Integrated Spatial and Frequency Domain 2D Motion Segmentation and Estimation

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

A video containing multiple objects in rotational and translational motion is analyzed through a combination of spatial and frequency domain representations. It is argued that the combined analysis can take advantage of the strengths of both representations. Initial estimates of constant, as well as time-varying, translation and rotation velocities are obtained from frequency analysis. Improved motion estimates and motion segmentation for the case of translation are achieved by integrating spatial and Fourier domain information. For combined rotational and translational motions, the frequency representation is used for motion estimation, but only spatial information can be used to separate and extract the independently moving objects. The proposed algorithms are tested on synthetic and real videos.

1. Introduction

The estimation of multiple motions in a video sequence, and its segmentation into independently moving objects, are required in numerous applications. Although motion analysis is often carried out directly in the spatial domain, it can also be performed using the frequency representation of the sequence [1], [3]. Motion estimation based on the Fourier Transform (FT) of a video sequence offers several advantages: (1) it is robust to global illumination changes. (2) The inaccuracies in motion estimation near object boundaries are avoided, since the estimates are not based on spatially local inter-frame luminance differences, e.g. involving optical flow, rather, they are derived from global intensity distribution. (3) Computationally, algorithms are available for efficient FT calculation.

There are a limited number of motion estimation methods that use the frequency domain. The contributions of the motion segmentation and estimation algorithms we present include the following. (1) It can process sequences undergoing time-varying rotations as well as translations. In contrast, the few existing FT based approaches use the FT only for translations [4], [2]. (2) It uses a novel way of deriving an initial motion segmentation in the frequency domain and integrating this result with spatial information. This yields much better results than the existing methods, spatial or FT based alone, in which motion segmentation is carried out entirely in the spatial domain.

2. Overview

The main components of our algorithm are as follows: (1) Translations with time-varying velocity (Sec. 3) are first estimated in the frequency domain (Sec. 4). These motion estimates, and the FT of the sequence, form an overdetermined linear system. By solving this system in a Least Squares (LS) sense, the position, shape and texture of the moving objects and the background are estimated in a novel manner (Sec. 5). The LS object estimates have blurry boundaries, but can be used as a good initial estimates, which are refined in the spatial domain (Sec. 6).

(2) For time-varying roto-translational motions (Sec. 7), the frame FTs are mapped to the polar domain. The rotation angles appear as translations in the polar domain, so they can be recovered through a procedure similar to that used for translations. The frames are then successively derotated by the estimated object rotation angles, to estimate the corresponding object translations. Thus, complex object motions are fully characterized using the FT data, contrary to existing FT based methods, that estimate only translations. Segmentation is performed in the spatial domain: each frame is predictively warped by the estimated motions and compared with the input to find the object that under-

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

went that motion. This object is removed from the sequence and the procedure is repeated for the rest of the moving objects. Sec. 8 presents experimental results.

3. Formulation for Translation

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 $\bar{u}_l(t)$. The FT of object l is $S_l(\bar{\omega}) = M_l(\bar{\omega})e^{j\Phi_l(\bar{\omega})}$, where $\bar{\omega} = [2\pi m/N_1, 2\pi n/N_2]^T$, $m, n \in Z$, is the 2-D frequency, $N_1 \times N_2$ the image size, $M_l(\bar{\omega})$ the FT magnitude, and $\Phi_l(\bar{\omega})$ the FT phase. Each object has velocity $\bar{u}_l(t)$, so for inter-frame time T, it is displaced by $\bar{r}_l^T = \int_t^{t+T} \bar{u}_l(\tau)d\tau$, i.e. object luminance s'_l after displacement becomes $s'_l(\bar{r}) = s_l(\bar{r} - \bar{r}_l^T) = s_l(\bar{r} - \int_t^{t+T} \bar{u}_l(\tau)d\tau)$, with FT $S'_l(\bar{\omega}) = S_l(\bar{\omega})e^{-j\bar{\omega}^T\bar{r}_l^T}$. For simplicity, let $T = 1$, so that $\bar{r}_l^k = \int_0^k \bar{u}_l(\tau)d\tau$ at frame k. The FT of frame k is $X_k(\bar{\omega})$, $1 \leq k \leq N$, the measurement noise is denoted by $V_{noise,k}(\bar{\omega})$, and the background area occluded by the moving objects is denoted by $V_{bck,k}(\bar{\omega})$. Then the FT $X_1(\bar{\omega})$ of the first frame is

$$X_1(\bar{\omega}) = S_b(\bar{\omega}) + S_1(\bar{\omega}) + ... + S_M(\bar{\omega}) - V_{bck,1}(\bar{\omega}) + V_{noise,1}(\bar{\omega}), \quad (1)$$

