

Joint Spatial and Frequency Domain Motion Analysis

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

Abstract

Traditionally, motion estimation and segmentation have been performed mostly in the spatial domain, i.e., using the luminance information in the video sequence. Frequency domain representation offers an alternative, rich source of motion information, which has been used to a very limited extent in the past, and on relatively simple problems such as image registration. We review our work during the last few years on an approach to video motion analysis that combines spatial and Fourier domain information. We review our methods for (1) basic (translation and rotation) motion estimation and segmentation, for multiple moving objects, with constant as well as time varying velocities; and (2) more complicated motions, such as periodic motion, and periodic motion superposed on translation. The joint space analysis leads to more compact and computationally efficient solutions than existing techniques.

1. Introduction

This paper presents an overview of our work in the last few years on motion analysis. The unifying theme of this work is simultaneous use of both spatial domain and frequency domain representations of a given video sequence. Our presentation is in two major parts. First, we review our approach to estimation of the two basic components of general image motion - translation and rotation. Second, we summarize our work on the estimation of specialized, more complex motions, illustrated here by the case of periodic motion. In both cases, we allow multiple objects under motion, segment each object out and estimate the parameters of each motion.

Motivation: Most work on video motion analysis has been carried out directly in the spatial domain. However,

the spatial locality of the representation has the disadvantages of lack of sensitivity to spatially diffuse phenomena such as gradual illumination changes. Some of these problems can be handled by using alternative representations. For example, a frequency domain representation, e.g., using the Fourier Transform, or Time-Frequency Distributions (TFDs) [6], [3], has strengths that are complementary to those of the spatial domain representations. Some advantages of the frequency approach include: (1) It is robust to global illumination changes. (2) Inaccuracies in motion estimation near object boundaries are avoided, since the estimates are not based on spatially local inter-frame luminance differences, rather on global intensity distribution. (3) Efficient algorithms are available for FT computation. (4) There is no need for the commonly used separate stages of feature detection and matching as in feature based spatial methods [11]. Some work has been done to take advantages of the frequency based representation. However, frequency domain methods have their own limitations. The main difficulty they encounter is the so-called "localization problem", which is a direct result of their global nature. Specifically, they are able to detect and estimate motions, but they do not directly relate each estimate to the associated frame pixels, thus not providing object localization or motion segmentation.

Our work has focused on joint use of both spatial and frequency representations. The advantages of one help overcome the shortcomings of the other, and lead to more versatile methods. We limit our overview to the case of rigid objects being imaged from a still camera. Sec. 2 first presents the case of two-frame translational and rotational motion, which is equivalent to the case of multiframe motion with constant velocity. Sec. 3 considers the problem of general, multiframe motion, with time varying velocities. Sec. 4 considers the problem of periodic motion analysis. Sec. 5 presents concluding remarks.

2. Constant Velocity Motion

In this case, the Fourier transform is initially used to estimate translational or roto-translational motions between pairs of frames. Motion segmentation follows, using both Fourier and spatial data.

2.1. Translation Estimation

Let each frame consist of M moving objects i , $1 \leq i \leq M$, with luminance $s_i(\bar{r})$ at pixel \bar{r} and displacement $\bar{b}_i(n)$ at frame n . The FT of object i is $S_i(\bar{\omega}) = M_i(\bar{\omega})e^{j\Phi_i(\bar{\omega})}$, where $\bar{\omega} = [2\pi m/N_1, 2\pi n/N_2]^T$, $m, n \in \mathbb{Z}$, is the 2-D frequency, $N_1 \times N_2$ the image size, $M_i(\bar{\omega})$ the FT magnitude, and $\Phi_i(\bar{\omega})$ the FT phase. The FT of frame k is $A(\bar{\omega}, k)$, $1 \leq k \leq N$, the measurement noise is $V_{noise,k}(\bar{\omega})$, and the background area hidden by the moving objects is denoted by $V_{bck,k}(\bar{\omega})$. Then the FT of frame k is $A(\bar{\omega}, k) = S_b(\bar{\omega}) + \sum_{i=1}^M S_i(\bar{\omega})e^{-j\bar{\omega}^T \bar{b}_i(k)} - V_{bck,k}(\bar{\omega}) + V_{noise,k}(\bar{\omega})$. Stacking the FT's of the N frames, we get the over-determined system

