Phase-Based Video Motion Processing

Based on SIGGRAPH 2013 "Phase based video motion processing" paper by Wadhwa et al.

Sidan Qi, Huai Wu, Ruicheng Xian, Wanxian Yang

Department of Computer Science, University of Illinois Urbana–Champaign

Figure: Results

Motivation and Overview

Many phenomena in real-life occurs at minuscule scales that are not perceptible with bare eyes, for instance, the swinging of skyscrapers due to winds, the pulsing change in skin color due to blood flow, or the wiggling of another person's eyes. The goal of video motion processing is to magnify such barely noticeable motions that are present in recorded videos.

Motion and Phase. The motivation for phase-based video motion processing comes from the Fourier shift theorem. A 1D image of unit length, denoted by $f \in \mathbb{R}^{[0,1]}$, can be written as a trigonometric polynomial using the definition of Fourier series. With this representation, the shift theorem states that

$$f(x+\delta(t)) = \sum_{k=-\infty}^{\infty} c_k \exp\left(i2\pi k(x+\delta(t))\right),$$

meaning that a translation on the image by distance $\delta(t)$ over time appears as a phase shift on the trigonometric polynomial.

By taking its phase component and use a temporal filter with zero DC response to extract the changing phase that corresponds to the translation, i.e. $\Delta\phi(t) = 2\pi k\delta(t)$,, we can directly magnify or attenuate this translation by increasing or decreasing the change in phase via multiplication, namely

$$f'(x+\delta(t)) = \sum_{k=-\infty}^{\infty} c_k \exp\left(i2\pi k(x+\delta(t))\right) \cdot \exp\left(i\alpha\Delta\phi(t)\right),$$

for some magnification factor α.

Usually, motions are localized such that we have to deal with $\delta(x,t)$, and they require multiple frequency bands to represent, meaning that by indexing the motions by j, we need to treat the phases over frequency bands $k \in A_j$ as a whole. Then we write

$$f(x+\delta(x,t)) = \sum_j \left(\sum_{k\in A_j} c_k \exp\left(i2\pi kx\right) \right) \cdot \exp\left(i2\pi \sum_{k\in A_j} \delta_j(x,t,k) \right),$$

and process each summand as a whose with temporal filters of appropriate pass-bands to isolate the motion of interest.

Complex Steerable Pyramid. Without further assumption, a simple method to partition the 2D frequency spectrum for motion extraction is the complex steerable pyramid. Ideally, the complex pyramid divides the 2D frequency domain into the following regions.

Each partition captures objects of a certain size and orientation, hence motions of a certain scale and direction. The aforementioned processing is performed on the signal contained in each partition.

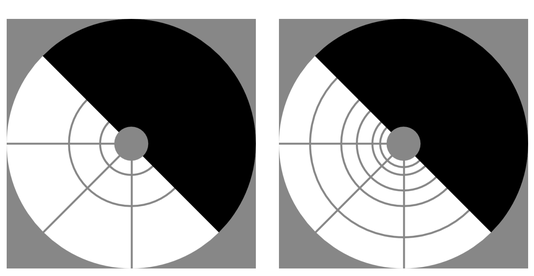

Figure: Ideal frequency response of the filters in a complex steerable pyramid of 3 levels and 4 orientations, meaning each of the 12 filters is an indicator function on a white partition. The right figure has suboctave bands.

The corners comprise the high-pass residual, and the center part is the low-pass residual. The black conjugate symmetric portion is discarded.

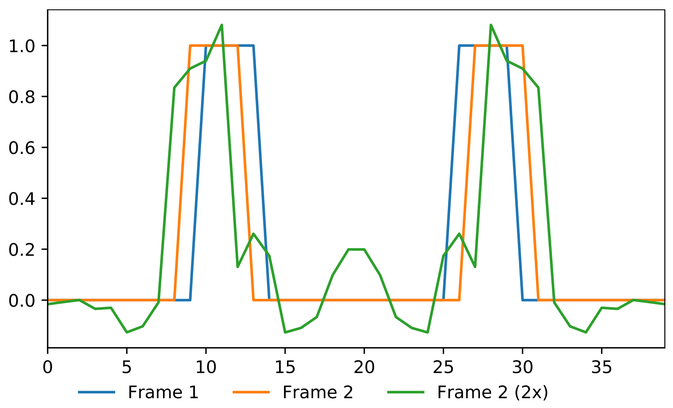

Figure: Toy example. In frame 2, the two boxes moved away from each other relative to frame 1. The green curve is the result of motion magnification by 2 using difference of phase and a pyramid with 3 levels.

Algorithm (Motion Modification).

Input:

- $I_{1:T}$ is a real-valued video sequence.
- D, K, N represents the depth, number of orientations and number of filters per octave of the complex steerable pyramid.
- f_s represents the sampling rate of the video sequence, and f_l, f_h represents the frequency range of the motion to be modified.
- α is the magnification factor.
- B, F represent the types of filters to be used to construct the pyramid and for temporal filtering, respectively.

Initialize:

- $P_{1:T}$, $R_{H1:T}$, $R_{L1:T}$ represent the pyramid sequence.
- $Q_{1:T}$ represents the pyramid for storing motion magnified frames.
- $J_{1:T}$ is the output video sequence.
- Filter F with f_s, f_h, f_l.

For $t = 1, \cdots, T$, set $(P_t, R_{Ht}, R_{Lt}) \leftarrow \text{PyramidAnalysis}(I_t, D, K, N, B)$, obtaining the complex steerable pyramid representation of each frame.

For $d, n, k = 1$ to D, N, K respectively:

- Collect $X_{1:T} := (P_1[d,n,k], P_2[d,n,k], \cdots, P_T[d,n,k])$, the evolution of I at scale (D, N) and direction K.
- Set $\Phi_{1:T}$ to be the phase component of $X_{1:T}$.
- Set $\Delta\Phi_{1:T} \leftarrow \text{TemporalFiltering}(\Phi_{1:T}, F)$, which is to perform temporal filtering on the phase.
- For $t = 1, \cdots, T$, set $Q_t[d,n,k] \leftarrow P_t[d,n,k] \circ \exp(i(\alpha-1)\Delta\Phi_t)$, which is to modify motion.

For $t = 1, \cdots, T$, set $J_t \leftarrow \text{PyramidSynthesis}(Q_t, R_{Ht}, R_{Lt}, D, K, N, B)$, obtaining the motion magnified video sequence.

Return $J_{1:T}$.

References

[1] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.

[2] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, 32(4):80, 2013.