Fusion of frequency and spatial domain information for motion analysis

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

Abstract

This paper presents an approach to the analysis of multiple motions in video, which combines frequency and spatial domain information in a new manner. The tasks of interest are finding the number of moving objects, velocity estimation, object tracking, and motion segmentation. We propose a novel, hybrid approach, which avoids problems of spatial domain methods, like sensitivity to illumination changes or great computational cost, but also uses spatial information for precise object localization. Experiments with synthetic and real sequences, show both the possibilities and limitations of this approach.

1. Introduction

This paper is concerned with object level motion analysis of video sequences, such as counting the number of moving objects, motion estimation, and motion based segmentation. Algorithms for these tasks are useful in many applications such as tracking, registration, and recognition.

Much of the previous work on motion analysis has used the spatial domain representation of images [1], [2] [3], which models motion phenomena in terms that directly correspond to how we ourselves may often describe the motion. Our ongoing work is aimed at combining multiple representations to explicitly capture the image characteristics most pertinent to various motion analysis tasks. In this paper, we focus on the Fourier Transform (FT) and the joint use of image and Fourier representations. In Sec. 2 we motivate the use of the FT. Sec. 3 presents the formulation of the FT method, and algorithms for the motion analysis tasks are presented in Sec. 4 - 5. In Sec. 6 we propose the integration of frequency and image-based methods. Sec. 7 presents experimental results and concluding remarks are given in Sec. 8.

2 Motivation for Frequency-Domain and Integrated Approaches

Since a video sequence depicts the temporal repetition of the moving objects, the requirements of certain motion analysis tasks may be well matched to a frequency representation. Frequency domain processing has several advantages over spatial domain methods. The motion estimation is based on the phase changes of the FT, so it is robust to global illumination changes. Its computational cost is significantly lower, making it more useful for practical applications. The size and shape of the moving objects does not affect the motion analysis, as the motion estimation and initial segmentation does not use spatial data.

FT processing alone is not adequate for all motion tasks, so spatial information is used for more accurate segmentation. Until now, there have been few approaches to the strongly inter-connected problems of motion estimation and segmentation that can take advantage of both frequency and spatial domain information [4]. Current frequency based methods [4], [5] use the FT *only* for the motion estimation, and not the segmentation. We propose a new approach, which estimates the number of moving objects and the independent motions in the Fourier domain, and fuses spatial and frequency data for the motion segmentation.

3 Frequency Domain Formulation

Let each frame consist of M moving objects l, $1 \leq l \leq M$, with luminance $s_l(\bar{r})$ at pixel $\bar{r}$ and velocity $\bar{u}_l$, initially considered constant. In Sec. 5.2 we show how non-constant translations can also be estimated. The FT of object l is $S_l(\bar{\omega}) = A_l(\bar{\omega})e^{j\Phi_l(\bar{\omega})}$, where $\bar{\omega} = [2\pi m/N_1, 2\pi m/N_2]^T$, $m, n \in Z$ is the 2-D frequency, $N_1 \times N_2$ the image size, $A_l(\bar{\omega})$ the FT magnitude and $\Phi_l(\bar{\omega})$ the FT phase. Each object is displaced by $\bar{r}_l = \bar{u}_l$ for inter-frame time $t = 1$, so its FT becomes $S'_l(\bar{\omega}) = S_l(\bar{\omega})e^{-j\bar{\omega}^T\bar{r}_l}$. The background

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

has FT $S_b(\bar{\omega})$, the FT of each frame is $X_k(\bar{\omega})$, $1 \leq k \leq N$, and the measurement noise is $V_{noise,k}(\bar{\omega})$, giving[1]

$$X_1 = S_b + S_1 + S_2 + ... + S_M + V_{noise,1}. \quad (1)$$

A moving object occludes a part of the background and unoccludes another. The FT's of the unoccluded and occluded parts of the background from frame 1 to k are denoted as $S_{un}^{1,k}$ and $S_{occ}^{1,k}$ respectively, so frame N has FT

$$\begin{aligned} X_N &= S_b + S_1 e^{-j(N-1)\bar{\omega}^T \bar{r}_1} + ... + S_M e^{-j(N-1)\bar{\omega}^T \bar{r}_M} \\ &+ S_{un}^{1,N} - S_{occ}^{1,N} + V_{noise,N}. \end{aligned} \quad (2)$$