$$A = HS + V_{noise} - V_{bck}, \quad (1)$$

where A is an $N \times 1$ data vector, with the values of each frame's FT, and H is an $N \times (M+1)$ matrix, containing the motion information, with row k $H_k(\bar{\omega}) = [1, e^{-j\bar{\omega}^T \bar{r}_1(k)}, \dots, e^{-j\bar{\omega}^T \bar{r}_M(k)}]$. V_{noise} is the $N \times 1$ vector of measurement noise, and $V_{bck,k}$ represents the occluded background areas.

2.2. Phase Correlation for Multiple Motions

For the translation between frames 1 and k , we consider the ratio $\Phi_{1,k}$ of their FTs:

$$\Phi_{1,k}(\bar{\omega}) = \frac{A(\bar{\omega}, k)}{A(\bar{\omega}, 1)} = \gamma_b(\bar{\omega}) + \sum_{i=1}^M \gamma_i(\bar{\omega})e^{-j\bar{\omega}^T \bar{b}_i(k)} + n_k(\bar{\omega}). \quad (2)$$

$\gamma_b(\bar{\omega}) = S_b(\bar{\omega})/(S_b(\bar{\omega}) + \sum_{i=1}^M S_i(\bar{\omega}) + V(\bar{\omega}, 1))$, $\gamma_i(\bar{\omega}) = S_i(\bar{\omega})/S_b(\bar{\omega}) + \sum_{i=1}^M S_i(\bar{\omega}) + V(\bar{\omega}, 1)$, and $n_k(\bar{\omega}) = (V_{noise,k}(\bar{\omega}) - V_{bck,k}(\bar{\omega}))/(S_b(\bar{\omega}) + \sum_{i=1}^M S_i(\bar{\omega}) + V(\bar{\omega}, k))$, $V(\bar{\omega}, k) = V_{noise,k}(\bar{\omega}) - V_{bck,k}(\bar{\omega})$. The inverse FT $\phi_{1,k}(\bar{r})$ is a weighted sum of delta functions, whose peaks give the M displacements between frames 1 and k .

$$\phi_{1,k}(\bar{r}) = a_b(\bar{r}) + \sum_{i=1}^M a_i(\bar{r})\delta(\bar{b} - \bar{b}_i(k)) + n_k(\bar{r}). \quad (3)$$

This yields the 2D trajectory $\bar{b}_i(k)$ of each object. Displacements between frames 1 and k can be estimated by simply dividing the motion estimates \bar{b}_i by $k-1$.

2.3. Roto-Translational Motion Estimation

Motivated by FFT-based techniques for image registration [9], we have developed a method for the estimation of multiple rotations and translations in a video sequence. Consider an image $f(x, y)$ and its rotated and translated version

$$f_r(x, y) = f((x-x_0)\cos(\theta_0) + (y-y_0)\sin(\theta_0), \\ - (x-x_0)\sin(\theta_0) + (y-y_0)\cos(\theta_0)). \quad (4)$$

The FT of the rotated image is given by

$$F_r(\omega_x, \omega_y) = F_1(\omega_x \cos(\theta_0) + \omega_y \sin(\theta_0), \\ -\omega_x \sin(\theta_0) + \omega_y \cos(\theta_0))e^{-j(\omega_x x_0 + \omega_y y_0)}. \quad (5)$$

In the log-polar domain, the magnitudes of the two Fourier transforms are $M_r(\rho, \theta) = M(\rho, \theta - \theta_0)$, where $\rho = \log \sqrt{\omega_x^2 + \omega_y^2}$ and $\theta = \tan^{-1}(\frac{\omega_y}{\omega_x})$, so a rotation is represented by a translation in the frequency domain, in log-polar coordinates. This indicates that rotational motion can also be extracted from frequency domain data in a similar manner to translational motion [1]. Therefore, for brevity, whenever possible we will avoid a separate discussion of the rotational component in the rest of the paper.