where $v_{bck,1}(\bar{r}) = s_b(\bar{r})m_{obj,1}(\bar{r}) = s_b(\bar{r})(m_1(\bar{r}) + ... + m_M(\bar{r}))$ is the background area hidden by the M objects in the spatial domain and $V_{bck,1} = S_b(\bar{\omega}) * M_{obj,1} = S_b * (M_1 + ... + M_M)$ is its FT. Here, $m_{obj,1}(\bar{r})$ is essentially the foreground mask, which zeroes out each non-object area $m_l(\bar{r})$. For simplicity, we will assume that the objects only move against the background and do not occlude each other. The methodology of this paper can be extended in a straightforward manner to relax this assumption, but we will not do so in this paper, for lack of space. The FT[1] of frame k is written as

$$X_k = S_b + ... + S_M e^{-j\bar{\omega}^T\bar{r}_M^k} - V_{bck,k} + V_{noise,k}, \quad (2)$$

where $V_{bck,k} = S_b * (M_1 e^{-j\bar{\omega}^T\bar{r}_1^k} + ... + M_M e^{-j\bar{\omega}^T\bar{r}_M^k})$. Stacking the FT's of the N frames, we get $X = Z + V_{noise} - V_{bck}$, where X is an $N \times 1$ data vector, with the values of each frame's FT, and Z is an $N \times 1$ vector, whose kth element (for frame k) is $Z_k = S_b + S_1 e^{-j\bar{\omega}^T\bar{r}_1^k} + ... + S_M e^{-j\bar{\omega}^T\bar{r}_M^k}$. V_{noise} is the $N \times 1$ vector containing the additive measurement noise, and $V_{bck,k}(\bar{\omega})$ is an $N \times 1$ vector that represents the occluded background areas for each frame, k given by $V_{bck,k} = S_b * M_{obj,k}$. We decompose Z as $Z = AS$, where $S(\bar{\omega})$ is the $(M+1) \times 1$ FT vector of the background and the objects, and $A(\bar{\omega})$ is a $N \times (M+1)$ matrix containing the motion information, with row k $A_k(\bar{\omega}) = [1,\ e^{-j\bar{\omega}^T\bar{r}_1^k},\ \ldots,\ e^{-j\bar{\omega}^T\bar{r}_M^k}]$. This leads to the over-determined system

$$X = AS + V_{noise} - V_{bck}, \quad (3)$$

where the objects' displacement information appears in the matrix A as a sum of weighted harmonics. Our formulation aims to extract the object displacements between frames 1 and any subsequent frame k, and segmenting the moving objects by solving (3).

4. Frequency Translation Estimation

We first consider the estimation of translation between two given frames, 1 and k. Consider the ratio $\Phi_{1,k}$ of the FTs of frames k and 1:

$$\Phi_{1,k}(\bar{\omega}) = \frac{X_k(\bar{\omega})}{X_1(\bar{\omega})} = a_b(\bar{\omega}) + \sum_{i=1}^{M} a_i(\bar{\omega})e^{-j\bar{\omega}^T\bar{r}_i^k} + n_k(\bar{\omega}), \quad (4)$$

where, for $1 \leq l \leq M$

$$a_b(\bar{\omega}) = \frac{S_b(\bar{\omega})}{S_b(\bar{\omega}) + S_1(\bar{\omega}) + ... + S_M(\bar{\omega}) + V_1(\bar{\omega})},$$
$$a_l(\bar{\omega}) = \frac{S_l(\bar{\omega})}{S_b(\bar{\omega}) + S_1(\bar{\omega}) + ... + S_M(\bar{\omega}) + V_1(\bar{\omega})} \quad (5)$$

and $n_k(\bar{\omega}) = (V_{noise,k}(\bar{\omega}) - V_{bck,k}(\bar{\omega}))/(S_b(\bar{\omega}) + S_1(\bar{\omega}) + ... + S_M(\bar{\omega}) + V_1(\bar{\omega}))$, with $V_1(\bar{\omega}) = V_{noise,1}(\bar{\omega}) - V_{bck,1}(\bar{\omega})$. From (4) we see that $\Phi_{1,k}(\bar{\omega})$ is just a sum of weighted exponentials, whose inverse FT $\phi_{1,k}(\bar{r})$ is a weighted sum of delta functions:

$$\phi_{1,k}(\bar{r}) = a_b(\bar{r}) + \sum_{i=1}^{M} a_i(\bar{r})\delta(\bar{r} - \bar{r}_i^k) + n_k(\bar{r}). \quad (6)$$

