Integration of Frequency and Space for Multiple Motion Estimation and Shape-Independent Object Segmentation

Alexia Briassouli, *Member, IEEE*, and Narendra Ahuja, *Fellow, IEEE*

***Abstract*—A video containing multiple objects undergoing independent translational and rotational motions is analyzed through a combination of spatial- and frequency-domain representations. The Fourier transform of the sequence is used to estimate the multiple translations and rotations in a computationally efficient manner, which is also robust to local inaccuracies and global illumination changes. A novel algorithm is presented for the simultaneous extraction of all objects undergoing translation and the background via a least squares technique that takes place entirely in the Fourier domain. Spatial information is combined with the frequency domain object extraction results, to further refine them. For the case of rotational or combined, rotational and translational motions, the moving objects are segmented using purely spatial information. We show that the combined analysis takes advantage of the strengths of both representations, by providing reliable and computationally efficient motion estimates and object segmentation. The proposed algorithm is shown to be robust to local noise and occlusion, because of its global nature. Experiments are performed on synthetic and real video sequences to demonstrate the capabilities of our approach.**

***Index Terms*—Fourier transform (FT), motion segmentation, phase-based motion estimation, video analysis.**

I. Introduction

APPLICATIONS of digital multimedia technology are becoming more widespread, as more devices are able to store, process and transmit images and videos over various networks. The advent of the semantic web [1], [2] and the latest MPEG video standards [3], [4] necessitate the efficient and reliable estimation of motions and the extraction of the corresponding moving objects. In this paper, we examine the case of a video containing multiple independently moving objects, which undergo 2-D planar translations, rotations, or a combination of both, over a static background. Such motions appear in many videos of practical interest, such as security applications, traffic surveillance, and sports videos.

A. Previous Work, Motivation

Many methods have been developed for the accurate extraction of motion information, using spatial domain data [5], [6], frequency transformed data [7]–[9], or a combination of both [10]. Traditionally, motion estimation takes place in the spatial domain, and is based on the changes in luminance between successive frames [5], [11]. The changes in the brightness pattern of an image sequence are known as the "optical-flow," to which an appropriate parametric motion model can be fitted [12]. Afterwards, pixels that undergo a similar motion are clustered together, thus defining objects, or groups of objects, in a scene [13], [14] (motion segmentation). This, of course, is based on the assumption that pixels with similar motion belong to the same object.

However, spatial methods for motion estimation are prone to errors, as they require very small inter-frame displacements and constant scene illumination. They are very sensitive to local inaccuracies, or small motion discontinuities, which make the motion estimates on the boundaries of moving objects, and the corresponding segmentation, unreliable [15]. Various methods have been developed to improve the results of spatial domain motion estimation, that involve regularization [16] and smoothness constraints on the extracted motion field [5], [17]. Nevertheless, these approaches remain local, and consequently still encounter difficulties when faced with local inaccuracies. Additionally, their computational cost is high, as they require pixel-wise or block-wise processing of each frame.

Frequency, or spatiotemporal-frequency domain methods have been developed to complement or overcome some difficulties of spatial-only methods [15], [18]. These approaches are based on the phase shift introduced to the frequency domain representation by spatial translations [19], [20]. They process the entire frames simultaneously, so they are inherently robust to local inaccuracies, like local occlusion, illumination changes and motion discontinuities. The global frame processing makes them less sensitive to smoothly textured areas than pixel-based approaches (such as optical flow), which are based on pixel-wise illumination differences. The numerous methods available for the fast computation of frequency transforms [21]–[23], and their noniterative nature [15], make them computationally efficient.

One of the main advantages of Fourier-domain methods, which is their global nature, is also a disadvantage, as it introduces the "localization problem:" there is no immediate connection between the motion estimates and their spatial location [15]. Spatiotemporal-frequency domain approaches, like [18], overcome this problem by simultaneously estimating the time-varying frequencies corresponding to each pixel. However, methods that are based on the use of the Wigner–Ville distribution (WVD) [18], [24] are unsuitable for the estimation of multiple motions, as the WVD contains many cross-corre-

Manuscript received October 5, 2006; revised April 4, 2007 and August 1, 2007. This paper was recommended by Associate Editor D. Schonfeld.