2.4. Initial Segmentation in Frequency Domain

Once the translational (and rotational) motions have been estimated, the linear system of Eq. 1 can be solved in a Least Squares sense to estimate object appearance and segmentation in all frames. Since this is an inverse problem, it is often ill-posed, making regularization necessary for reliable solution [2], [10]. As the experiments described later in this section show, the least squares estimates retain significant information from the data, such as the shape and texture of the objects. They also separate the background from the objects and the multiple objects amongst themselves.

2.5. Final Segmentation by Fusion of Spatial and Frequency Domains

The LS object estimates presented above differ from their true values due to the occlusions of the background ($V_{bck,k}(\bar{\omega})$ terms, unaccounted for in the LS estimates), and the approximation error introduced by the regularization. To overcome these errors, we fuse the Fourier domain object estimates with the spatial data. The fusion is performed by comparing the least squares object estimates with the original video frames, and refining the object and background areas accordingly.

Experiments Below we present representative results of the motion estimation and segmentation algorithm presented above.

1. Parking Lot Sequence: Experiments are conducted with a sequence of a car translating in a parking lot (Fig. 1(a)). Its translation is correctly estimated via Eq. (2) and the LS segmentation successfully separates the background from the car (Figs. 1(b)-(c)). Fusion with spatial information leads to the final car segmentation in Fig. 1(d).

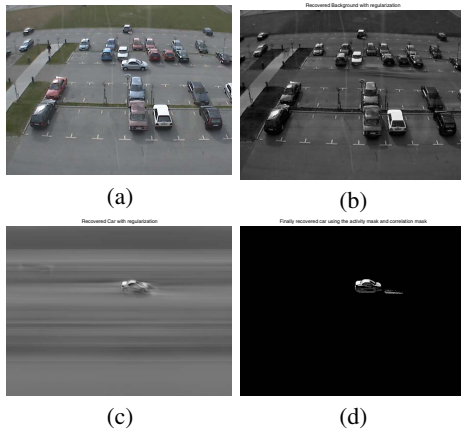


Figure 1. Parking Lot Sequence: (a) Original Frame (b) LS recovered background. (c) LS recovered car. (d) Recovered car after spatial fusion.

2. Traffic Sequence: Experiments are conducted with a real traffic sequence, consisting of two cars that are turning (Fig. 2(a)). The angles of rotation for each car are estimated between successive frames and compared against the ground truth, which is obtained through manual feature point tracking. The angles are estimated with errors up to 0.05% of the range of rotation values. The estimated translations in the horizontal and vertical directions are also close to their true values with errors up to 0.075% of the range of translational values. Finally, as Fig. 2 shows, the bottom right and top right cars are accurately recovered.

3. Time-Varying Velocity Motion

A straight forward approach to estimation of time-varying velocities that we have reported in [1] is to recursively divide the video sequence into subsequences until the total displacements across the subsequences becomes linear in each number of frames, suggesting that the motion in the subsequence can be estimated as having a constant velocity.

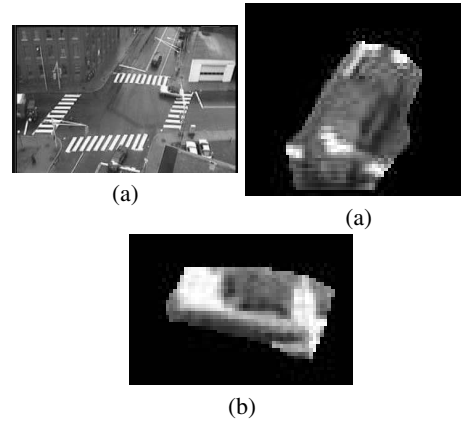


Figure 2. Traffic sequence: (a) Original frame. (b) Reconstructed bottom right car. (c) Reconstructed top right car.

