6}$$

which is essentially a complex wave $e^{i2\pi f_c t}$ with a Guassian 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" the signal $s(t)$ being analyzed.

3. μ-PROPAGATION FOR FM SIGNAL CONSTRUCTION

The time-varying trajectories in a video are extracted through the application of the Morlet wavelet to an intermediate transformed version of the input video signal. This transformed version is obtained by applying a method similar to the subspace-based line detection algorithm (SLIDE) of [14], and in particular the μ-propagation scheme.

We consider that the video is formed by a time series of two-dimensional frames, with luminance $f(\bar{r},t)$ at each pixel $\bar{r} = [x,y]$, at frame t. In this work we consider that each frame consists of a static background $s_b(\bar{r},t)$ and one moving object $s_o(\bar{r},t)$ for simplicity. Future extensions can handle the case of a moving camera through by simply compensating for its motion, or by removing the background from the data using state-of-the-art background removal techniques. Then, frame 1 of an N-frame video is given by:

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

and frame $t,\ 1 \leq t \leq N$ is:

$$f(\bar{r},t) = s_b(\bar{r}) + s_o(\bar{r} - \bar{d}(t),1) + v(\bar{r},t), \tag{8}$$

where $\bar{d}(t) = [d_x(t), d_y(t)]$ is the object's displacement, and $v(\bar{r},t)$ represent measurement noise and modeling errors, from frame 1 to t.

In practice, as the object moves, it occludes different background pixels in each frame $t,\ 1 \leq t \leq N$. However, neither the precise background pixels, nor the precise location of the object pixels are known, so there is no way to model or handle this occlusion. These modeling errors are thusly incorporated in the noise term $v(\bar{r},t)$, which also includes measurement noise [7]. The modeling error is usually insignificant and does not introduce large inaccuracies to the frequency estimation, as the moving objects cover and uncover small areas of the background. Nevertheless, the accuracy and reliability of the system can be increased by removing the background with any of the well-known background removal methods [15]. This reduces the effect of the noise term $v(\bar{r},t)$ (which now only represents measurement noise), since there is no inaccuracy in the mathematical model of Eqs. (7), (8). It also allows us to project the video frames in the x and y directions with less interference from the background luminance values, as follows:

$$\begin{aligned} f_x(x,t) &= \textstyle\sum_y \Big(s_o(x - r_x(t), y - d_y(t), 1) + v(x,y,t)\Big) \quad (9)\\ &= s_{x,o}(x - d_x(t),1) + v_x(x,t).\\ f_y(y,t) &= \textstyle\sum_x \Big(s_o(x - r_x(t), y - d_y(t), 1) + v(x,y,t)\Big)\\ &= s_{y,o}(y - d_y(t),1) + v_y(y,t) \quad (10) \end{aligned}$$

The method of μ-propagation [14], [16] is applied 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:

$$F_x(\mu, t) = \sum_x f_x(x, t)e^{j\mu x} = S_{x,o}(\mu)e^{j\mu d_x(t)} + V_x(\mu, t), \tag{11}$$

where:

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

The signal $F_x(\mu, t)$ of Eq. (11) is essentially an FM signal, as it contains the time-varying displacement information $d_x(t)$ in its phase $\phi(t) = \mu r_x(t)$. The time-varying frequency of this signal can be extracted from the wavelet transform described in Sec. 2, where the signal being processed is $s(t) = F_x(\mu, t)$. It should be noted that we have presented the case of μ-propagation only for the x-projection, as the same analysis applies to the y-projection $f_y(y, t)$.

3.1. Selection of the μ parameter

An accurate representation of the time-varying frequency of the signals $F_x(\mu, t)$ and $F_y(\mu, t)$, could be obtained if the optimal value for the parameter μ were known. Higher values of μ lead to a higher resolution in the velocity estimation, but limit the magnitude of the frequencies, and consequently velocities [16], that can be found. To examine the role of μ, we focus on the horizontal projection of the video; the same principles apply for the vertical projection. The scales $a(t)$, for the CWT are related to the actual frequencies [11] as follows:

$$f(t) = \frac{f_c}{a(t)\Delta}, \tag{12}$$

where f_c is the central frequency for the wavelet used, $a(t)$ is the scale, extracted by the wavelet, and Δ is the sampling period. The frequency that we are trying to estimate, after μ-propagation, is related to the displacement and velocity as follows:

$$f_{est}(t) = \frac{d\phi(t)}{dt} = \frac{d\mu d_x(t)}{dt} = \mu\frac{d(d_x(t))}{dt} = \mu u_x(t), \tag{13}$$

where $u_x(t) = \frac{dd_x(t)}{dt}$ is the velocity in the x direction. The maximal frequency obtained is given by $f_{max} = \mu u_{max}$. Since the wavelet transform is being used, in the actual experiments a range of values for the scales a is 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. (12) and (13) become:

$$\mu u_x(t) = \frac{0.8125}{a(t)}. \tag{14}$$

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

4.1. Jogging Sequence

Experiments took place with a real video of a person jogging to the right and left, as shown in Fig. 1(a), (b). The horizontal projection (Fig. 1(c)) shows that during some frames the person is not present in the video (he disappears and reappears later, running in the opposite direction). Fig. 1(d) shows the result of the frequency modulation on the horizontal projection: the FM modulation of Eq. (9) is visible in the periodically varying colors in this figure. The real part of the FM modulated signal is shown in Fig. 1(e): the sinusoid has a high frequency when the person is running, and is constant in the intermediate frames, as expected. After applying the CWT transform and estimating its energy, we obtain the result of Fig. 1(f), where, again, there is no energy when the person is outside the frames, and high energy when the person is running.