A. Briassouli is with the Department of Computer and Communication Engineering, University of Thessaly, 38221 Volos, Greece (e-mail: briassou@uth.gr).

N. Ahuja is with the Beckman Institute, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: ahuja@vision.ai.uiuc.edu).

Digital Object Identifier 10.1109/TCSVT.2008.918799

lation terms even in the case of one object motion, and would consequently be unable to extract different velocities. Also, time-frequency distribution based methods deal with translations, but not with more complex, rotational-translational motions. Wavelet domain methods [10] are able to localize motions extracted from phase information, but they suffer from inaccuracies at motion discontinuities. Finally, spatiotemporal filtering approaches [7] are able to successfully estimate and localize translations, but they have a high computational cost, as they require the application of many different velocity-tuned filters to the video, in order to give accurate estimates.

B. Contributions

In order to address these limitations, we propose a novel approach that combines the results of frequency and spatial domain processing in a novel manner. We present a Fourier domain method for motion estimation, that is based on phase correlation techniques, which have proven to give robust results for image registration [9], [25], [26]. However, phase correlation for registration deals with only one moving "object" (the image to be registered), whereas our method is designed for the estimation of multiple translational and rotational-translational motions. Also, in contrast to existing Fourier domain methods for multiple motion estimation [15], [18], [27] that are limited to the case of translational displacements, we propose a system that estimates both translations and rotational-translational motions [28], [29].

For the special case of multiple, temporally constant translations, our problem formulation also allows the estimation of the number of moving objects, as in harmonic retrieval problems [30], [31]. For the case of multiple translations (that may be temporally varying) we provide a novel, elegant method for the simultaneous segmentation of the background and the moving objects, achieved by solving a linear system in the least squares (LS) sense. This solution is both spatially and temporally global, as it uses the FT of all video frames simultaneously. This makes it robust to temporally local occlusions that may occur, for example, when an object hides another over a subsequence of the video. The frequency based segmentation results are further refined by appropriate fusion with spatial domain data following non-ad-hoc, principled approaches, for the accurate object extraction. When the motions are rotational-translational, the segmentation is achieved using only spatial information: the motions that have been estimated in the frequency domain are compensated for, and the result is compared against the original video frame pixel values.

The paper is organized as follows. In Section II, we present the basic problem formulation for multiple translating objects in a video sequence. Section III presents a new method for the extraction of multiple translating objects using frequency domain information. The motion segmentation is formulated as a LS problem in the Fourier domain, and the results are refined by using spatial information, as described in Section IV. In Section V, we examine the problem of multiple complex motions, involving both rotations and translations. Experiments with numerous sequences, synthetic and real, display the effectiveness of the proposed motion estimation and segmentation methods in Section VI. Finally, conclusions and future directions of research are presented in Section VII.

II. Fourier Domain Translation Estimation

In this section, we present the problem formulation for a video that contains multiple objects that are translating against a static background, with a translation that can vary with time. Each frame contains M objects, with luminance $s_i(\bar{r})$, $1 \leq i \leq M$ at pixel $\bar{r} = (x, y)$, and displacement $\bar{d}_i(k)$ from frame 1 to k. The Fourier transform (FT) of each object i is $S_i(\bar{\omega}) = |S_i(\bar{\omega})|e^{j\Theta_i(\bar{\omega})}$, where $\bar{\omega} = [2\pi m/N_1, 2\pi n/N_2]^T$, $m, n \in Z$, is the 2-D spatial frequency, $N_1 \times N_2$ the image size, $|S_i(\bar{\omega})|$ the FT magnitude, and $\Theta_i(\bar{\omega})$ the FT phase. Each video frame is represented in the spatial domain as the sum of a static background, denoted by $s_b(\bar{r})$, and the M objects, $s_i(\bar{r}), 1 \leq i \leq M$, so frame 1 is

$$a(\bar{r}, 1) = s_b(\bar{r}) + \sum_{i=1}^{M} s_i(\bar{r}) + e_{\mathrm{mod}}(\bar{r}, 1) + v_{\mathrm{noise}}(\bar{r}, 1) \quad (1)$$

where $a(\bar{r}, 1)$ denotes the luminance value of frame 1, at pixel location $\bar{r}$. In reality, an additive model for the video frames is not entirely accurate: the background pixel values are not added to the object pixels in the object areas, but they mask their values. Thus, the additive model of (1) includes a modeling error e_{mod}, which comprises of the background pixels that are masked by the objects in one frame, and uncovered in other frames. Since e_{mod} consists of pixel luminance values, it is not random. However, it is unknown, since the object segmentation is unknown, and changes from frame to frame, as different regions of the background are masked or revealed.