We will omit the details of this simple method here. Below we present an alternative approach, in which we extract time-varying velocity profiles using Time-frequency distributions (TFDs) analysis of the non-stationary spectral content of the video sequence. TFDs are used as they allow simultaneous estimation of the entire trajectories. Fusion with spatial information is needed for separation of the multiple time-varying motion trajectories into differently moving objects, and motion based segmentation.

3.1. Short Term Fourier Transform

The time-varying spectrum content of non-stationary signals can be captured by TFDs, which are essentially the FT, applied over small, windowed, time segments of a signal. The most common and simplest TFD is the Short-Term Fourier Transform (STFT). By filtering the signal with an appropriate low-pass function around time t , we obtain an approximate representation of the signal's spectral content at that instant. For a one-dimensional signal $s(t)$, in discrete time, the STFT is defined as

$$STFT_s(t, \omega; h) \equiv \sum_{-\infty}^{+\infty} s(\tau + t)h^*(\tau)e^{-j\omega\tau}, \quad (6)$$

which means it is essentially the FT of a windowed signal. The window function controls the relative weight imposed on different parts of the signal, leading to an inherent trade-off between time and frequency resolutions. When $h(t)$ has higher values near the center of the interval (the observation point t), the STFT emphasizes local quantities. A window that is compact in time leads to higher time resolution,

whereas a window peaked in the frequency domain gives better frequency resolution.

3.2. Velocity Estimation

Consider a scene containing M objects, with FTs $S_i(\bar{\omega})$, $1 \leq i \leq M$, moving independently against a background which has been detected (e.g. by median filtering over time) and eliminated (replaced by 0) in the images. Frame k is $A(\bar{\omega}, k) = \sum_{i=1}^M S_i(\bar{\omega}, 1)e^{-j\bar{\omega}^T \bar{b}_i(k)}$. A slice of $A(\bar{\omega}, k)$ at $(\omega_x, 0, k)$ has a time-varying spectral content, consisting of M time-varying frequencies at each time instant k :

$$\bar{\Omega}_x = \omega_x \bar{B}_x = \omega_x \begin{pmatrix} b_x^1(1), & b_x^1(2), & \dots, & b_x^1(N) \\ \vdots & \vdots & \ddots & \vdots \\ b_x^M(1), & b_x^M(2), & \dots, & b_x^M(N) \end{pmatrix}. \quad (7)$$

Thus, at each time instant $1 \leq k \leq N$, the M translations can be estimated from the corresponding time-varying frequencies. By applying the STFT to $A(\bar{\omega}, k)$, we can extract the time-varying power spectrum of this signal, which is a 2D function of frequency and time, with its power concentrated along ridges in the time-frequency plane. The TFD maxima for the signal A give the time-varying frequency content $\bar{\omega}_x(k)$ of A . Since $\bar{\omega}_x(k)$ is directly proportional to the time-varying object displacement, Eq. (7) enables the estimation of the horizontal displacement ($b_x(1) = 0$) $\bar{b}_x(t) = [0, b_x(2), \dots, b_x(N)]$. By setting $\omega_x = 0$, the same method can be used to estimate the vertical displacements $b_y(k)$ from $A(0, \omega_y, k)$.

Experiments We present results on a video sequence of people walking. We demonstrate how our method can be used to estimate multiple motions with the number of moving objects changing with time: There are initially three people in the scene, but one of them exits the scene near the middle of the video (Fig. 3). Two of the people are walking together, so they are considered as one moving entity, whereas the third person (who eventually leaves) is the second moving entity. The STFT indeed has two ridges for the first part of the sequence, and only one ridge for the rest of it, so its maxima trace two curves with different lengths (Fig. 4(a)): the first corresponds to the two people walking to the left, throughout the entire sequence, and the second to the person who is walking to the right, and exits after several frames.