The term $n_k(\bar{r})$ contains the measurement noise $v_{noise,k}(\bar{r})$ and the background occlusion $v_{bck,k}(\bar{r})$. The FT $V_{bck,k} = S_b(\omega) * (\sum_{i=1}^{M} M_i(\omega)e^{-j\bar{\omega}^T\bar{r}_i^k})$ does not include significant harmonics, which could affect the accuracy of the motion estimates, so $\phi_{1,k}(\bar{r})$ has peaks at the displacements $\bar{r}_l^k$, $1 \leq l \leq M$. For better resolution, we estimate the energy $|\phi_{1,k}(\bar{r})|^2 = \sum_{i=1}^{M} a_i^2(\bar{r})\delta^2(\bar{r} - \bar{r}_i^k) + \sum_{i \neq j} a_i(\bar{r})a_j(\bar{r})\delta(\bar{r} - \bar{r}_i^k)\delta(\bar{r} - \bar{r}_j^k) + [a_b(\bar{r}) + n_k(\bar{r})]^2 + 2\sum_i [a_b(\bar{r}) + n_k(\bar{r})] \sum_{i=1}^{M} a_i(\bar{r})\delta(\bar{r} - \bar{r}_i^k)$. The cross terms in the sum are zero for $\bar{r}_i^k \neq \bar{r}_j^k$, so only the peaks at the true displacements remain. By taking the squared magnitude of $\phi_{1,k}(\bar{r})$, the peaks at $\bar{r}_l^k$ are enhanced. Detection of these peaks yields estimates of the object displacements. The last two lines of $|\phi_{1,k}(\bar{r})|^2$ contain noise terms, which depend on each frame's measurement noise $V_{noise,k}$, and the occluded background terms $V_{bck,k}$. These noise terms do not affect the accuracy of the motion estimates, since they do

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

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 IEEE

not introduce significant peaks at $\bar{r}_i^k$. Estimating the displacements between frame 1, and all other frames yields the 2D trajectory $\bar{r}_l(t)$ of each object, whose velocity is the time-derivative of $\bar{r}_l(t)$, i.e. $\bar{u}_l(t) = \frac{\partial \bar{r}_l(t)}{\partial t}$, defined over all t, if the trajectories are continuous functions of time.

5. LS Motion Segmentation

5.1. Effect of Background

For the case of pure translations, with constant or time-varying velocity, an estimate of motion and segmentation can be obtained in the frequency domain. Equation (3) can be solved to give the vector $S = [S_b, S_1, ..., S_M]^T$ containing the FTs of the background and the M moving objects at frequency $\bar{\omega}$. A straightforward (but erroneous) estimate of the solution can be obtained by neglecting the term $V_{bck}(\bar{\omega})$ and obtaining an LS estimate of the solution of (3) assuming $V_{noise}(\bar{\omega})$ is zero mean. To understand how the error of neglecting $V_{bck}(\bar{\omega})$ affects the object FT estimation, consider the simple example of two objects in frame 1 that move by $\bar{r}_l$, $l = \{1, 2\}$, in frame 2: $X_2 = S_b + S_1(\bar{\omega})e^{-j\bar{\omega}^T\bar{r}_1} + S_2 e^{-j\bar{\omega}^T\bar{r}_2} + V_{all,2}$, where $V_{all,2}(\bar{\omega}) = V_{noise,2}(\bar{\omega}) + V_{bck,2}(\bar{\omega})$. The $N_1 \times N_2$ LS solutions $\hat{S}_b(\bar{\omega})$, $\hat{S}_1(\bar{\omega})$, $\hat{S}_2(\bar{\omega})$ also include the effects of $V_{all,2}(\bar{\omega})$. If we identify the deviations caused by the LS approximation and $V_{all}(\bar{\omega})$ by $V_b(\bar{\omega})$, $V_1(\bar{\omega})$, and $V_2(\bar{\omega})$ respectively, then we can write:

$$\begin{aligned} X(\bar{\omega}) &= \hat{S}_b(\bar{\omega}) + \hat{S}_1(\bar{\omega})e^{-j\bar{\omega}^T\bar{r}_1} + \hat{S}_2(\bar{\omega})e^{-j\bar{\omega}^T\bar{r}_2} \\ &= (S_b(\bar{\omega}) + V_b(\bar{\omega})) \\ &+ (S_1(\bar{\omega}) + V_1(\bar{\omega})e^{j\bar{\omega}^T\bar{r}_1})e^{-j\bar{\omega}^T\bar{r}_1} \\ &+ (S_2(\bar{\omega}) + V_2(\bar{\omega})e^{j\bar{\omega}^T\bar{r}_2})e^{-j\bar{\omega}^T\bar{r}_2}, \end{aligned} \quad (7)$$

If we write $V_l(\bar{\omega}) = M_{V,l}(\bar{\omega})e^{-j\phi_v}$, the error in each LS estimate is $V_l(\bar{\omega}) = M_{V,l}(\bar{\omega})e^{-j\phi_v(\bar{\omega})}e^{j\bar{\omega}^T\bar{r}_l}$, $l = \{b, 1, 2\}$. The term $\phi_v(\bar{\omega})$ changes for each $\bar{\omega}$ in an unknown way, whereas the term $\bar{\omega}^T\bar{r}_l$ is a fixed repeated pattern given by a plane whose normal is $\bar{r}_i$. Thus, $e^{j\bar{\omega}^T\bar{r}_l}$ forms a conspicuous part of the solution error since the other part $e^{-j\phi_v(\bar{\omega})}$ is not necessarily periodic and as noticeable. This suggests that the recovered background will have a periodic component in the errors, and, more importantly, that there is a form of spectral leakage (frequency space blurring) in the object FTs $\hat{S}_l(\bar{\omega})$, which will appear as sinusoidal artifacts in the estimates. This prediction is verified in the experiments we report in Sec. 8.