Stacking the FT's of the N frames, we get $X = Z + V_{noise} + V_{bck}$, where Z is the $N \times (M+1)$ data matrix, V_{noise} is the additive measurement noise, and V_{bck} represents the occluded and unoccluded background areas, with elements $V_{bck,k} = S_{un}^{1,k} - S_{occ}^{1,k}$. We can decompose Z as $Z = AS$, where $S = [S_b, S_1, ..., S_M]^T$, and A is a Vandermonde matrix containing the motion information, with rows:

$$A_k = [1,\ e^{-j(k-1)\bar{\omega}^T \bar{r}_1},\ \ldots,\ e^{-j(k-1)\bar{\omega}^T \bar{r}_M}]. \quad (3)$$

Then we have

$$X = AS + V_{noise} + V_{bck}. \quad (4)$$

This is an over-determined system, which can be solved in a Least Squares (LS) sense to give S.

4 Object Number - Counting of Independent Motions

The rank of the noiseless data correlation matrix $R_Z = AR_SA^H$, where R_S is the correlation matrix of S, is equal to the rank of A. Due to its Vandermonde structure, A has M independent columns, so its rank gives the number of independently moving objects. For noise with $R_V = \sigma^2 I$ (w.l.o.g.), the singular values of the sample correlation matrix R_X are $\{\sigma_1^2 + \sigma^2, ..., \sigma_M^2 + \sigma^2, \sigma^2, ..., \sigma^2\}$, where $\sigma_k^2, 1 \leq k \leq M$ are the singular values of R_Z. In practice, $\sigma_k^2 \gg \sigma^2$, so M can still be determined from them. This estimate is useful for motion estimation, in addition to being a motion parameter of interest by itself.

5 Retrieval of Harmonics - Motion Estimation

The FT of the frames contains the motion information in the form of a sum of weighted harmonics. Existing transform domain methods (which assume constant translational motion) perform spatiotemporal filtering [6], estimate the 3D FFT [7] or use harmonic retrieval methods [5]. We propose a simpler, computationally less costly method for motion estimation. Unlike existing FT methods, it has the significant advantage of not being restricted to constant translations.

5.1 Constant Motion

Consider the phase change $\Phi_{1,k}$ of frames $1-k$:

$$\Phi_{1,k} = \frac{S_b + S_1 e^{-j(k-1)\bar{\omega}^T \bar{r}_1} + ... + S_M e^{-j(k-1)\bar{\omega}^T \bar{r}_M}}{S_b + S_1 + ... + S_M}. \quad (5)$$

Its inverse FT $\phi_{1,k}$ is a weighted sum of delta functions, that displays peaks corresponding to the harmonics $\bar{\omega}^T \bar{r}_l = [\omega_x x_l, \omega_y y_l]$ for each object l, from which we can extract the motion ($\bar{r}_l = (x_l, y_l)$).

In practice, the resolution of the peaks can be degraded by the aliasing from neighboring side-lobes. Thus, we identify the motions in a scene sequentially: the stronger motion components and a set of values close to them are estimated first and removed [5], so that the weaker harmonics can be detected more easily. The previously estimated number of motions helps, since we repeat this process until M motions are extracted.

5.2 Time-Varying Motion

For time-varying translations, the initially estimated $\bar{r}_{1,k}$ gives the average velocity $\bar{u}_{1,k}^{avg} = \bar{r}_{1,k}/T_{1,k}$, where $T_{1,k}$ is the time from frame 1 to k. This can be repeated for increasingly shorter subsequences, until their velocities become similar. Then, the constant components of the time-varying motion have been found.

If the velocity of object l from frame 1 to k is $\bar{u}_{1,k}^l$, $Z = \hat{A}S$, and the rows of $\hat{A}$ are

$$\hat{A}_k = [1,\ e^{-j(k-1)\bar{\omega}^T \bar{u}_{1,k}^1},\ \ldots,\ e^{-j(k-1)\bar{\omega}^T \bar{u}_{1,k}^M}]. \quad (6)$$

$\hat{A}$ does not have a Vandermonde structure, so the number of independent motions cannot be estimated beforehand, but the displacements between frames 1 and k can still be estimated as in Sec. 5.1. From these motion estimates, the number of motions is found a posteriori, and S can be retrieved from the LS solution of (4), with $\hat{A}$ instead of A.