The successful joint use of frequency and spatial domain information for motion estimation and segmentation motivates the use of both domains for more complicated or specialized types of motions. Periodic motion is one such class of motions which are common in nature and often associated with biological movements, e.g., walking, speaking and gesturing.

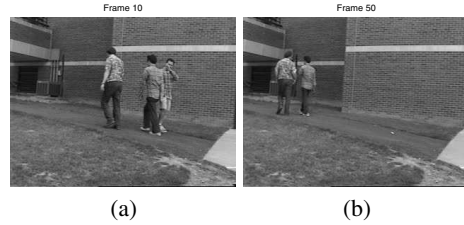


Figure 3. A sequence with the number of moving objects changing: (a) Frame 10. (b) Frame 50.

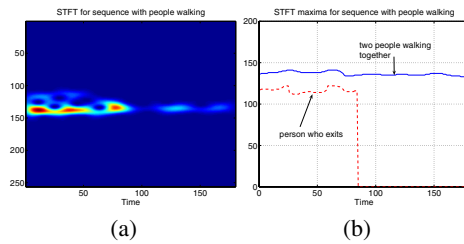


Figure 4. Sequence with a changing number of people: (a) STFT. (b) The maxima of the STFT extract the two trajectories simultaneously.

4. Periodic Motion

There has been extensive work on estimation of periodic motions, but the existing research is limited to simple cases, involving a single, pure periodic motion [8]. Also, current methods involve elaborate feature or region correspondences [5] and manual intervention. We take advantage of the frequency space signature of multiple periodic motions to extract them, using TFDs, which leads to a simple formulation. Unlike any of the previous methods, we extract multiple periodic trajectories simultaneously. Our approach is also robust to deviations from strict periodicity, which is necessary in many real life sequences. Below we first present our work on pure periodic motions, and then on periodic motion superposed on translation.

4.1. Pure Periodic Motion

Consider frame $a(x, y, n)$. We construct a frequency modulated (FM) signal, whose 2D frequency is modulated by the time-varying displacements of the objects, via the technique of constant μ propagation [6]. Essentially, we estimate the 2D FT at a constant 2D “spatial frequency”

$\bar{\mu} = [\mu_1, \mu_2]$, as follows:

$$A(\mu_1, \mu_2, k) = \sum_{\bar{r}} \sum_{i=1}^M [s_i(\bar{r} - \bar{b}_i(k)) + v_{noise}(\bar{r}, k)] e^{j(\mu_1 x + \mu_2 y)}$$

$$= \sum_{i=1}^M S_i(\mu_1, \mu_2) e^{j(\mu_1 \bar{b}_i^x(k) + \mu_2 \bar{b}_i^y(k))} + V_{noise}(\mu_1, \mu_2, k)$$

The frequencies $\omega_i(k) = \mu_1 \bar{b}_i^x(k) + \mu_2 \bar{b}_i^y(k)$ in $A(\mu_1, \mu_2, k)$ are extracted by applying the STFT (Sec. 3.1). The motion appears in each $\omega_i(k)$ as a weighted sum of the horizontal and vertical displacements, but this can be overcome by estimating $A(\mu_1, \mu_2, k)$ at $\mu_2 = 0$ and $\mu_1 = 0$. This gives $\omega_i(k) = \mu_1 \bar{b}_i^x(k)$ and $\omega_i(k) = \mu_2 \bar{b}_i^y(k)$ respectively, so the horizontal and vertical displacements are separated.