5.2. Regularization of LS Solution

If the SVD of A is $A = U\Sigma V^H$, the LS solution to (3) for the noiseless case is given by $S = (A^H A)^{-1}A^H X = V\Sigma^{-1}U^H X$, where Σ^{-1} is diagonal with values $1/\sigma_l$ for $\sigma_l \neq 0$ and 0 for $\sigma_l = 0$. The smaller singular values correspond to the high-frequency components of the solution, so small errors in small σ_l's introduce large changes in the corresponding high-frequency components of S. These errors appear in the form of large oscillations superimposed on the solution, and render it useless. To deal with this instability, the Tikhonov regularization algorithm is used, which, for a system $X = AS$ minimizes $||AS - X||_2^2 + \lambda||S||_2^2$, where λ is a positive constant that controls the size of the solution vector[2]. The LS solution then becomes

$$S = (A^H A + \lambda I)^{-1}A^H X = \sum_{l=1}^{M} \frac{\sigma_l}{\sigma_l^2 + \lambda}\bar{v}_l\bar{u}_l^H X, \quad (8)$$

and the effect of $\sigma_l \simeq 0$ is dampened by the regularization parameter λ. From (8) we can see that the regularized solution will be more stable, but also biased, due to the addition of λ in the denominator, making the LS estimates slightly smaller, and therefore the object appearance darker than the actual objects. On the other hand, large values of λ reduce the accuracy of the LS solutions. Thus, there is a trade-off in the choice of this regularization parameter, whose ideal value cannot be determined à priori, as it requires knowledge of the actual solution. For this reason, we empirically tested numerous values of λ on over 10 different sequences (including the ones in the Sec. 8), and found that a value around 1 gave consistently good object estimates. Also, the LS estimates were robust to small deviations of λ around 1.

5.3. Effect of Noise

As described in Sec. 3, the background areas that are occluded contribute to the overall errors in the LS solution of (3). If we take the background occlusion effect and measurement noise $V_{all} = V_{bck} + V_{noise}(\bar{\omega})$ into account in Eq. (8), the LS estimates become $\hat{S} = (A^H A + \lambda I)^{-1}A^H(X - V_{all})$, so there is an error $\mathcal{E} = \hat{S} - S = -\sum_{l=1}^{M} \frac{\sigma_l}{\sigma_l^2+\lambda}\bar{v}_l\bar{u}_l^H V_{all}$, which depends on V_{all}, whose values over all frequencies and frames follow a distribution that changes with each video. This distribution is not known à priori, but it affects the error in the LS solution. The mean error for each frequency $\bar{\omega}$ is then given by $E[\mathcal{E}] = -\sum_{l=1}^{M} \frac{\sigma_l}{\sigma_l^2+\lambda}\bar{v}_l\bar{u}_l^H E[V_{all}]$, where the expectation is taken over the entire video sequence, i.e. over all the frames. Since V_{all} is not zero mean, due to the presence of V_{bck} term, $\hat{S}$ is biased.

[2] We use the L_2 norm $||x||_2 = \sqrt{x_1^2 + ... + x_N^2}$ for the $N \times 1$ vector x.

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

6. Integration with Spatial Estimates

The LS object estimates presented above differ from their true values due to the effect of the background (Sec. 5.1), and the approximation error introduced by the regularization (Sec. 5.2). To reduce these errors, we fuse the results from the frequency domain with complementary spatial information, as described in the sequel.

6.1. Correlation of LS Solution and Original

Each LS solution $s_l(x,y)$ is correlated with the original frame $x_1(x,y)$ in the spatial domain. The normalized cross-correlation $c_b(x,y)$ of $s_l(x,y)$ with $x_1(x,y)$ is computed over a square neighborhood $\mathcal{N}_b(x,y)$ around pixel (x,y), and it is high at pixels that belong to object l and low elsewhere. The correlation of the LS solution with the original at all (x,y) gives a "correlation map" $C(x,y)$ containing the coefficient values $c_b(x,y)$ at each pixel (x,y).