6 Motion Segmentation - Object Tracking

6.1 Iterative LS for Motion Segmentation

A more accurate solution for S could be obtained from $X' = X - V_{bck} = Z + V_{noise}$, if V_{bck} was known. We approximate V_{bck} using an "object mask" (Sec. 6.1.1- 6.1.3),

[1] For simplicity we omit $\bar{\omega}$ from the function arguments in the sequel.

which can be iteratively improved to get to better approximations $\hat{V}_{bck}$ of V_{bck}.

6.1.1 Difference Mask for Each Object from Frequency Domain Solutions

From each LS solution $\hat{s}_l$ and the frame luminance s at each pixel we get $D_l(x,y) = |s(x,y) - \hat{s}_l(x,y))|$, $D_{max,l} = max_{x,y}\{D_l(x,y)\}$ and:

$$D_{mask,l}(x,y) = \frac{D_{max,l} - D_l(x,y)}{D_{max,l}}, \qquad (7)$$

which is closer to 1 for pixels belonging to object l, since $D_l(x,y) \simeq 0$ in those positions. In pixels that don't belong to object l, $D_l(x,y)$ has higher values, so $D_{mask,l}$ is closer to 0. Thus, the LS solutions S_l lead to a measure of the probability that pixel (x,y) belongs to object l.

6.1.2 Probability Mask from Velocity Mapping

The frame pixels are tracked by assigning the M object velocities and the background velocity 0 to each of them. When a pixel $s_l(x,y)$ is tracked with its correct velocity $\bar{u}_l$, its luminance remains mostly constant over the frames, i.e. it displays a small variance σ_l^2. If it is tracked with the wrong velocity $\bar{u}_k$, the variance σ_k^2 of the incorrectly tracked pixels increases. We denote by F_l the probability that the tracked pixel's distribution fits the small-variance distribution, i.e. that the pixel belongs to object l. This gives a "spatial probability mask" $P_l(x,y) = F_l(x,y)/\sum_{i=1}^{M} F_i(x,y)$, the probability that (x,y) belongs to object l, $1 \leq l \leq M$.

6.1.3 Final Probability Mask for Motion Regions

The frequency-based and spatial "probability masks" are combined via pixel-wise multiplication and the result is normalized (so it takes values between 0 and 1). This gives the best possible "object probability mask" that helps find $\hat{V}_{bck}$ and S.

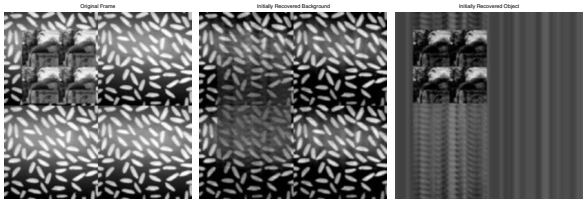
7 Experiments

7.1 Number of Motions: Synthetic Data

The first sequence consists of two squares with parts of "Cameraman" translating against a black background with $\bar{u}_1 = [3,7]$, $\bar{u}_2 = [15,25]$. From the SVD of R_X, we see that two harmonics are present, as two of its singular values are $118, 117$, i.e. much higher than the rest, which are under 50. The periodogram peaks give the correct velocities and the moving objects are retrieved accurately from (4).

Table 1. Estimates for varying velocity

Real $\bar{u}$ - 1st part	$\bar{u}_{1a} = [1,3]$	$\bar{u}_{2a} = [2,5]$
frames $1-16$	[3.21, 4.93]	[5.29,7.57]
frames $1-3$	[1.33,3.33]	[2,4.67]
Real $\bar{u}$ - 2nd part	$\bar{u}_{1b} = [5,7]$	$\bar{u}_{2b} = [8,10]$
frames $1-16$	[3.21, 4.93]	[5.29,7.57]
frames $4-16$	[4.18,5.64]	[6.18, 8.36]



(a) (b) (c)

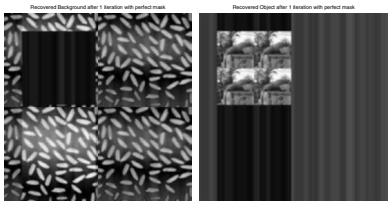
Figure 1. (a) Original frame. (a) Initially recovered background from LS solution. (b) Recovered object from the initial LS frequency domain solution.