Thus, the dominant frequencies at each time k are extracted, leading to a multicomponent signal, consisting of the M periodically varying frequencies $\omega_i(k)$. The periodic nature of the motions allows their separation in an efficient manner. For object i , we get the periodic signal $\bar{b}_i^x = [b_i^x(1), \dots, b_i^x(N)]$, which represents its motion over time. We sum the M signals \bar{b}_i^x of all objects i at each instant k , to form $\bar{g}_x = [g^x(1), \dots, g^x(N)] = \sum_{i=1}^M \bar{b}_i^x$, with values at each frame k ($1 \leq k \leq N$) given by $g^x(k) = \sum_{i=1}^M b_i^x(k)$. The resulting 1D function \bar{g}_x is a sum of periodic functions \bar{b}_i^x , with different periods T_i^x ($1 \leq i \leq M$). By applying traditional spectral analysis methods [7] to \bar{g}_x , we can obtain its M frequencies ω_i^x ($1 \leq i \leq M$), and the corresponding periods $T_i^x = 1/\omega_i^x$.

Segmentation Once the different periods are estimated in the sequence, the moving objects can also be extracted using this information. By correlating frames that are separated by an integer number of periods, we expect to get higher correlation values in the area of the periodically moving object. Since the motions are periodic, we have $b_i^x(k) = b_i^x(k + T_i^x)$, $b_i^y(k) = b_i^y(k + T_i^y)$ for object i . We consider $T_i^x = T_i^y = T_i$ for simplicity, but the same analysis can be applied when $T_i^x \neq T_i^y$. Therefore, we can extract the j_{th} object by correlating frames k and $k' = k + T_j$: since only that object is expected to re-appear in the same position in those frames, the correlation values will be highest in the pixels in its area.

4.2. Periodic Motion Superposed on Translation

The formulation we have used allows easy estimation of periodic motions superposed on translations. An example is walking, where the legs move periodically but the moving entity is also translating. Correlation-based methods cannot deal with such motions, because of the shifting position of

the periodically moving object. The time-varying trajectory $b(k)$, which is used to create the FM signal, is of the form $b(k) = \alpha \cdot k + b_P(k)$, where $1 \leq k \leq N$, α is a constant and $b_P(k)$ is the periodic component of the motion. The FM signal we create via μ -propagation is: $z(k) = e^{j\mu(\alpha \cdot k + b_P(k))}$, whose phase is $\phi_z(k) = \mu(\alpha \cdot k + b_P(k))$. The TFDs estimate its frequency, i.e. the time-derivative of $\phi_z(k)$, given by:

$$\omega_z(k) = \frac{\partial(j\mu(\alpha \cdot k + b_P(k)))}{\partial k} = j\mu\alpha + \frac{\partial b_P(k)}{\partial k}. \quad (8)$$

Consequently, the translational component of the motion becomes a simple additive term, whereas the periodicity of $b_P(k)$ is retained in the extracted frequency. This eliminates the need to align the video frames as is essential in traditional methods.

Segmentation Segmentation of periodically moving objects cannot be performed directly in terms of the periodic motion parameters, since the object has also translated. This difficulty can be easily overcome by estimating the “mean” translation between frames, via their FT [1], [4]. If there are M objects in the sequence, where object i is displaced by $\bar{b}_i(n)$ from frame 1 to k , the “mean” translation can be estimated by Eq. (3). Thus, the peaks of $\phi_{1,k}(\bar{r})$ estimate the “mean” translations $\bar{b}_i(k)$ of object centroids, between frames 1 and k (e.g. a walking person’s body but not the periodically moving legs and arms).

Experiments We now present the results of our method on two sequences.

1. Walking Sequence: This sequence contains a pure periodic motion superposed on translation. We estimate the ground truth periods of the arms and legs to be 5 from observation. Our algorithms extract these parameters correctly using STFT. Fig. 6 shows the results only for the horizontal motion of legs, as the arm motion results are almost identical. The mean translation is then estimated to be 135 pixels via Eq. (2), and the image is shifted back to a common location in all frames in order to extract the periodically moving legs.