The correlation coefficient $c_b(x,y)$ follows Student's $t-$distribution, which approaches a Normal distribution as the number of samples increases. This is expected from the Weak Law of Large Numbers, since in that case $c_b(x,y)$ is the sum of a large number of samples, and has been verified by us experimentally. For normally distributed correlation coefficients, we have $Prob(x \in object) = P(c_b > \eta) = Q\left(\frac{\eta-\mu}{\sigma}\right)$, where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt$ gives the tail probability of a Gaussian distribution and μ, σ^2 are its mean and variance, respectively. The 90th percentile of the correlation coefficients is equivalent to $P(c_b > \eta) = \alpha$, where $\alpha = 0.1$, so the threshold for these correlation values is given by $\eta = \mu + \sigma Q^{-1}(\alpha)$. The correlation coefficients higher than this threshold are the highest $\alpha\%$ of all correlation values, and belong to the moving object.

6.2. Activity Areas

The movement of each pixel is followed over the entire sequence with a local velocity estimate $\bar{v}_l$, to determine if it belongs to object l. If its positions are correctly predicted, its values across the sequence differ only by measurement noise, so its luminance value remains approximately the same over all frames. Its value at frame k is $x_k(\bar{r}) = x_1(\bar{r}) + z_k(\bar{r})$, where $x_1(\bar{r})$ is its luminance value in the original frame, and $z_k(\bar{r})$ is additive zero mean Gaussian measurement noise at frame k. If a pixel is incorrectly followed, its value changes by $m_k(\bar{r})$, so $x_k(\bar{r}) = x_1(\bar{r}) + m_k(\bar{r}) + z_k(\bar{r})$. Then the problem of determining if a pixel belongs to object l or not is formulated as a binary hypothesis test:

$$\begin{aligned} H_0 : d_k(\bar{r}) &= z_k(\bar{r}) \\ H_1 : d_k(\bar{r}) &= m_k(\bar{r}) + z_k(\bar{r}), \end{aligned} \quad (9)$$

where $d_k(\bar{r}) = x_k(\bar{r}) - x_1(\bar{r}) \neq 0$. When the pixel is correctly followed, $d_k(\bar{r})$ follows the noise distribution (Gaussian). Under H_1, the data distribution changes significantly, since a pixel-dependent random quantity ($m_k(\bar{r})$) is added to the samples. Its variance increases greatly, because incorrect prediction of a pixel introduces abrupt changes in $d_k(\bar{r})$.

To determine whether $d_k(\bar{r})$ belongs to H_0 or H_1, it suffices to test the nongaussianity of the data. The classical measure of nongaussianity of a random variable y is the kurtosis, defined as $kurt(y) = E\{y^4\} - 3(E\{y^2\})^2$. The fourth moment of a Gaussian random variable is $E\{y^4\} = 3(E\{y^2\})^2$, so its kurtosis is equal to zero. Non-zero values of the kurtosis $kurt(d_k(\bar{r}))$ show that the pixel has been incorrectly displaced, so it does not belong to object l, and zero values show that the pixel indeed moved with $\bar{v}_l(\bar{r})$.

7. Translation and Rotation

In this section we consider M independently moving objects, where each object l first undergoes a translation, and then a rotation, with respect to a global coordinate system whose origin is in the center of the image. Frame 1 is $x_1(x,y) = s_b(x,y) + \sum_{i=1}^M s_i(x,y) + v_{noise,1}(x,y)$ and frame k is $x_k(x,y) = s_b(x,y) + \sum_{i=1}^M s_i((x - x_{i,k})\cos(\theta_{i,k}) + (y - y_{i,k})\sin(\theta_{i,k}), -(x - x_{i,k})\sin(\theta_{i,k}) + (y - y_{i,k})\cos(\theta_{i,k})) + v_{bck}^{1,k}(x,y) + v_{noise,k}(x,y)$, where $(x_{l,k}, y_{l,k})$ includes the distance of each pixel (x,y) from the center of the video frame, and its translation. The rotation appears in the magnitude of each object's FT, so the FT $S_i(\bar{\omega}) = M_i(\bar{\omega})e^{j\Phi_i(\bar{\omega})}$ of object l, from frame 1 to k, is

$$\begin{aligned} S_i^k(\omega_x,\omega_y) &= M_i(\omega_x\cos(\theta_{i,k}) + \omega_y\sin(\theta_{i,k}), \\ &- \omega_x\sin(\theta_{i,k}) + \omega_y\cos(\theta_{i,k}))e^{j(\Phi_i(\bar{\omega}) - \bar{\omega}^T\bar{r}_{i,k})}, \end{aligned}$$