7.2 Non-constant translational motion

Using the same synthetic data as before, we generate 16 frames: for frames $1-5$ we have $\bar{u}_{1a} = [1,3]$ and $\bar{u}_{2a} = [2,5]$, and for $6-16$, $\bar{u}_{1b} = [5,7]$ and $\bar{u}_{2b} = [8,10]$. Table 1 shows that if we use all the frames we get estimates between the true values of the two different pairs of velocity vectors. If we break the sequence into two parts, there are still errors, but there is a clear separation of the time varying velocities.

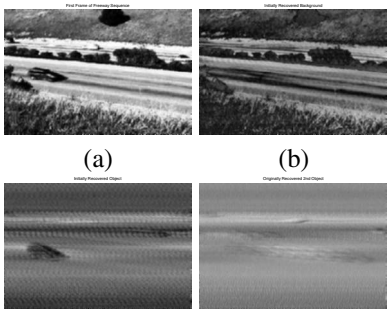
7.3 Real-Image Synthetic Sequence

To effectively test the method in a controlled but realistic environment, we perform experiments with the synthetic 20 frame sequence of Fig. 1(a). The motion is accurately estimated to be $\bar{u} = [15,0]$. In Fig. 1 there are artifacts because of V_{bck}, which become black (zero) when $\hat{X}$ is used (Fig. 2). There still are vertical stripes outside the "motion region", because of the regularization involved in the LS solution of (4). The MSE also gives an indication of the improvement achieved by accounting for V_{bck}. The original S_b estimate had $MSE_b = 4.8581 \times 10^{-4}$, but after compensating for V_{bck}, it becomes $MSE_b = 4.6982 \times 10^{-4}$, and the object MSE decreases from $MSE_1 = 7.8937 \times 10^{-4}$ to 6.4361×10^{-4}. The results of Fig.2, however, give a better indication of the improvement achieved.



(a) (b)

Figure 2. Ideally extracted background and object with perfect knowledge of V_{bck}. (a) Recovered background. (b) Recovered object.



(a) (b)

(c) (d)

Figure 3. (a) First frame. (b) Initially recovered background: the cars have been "erased"! (c) Originally extracted first car with artifacts from V_{bck}. (d) The second car is more difficult to extract: its color is similar to the background.

7.4 Real Sequence with Multiple Objects

We examine a sequence with a dark car moving to the right, and a white one moving to the left in the other lane (Fig. 3(a)). In the initial LS solution (Fig. 3(b)-(c)), the background and the two independently moving objects have been separated, but the frequency domain results need improvement, so the spatial techniques of Sec. 6 are used. We get separate masks (Fig. 4) for the two objects as in Sec. 6.1. The iterative approach improves the MSE, which decreases from $MSE_b = 7.5062 \times 10^{-4}$ to 6.9172×10^{-4} for the background, from $MSE_1 = 10.3238 \times 10^{-4}$ to $MSE_1 = 10.0204 \times 10^{-4}$ for the first object and from $MSE_2 = 9.8499 \times 10^{-4}$ to $MSE_2 = 9.0504 \times 10^{-4}$ for the second object.

(a) (b)

Figure 4. The cars after masking. The estimated position and shape are essentially correct for both cars.

8 Future Directions - Conclusions

The method presented is promising, as it can achieve object counting, simultaneous motion estimation and segmentation almost entirely in the frequency domain, and thus avoid many problems of spatial methods. It is extended to varying translations, unlike existing frequency-domain methods. More generalizations, for the estimation of multiple rotations are under current investigation.

References

[1] J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Trans. on Image Processing*, Vol. 3, Sept. 1994, pp. 625-638.

[2] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, Issue 8, Aug. 2000, pp. 747-757.

[3] C. Debrunner and N. Ahuja, "Segmentation and Factorization-Based Motion and Structure Estimation for Long Image Sequences," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 20, Issue 2, Feb. 1998, pp. 206 -211.

[4] W. Yu, G. Sommer and K. Daniilidis "Multiple Motion Analysis: in Spatial or in Spectral Domain?," *Computer Vision and Image Understanding*, Vol. 90, 2003, pp. 129-152.

[5] W. Chen, G. B. Giannakis and N. Nandhakumar "A Harmonic Retrieval Framework for Discontinuous Motion Estimation," *IEEE Trans. on Image Processing*, Vol. 7, No. 9, Sept 1998, pp. 1242-1257.

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

[7] A. Kojima, N. Sakurai, J. I. Kishigami, "Motion detection using 3D-FFT spectrum," *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, Vol. 5, 27-30 April 1993, pp. 213-216.