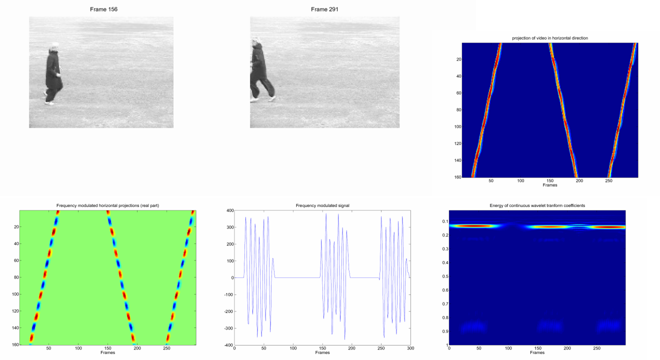

Fig. 1. Jogging Sequence. (a) Frame 156. (b) Frame 291. (c) Horizontal projection. (d) Frequency modulated horizontal projection. (e) Resulting FM signal. (f) Energy of CWT coefficients: there is a constant velocity when the person is moving and zero activity when the person is not in the frames.

4.2. Pendulum Sequence

A video of a pendulum swinging was examined here (Fig. 2(a), (b)). The horizontal projection (Fig. 2(c)) shows that the pendulum's motion has a sinusoidal form as it swings back and forth. As in Sec. 4.1, the frequency modulation on the horizontal projection leads to periodically varying colors in Fig. 2(d). The real part of the FM modulated signal in

Fig. 2(e): the sinusoid has a frequency that increases, decreases with the pendulum's swinging motion. The energy of the CWT is shown in Fig. 2(f): it follows the sinusoidal form of the pendulum's trajectory, but transforms its negative parts to positive, since it is a (positive) energy term.

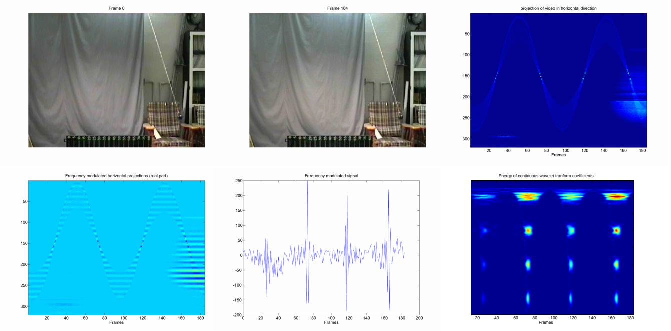

Fig. 2. Jogging Sequence. (a) Frame 156. (b) Frame 291. (c) Horizontal projection. (d) Frequency modulated horizontal projection. (e) Resulting FM signal. (f) Energy of CWT coefficients: there is a constant velocity when the person is moving and zero activity when the person is not in the frames.

5. CONCLUSIONS

The wavelet transform has been proposed as an alternative method for the extraction of motion trajectories from video. The motivation for the use of the CWT is that it allows the simultaneous processing of all video data, both in space and time, thus addressing inaccuracies of local methods. It has the additional advantage of being able to handle non-stationary signals, and thus track time-varying object motions. The algorithm creates an FM signal, modulated by the time-varying displacement, via μ-propagation. The wavelet transform of the resulting FM signal is then estimated and its energy peaks provide the desired trajectory. Experiments with real video sequences lead to accurate trajectory extraction, and demonstrate the tracking capabilities of this method. In future research, the extracted trajectories can be parameterized and employed in classification/recognition systems.

6. REFERENCES

[1] B.K.P. Horn and B.G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

[2] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of Imaging understanding workshop*, 1981, pp. 121–130.

[3] J.L. Barron and R. Eagleson, "Recursive estimation of time-varying motion and structure parameters," *Pattern Recognition*, vol. 29, no. 5, pp. 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*, May 1993, pp. 231–236.

[5] J. Weijer and T. Gevers, "Robust optical flow from photometric invariants," in *Proc. IEEE International Conference on Image Processing, 2004*, Oct. 2004, vol. 3, pp. 1835–1838.

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

[7] 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.

[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*, April 1993, vol. 5, pp. 213–216.

[9] E. H. Adelson and H. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt Soc. Amer. A*, vol. 2, pp. 284299, 1985.

[10] 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, pp. 4347, 1994.

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

[12] 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.

[13] M. Vetterli and C. Herley., *Wavelets And Filter Banks: Theory And Design*, vol. 40, 1992.

[14] H.K. 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.

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

[16] I. Djurovic and S. Stankovic, "Estimation of time-varying velocities of moving objects by time-frequency representations," *IEEE Transactions on Image Processing*, vol. 12, no. 5, pp. 550–562, May 2003.