For $W_{l,k}(\bar{\omega}) = \Phi_l(\bar{\omega}) - \bar{\omega}^T\bar{r}_{l,k}$, frame $k(\,1 \leq k \leq N)$ is

$$\begin{aligned} X_k(\omega_x,\omega_y) &= S_b(\omega_x,\omega_y) \\ &+ \sum_{i=1}^M M_i(\omega_x\cos(\theta_{i,k}) + \omega_y\sin(\theta_{i,k}), \\ &- \omega_x\sin(\theta_{i,k}) + \omega_y\cos(\theta_{i,k}))e^{jW_{i,k}(\bar{\omega})} \\ &+ V_{bck}^{1,k}(\omega_x,\omega_y) + V_{noise,k}(\omega_x,\omega_y). \end{aligned} \quad (10)$$

In log-polar coordinates, this becomes

$$\begin{aligned} X_k(\rho,\theta) &= S_b(\rho,\theta) + \sum_{i=1}^M M_i(\rho,\theta-\theta_{i,k})e^{jW'_{i,k}(\rho,\theta-\theta_{i,k})} \\ &+ V_{bck}^{i,k}(\rho,\theta) + V_{noise,k}(\rho,\theta) \\ &= D_b(\rho,\theta) + \sum_{i=1}^M D_i(\rho,\theta-\theta_{i,k}) \\ &+ V_{bck}^{1,k}(\rho,\theta) + V_{noise,k}(\rho,\theta), \end{aligned} \quad (11)$$

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

where we define $W'_{i,k}(\rho,\theta) = W_{i,k}(\rho,\theta+\theta_i)$ to represent the phase terms, $D_b(\rho,\theta) = S_b(\rho,\theta)$, and $D_i(\rho,\theta) = M_i(\rho,\theta)e^{jW'_{i,k}(\rho,\theta)}$. Thus, the rotation of each object appears as a translation along the θ axis, analogous to pure translation case in the x-y space in Sec 3. The inverse transform of Eq. (11) is

$$\begin{aligned} x_k(x_\rho, y_\theta) &= d_b(x_\rho, y_\theta) + \sum_{i=1}^{M} d_i(x_\rho, y_\theta) e^{j y_\theta \theta_{i,k}} \\ &+ v^{1,k}_{bck}(x_\rho, y_\theta) + v_{noise,k}(x_\rho, y_\theta), \quad (12) \end{aligned}$$

i.e. a weighted sum of exponentials, similar to Eq. (4). Consequently, the angles of rotation from frames 1 to k can now be found using the method presented in Sec. 4.

7.1. Translation Estimation

Once each object's rotation angle is estimated, it is used to estimate the corresponding translation. The rotations and translations that are found first are the dominant motions. Following this, the next strongest motions are estimated, until all motions have been found. Thus, initially frame k is de-rotated by the dominant angle $\theta_{1,k}$. This will lead to the extraction of the corresponding translation $\bar{r}_{1,k}$. The de-rotated frame k is expressed in spatial coordinates as

$$\begin{aligned} x'_k(x,y) &= s'_b(x,y) + s_1(x - x_{1,k}, y - y_{1,k}) + \\ &+ \ ... + s'_k(x,y) + ... + s'_M(x,y) \\ &+ v^{1,k}_{bck,rot}(x,y) + v^{k}_{noise,rot}(x,y), \quad (13) \end{aligned}$$

where $v^{1,k}_{bck,rot}$, $v^{k}_{noise,rot}$ represent $v^{1,k}_{bck}$ and v^{k}_{noise} in the de-rotated image, s'_b the de-rotated background, and s'_k are the de-rotated objects. The transform of (13) is

$$\begin{aligned} X'_k(\bar{\omega}) &= S'_b(\bar{\omega}) + S_1(\bar{\omega}) e^{-j\bar{\omega}^T \bar{r}_{1,k}} + ... \\ &+ S'_k(\bar{\omega}) + ... + S'_M(\bar{\omega}) \\ &+ V^{1,k}_{bck,rot}(\bar{\omega}) + V_{noise,k,rot}(\bar{\omega}), \quad (14) \end{aligned}$$

where $\bar{r}_{1,k} = (x_{1,k}, y_{1,k})$. This expression contains a harmonic corresponding to the desired displacement $\bar{r}_{1,k}$, around which most of the energy of (14) is concentrated.

The de-rotated background term $S'_b(\bar{\omega})$ complicates the extraction of the dominant motion, as it can introduce significant aliasing, so it should be suppressed before the translations are estimated. The other terms in this expression do not create aliasing problems, because $\bar{r}_{1,k}$ corresponds to the dominant motion. This is validated in the experiments, where it is shown that this translation can be reliably extracted when the background is suppressed.

8. Experiments

We now present the results of experiments with five video sequences, chosen to test the various aspects of our algorithm. Before we do that, it is useful to review the **main sources of inaccuracy** in the resulting estimates. First, the contrast of a moving object against its background determines the strength of its motion signatures in the FT domain: the changes in FT are smaller for smaller contrast, so the motion estimates are poorer for lower contrast objects. Second, when an object moves across a homogeneous background, the motion boundaries are ambiguous and hard to detect. Both of these disadvantages apply to any motion analysis method. Third, when certain background areas are never revealed throughout the sequence, or are visible for short periods of time, then those parts of the background are not recovered as well as expected from the analysis of Sec. 5.3.