Figure 5. Walking Sequence: (a) Frame 12. (b) Frame 60.

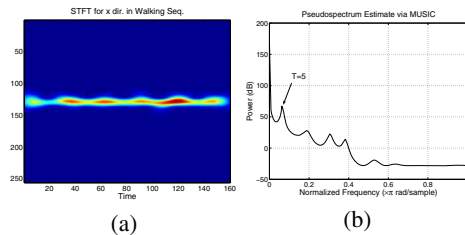


Figure 6. Walking Sequence: (a) 2D STFT for the horizontal (leg) motion ($\mu_2 = 0$). (b) The power spectrum for the horizontal direction correctly finds $T = 5$.

2. Swings Sequence: This sequence shows two children on swings (Fig. 7(a)), moving with the same period, $T = 2.5$, but different phase, as they start off from different positions. The period estimate $T = 2.875$ is quite close to its observed value of $T = 2.5$, and allows the successful segmentation of the children (Fig. 7(b), (c)). It should be noted that the method succeeds despite the fact that the children bodies change non-rigidly during the motion (e.g. legs folding or extending). The results of the segmentation are shown by drawing the circumscribing window which contains them in all frames, instead of a tight segmentation.

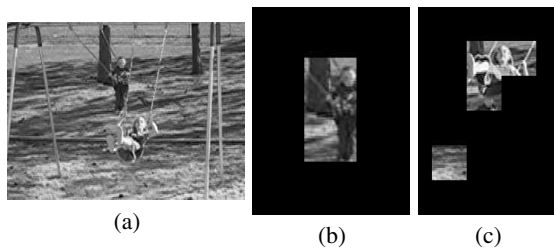


Figure 7. Swings sequence: (a) Frame 10. Segmentation results for (b) boy. (c) girl.

5. Conclusions

Our recent approach to motion analysis has been presented. Motion estimation is achieved in the frequency domain, and the segmentation of the sequence is based on both frequency and spatial data. The proposed approach avoids problems of spatial methods, such as sensitivity to global illumination changes, problems at moving object boundaries, and high computational cost. The joint space analysis can also be used for efficient and robust estimation of multi-

ple time-varying motions, as well as more specialized motions, such as periodic ones, via the use of time-frequency distributions. Our approach overcomes many limitations of existing methods, such as sensitivity to noise or local illumination changes, because of the global nature of the time-frequency domain processing and the joint use of spatial and frequency information. Future directions of research include extending this method to the problem of motion estimation and segmentation with a moving camera, and to non-rigid objects.¹

References

- [1] A. Briassouli and N. Ahuja. Fusion of frequency and spatial domain information for motion analysis. In *ICPR 2004, Proceedings of the 17th International Conference on Pattern Recognition*, volume 2, pages 175–178, Aug. 2004.
- [2] A. Briassouli and N. Ahuja. Integrated spatial and frequency domain motion segmentation and estimation. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, volume 1, pages 244 – 250, Oct. 2005.
- [3] A. Briassouli and N. Ahuja. Estimation of multiple periodic motions from video. In *9th European Conference on Computer Vision*, May 2006.
- [4] 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.
- [5] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):781–796, Aug. 2000.
- [6] 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.
- [7] S. M. Kay. *Modern Spectral Estimation, Theory and Applications*. Prentice-Hall, Englewood Cliffs, NJ, 1988.
- [8] R. Polana and R. Nelson. Detection and recognition of periodic, nonrigid motion. *International Journal of Computer Vision*, 23(3):261–282, 1997.
- [9] B. S. Reddy and B. N. Chatterji. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, Aug. 1996.
- [10] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. Wiley, New York, 1977.
- [11] P. Tsai, M. Shah, K. Keiter, and T. Kasparis. View-invariant analysis of cyclic motion. *International Journal of Computer Vision*, 25(12):1–23, 1997.

¹The support of the Office of Naval Research Under Grant N00014-03-1-0107 is gratefully acknowledged.