1. Constant Translation Sequence: Experiments were first conducted with a synthetic sequence of a helicopter (Fig. 1(a)) translating with a constant velocity to the right. The motion estimation method of Sec. 4 gives a correct estimate $\bar{u} = (0, -8.6)$, which is constant throughout the sequence. The moving object is then separated from the background in the frequency domain, by finding the LS solution of Eq. 3. In Fig. 1, there is no sign of the helicopter on the recovered background, and the shape and texture of the helicopter have been retrieved correctly. There are some horizontal artifacts in the recovered object, because of the spectral leakage from the background occlusion areas (Sec. 5.1). The recovered background is slightly darker than in the original frame, due to the regularization (Sec. 5.2). Correlation in the spatial domain is then used to extract the moving object more accurately (Sec. 6). Fig. 1(d) shows an intermediate result, from the correlation of the LS solution for the object with the original frame using 10×10 blocks, and Fig. 1(e) shows the "activity areas" for this sequence (Sec. 6.2) after following each pixel. Here, the spatial information not only complements the LS solution, but the correlation and activity masks also complement each other. This is obvious in Figs. 1(f), where the background artifacts in the correlation and activity masks are in different areas of the image. The Mean Squared Error (MSE) for the recovered object is $MSE_{heli} = 46.7933$.

2. Time-Varying Translation: Experiments were conducted with a sequence of two objects undergoing time-varying translations (Fig. 2(a)). The motion of the two objects is estimated accurately and the estimates are used to get the LS solutions for the background and the moving objects, shown in Figs. 3(a)-(c). In Fig. 3(a) we can also see a characteristic case of the background not being perfectly recovered, as it was never completely revealed for a long enough period of time by the moving squares. This inaccuracy in the background estimate was also expected from the

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

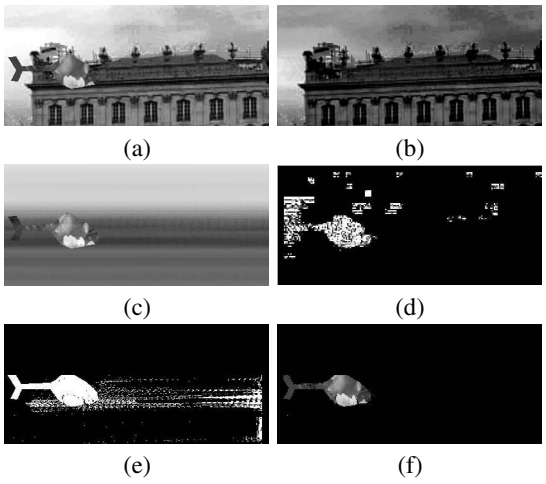


Figure 1. (a) Original frame. (a) Initially recovered background from LS solution. (b) Recovered object from the initial LS frequency domain solution. (c) Correlation with 2×2 blocks. (d) Activity mask obtained from pixel following. (e) Recovered object after fusion of LS solution and spatial domain information.

analysis of Sec. 5, since the background occlusion and measurement noise introduce an error $V_b(\bar{\omega})$ in that estimate.

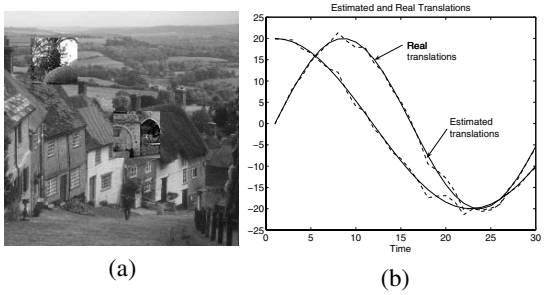


Figure 2. (a) Frame 1. (b) Estimated and real time varying translations as functions of time.

The original frame is first correlated with the LS estimated background to identify the background using both spatial and FT information. The extracted object areas of the original frame are then correlated with the LS solution for object 1, giving an estimate of its area in the frame (Fig. 4(a)). An activity mask is also calculated for each estimated velocity, giving possible areas for the corresponding objects (Fig. 4(b)). The combination of these masks leads to an accurate estimate of the object, as Fig. 4(c) shows, with

Figure 3. (a) Initially recovered background. (b) Initially recovered object 1. (c) Initially recovered object 2.

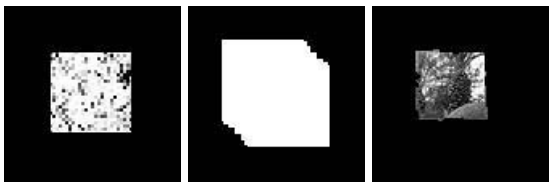


Figure 4. (a) Correlation of original frame with LS solution for object 1 over 2×2 blocks. (b) Activity mask for object 1. (c) Final recovered object 1 after integration of LS solution with spatial information.

errors $MSE_{o1} = 340.7229$, $MSE_{o2} = 201.784$.

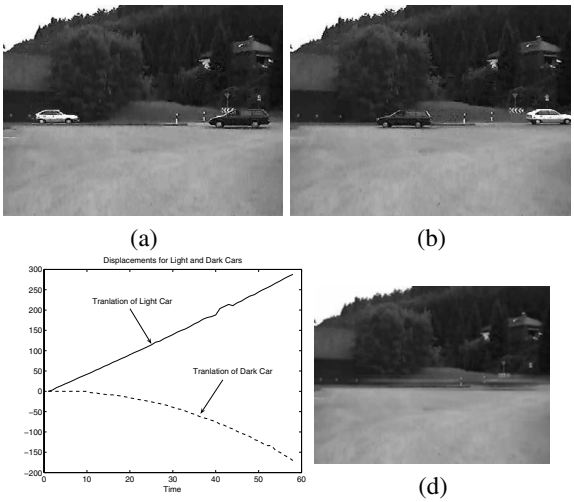
3. Rotation and Translation Sequence: We use the same initial frame as before (Fig. 2(a)) to create a synthetic sequence with two objects that undergo time-varying roto-translational motions. The dominant angle is found first, and the frame is derotated by it. The angles are estimated quite accurately with errors up to 0.08% of the range of rotation values.

4. Real Traffic Sequence: Experiments were conducted with a real traffic sequence, consisting of two cars that are turning (Fig. 5(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 quite accurately (Sec. 7) with errors up to 0.05% of the range of rotation values. The frames are de-rotated by the estimated angles to extract the translations of each object (Sec. 7.1). 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. 5 shows, the bottom right and top right cars are accurately recovered

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

(a) (a) (b)

Figure 5. Real traffic sequence: (a) Reconstructed bottom right car. (b) Reconstructed top right car.



(a) (b) (c) (d)

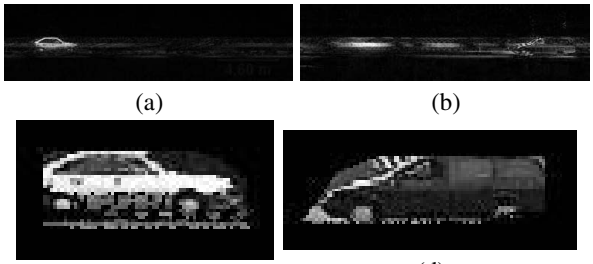
Figure 6. (a) Frame 1 of the car sequence. (b) Frame 58 of the car sequence. (c) Horizontal translation estimates as functions of time for both cars. (d)LS estimate of Background.

with $MSE_{o1} = 97.7$, $MSE_{o2} = 57.65$.

5. Real Car Sequence: Experiments were conducted with a real sequence of two cars (Fig. 7(a)-(b)) translating with time varying velocities (Fig. 7(c)).

The time varying translational velocities for each car are estimated using the FT of the frames, and these estimates are then used for the LS estimates of the moving cars (Sec. 3). As Figs. 7(d)-(f) show, the background and the two objects are recovered quite accurately. The solution for the second car (Fig. 7(f)) is less accurate, as expected, since this is a dark object moving against a dark background. Nevertheless, the shape and even details of the car (its wheels, windows and bumper) have been captured.

The LS solutions are correlated with the actual frames to find candidate object areas. "Activity masks" corresponding to each velocity are also extracted, and help eliminate artifacts from the correlation masks. The cars are finally recovered accurately, as shown in Figs. 7(g)-(h) ($MSE_{o1} = 71.9371$, $MSE_{o2} = 77.5467$). Some parts of the road around each moving car are extracted along with it, since they do not change significantly after motion compensation.



(a) (b) (c) (d)

Figure 7. LS estimates of: (a) White Car. (b) Dark car. (c) Finally recovered white car. (d) Finally recovered dark car.

9. Conclusions/Discussion

A novel hybrid method for motion analysis has been presented. The 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, or high computational cost.

References

[1] J. I. K. A. Kojima, N. Sakurai. Motion detection using 3d-fft spectrum. In *ICASSP-93, 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 213–216, April 1993.

[2] Anonymous.

[3] P. Milanfar. Two-dimensional matched filtering for motion estimation. *IEEE Transactions on Image Processing*, 8(3):438–444, March 1999.

[4] G. B. G. W. Chen and N. Nandhakumar. A harmonic retrieval framework for discontinuous motion estimation. *IEEE Transactions on Image Processing*, 7(9):1242–1257, Sept. 1998.

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**