The term v_{noise} represents the measurement noise, introduced during the image (or video) acquisition process [32]. The measurement noise, which is a random quantity, has been modeled statistically in the literature [33], [34], and numerous methods have been developed for its removal [35]. We consider that the data being processed either contains negligible amounts of random measurement noise, or that this noise has been removed, as is the case in related work [15], [25], [26], so v_{noise} does not appear in the methods for motion estimation. The rest of the terms in (1) are not random, so in the sequel we will be referring to deterministic quantities for motion estimation. Frame k, $1 \leq k \leq N$ then is

$$a(\bar{r}, k) = s_b(\bar{r}) + \sum_{i=1}^{M} s_i(\bar{r} - \bar{d}_i(k)) + e_{\mathrm{mod}}(\bar{r}, k). \quad (2)$$

It should be noted that the number of objects M is initially unknown, so the only information used in the system are the video frames themselves. For the special case of temporally constant inter-frame translations, the number of moving objects can be estimated *a priori* (Section III-A). For translations that vary with time, as in (2), the number of objects is estimated at a later stage, by simply counting the number of translations that are extracted. The method we propose for translation estimation is based on the phase-shift property of the FT, similarly to phase correlation techniques [9], [15], [26], that have been used for image registration. In our case the problem is more complex, due to the presence of multiple moving objects. The 2-D spatial FT of (2) is

$$A(\bar{\omega}, k) = S_b(\bar{\omega}) + \sum_{i=1}^{M} S_i(\bar{\omega}) e^{-j\bar{\omega}^T \bar{d}_i(k)} + E_{\mathrm{mod}}(\bar{\omega}, k). \quad (3)$$

We then examine the ratio of the FTs of frames 1 and k

$$\Phi_k(\bar{\omega}) = \frac{A(\bar{\omega},k)}{A(\bar{\omega},1)} = \Gamma_b(\bar{\omega}) + \sum_{i=1}^{M} \Gamma_i(\bar{\omega}) e^{-j\bar{\omega}^T \bar{d}_i(k)} + \Gamma_{\mathrm{mod}}(\bar{\omega},k) \quad (4)$$

where

$$\Gamma_b(\bar{\omega}) = \frac{S_b(\bar{\omega})}{A(\bar{\omega},1)}, \quad \Gamma_i(\bar{\omega}) = \frac{S_i(\bar{\omega})}{A(\bar{\omega},1)},$$
$$\Gamma_{\mathrm{mod}}(\bar{\omega},k) = \frac{E_{\mathrm{mod}}(\bar{\omega},k)}{A(\bar{\omega},1)}. \quad (5)$$

In (4), the displacements appear in a sum of weighted exponentials, so the inverse FT of $\Phi_k(\bar{\omega})$ is

$$\varphi_k(\bar{r}) = \gamma_b(\bar{r}) + \sum_{i=1}^{M} \gamma_i(\bar{r}) \delta(\bar{r} - \bar{d}_i(k)) + \gamma_{\mathrm{mod}}(\bar{r},k). \quad (6)$$

Since $\varphi_k(\bar{r})$ is a sum of weighted delta functions, it has M peaks at $\bar{r} = \bar{d}_i(k)$, from which we can estimate the M translations. In practice, digital images (frames) are being processed, so the FT is actually a discrete FT (DFT), which is implemented via the fast FT (FFT). Also, in practice the delta functions are impulse functions, and the image is considered to be periodically replicated in space. Despite this, the peaks around $\bar{d}_i(k)$ are pronounced and provide reliable localization of the actual displacements, as also reported in [9] and [26]. In order to further enhance the peaks' resolution, we examine the squared magnitude of (6), given by

$$|\varphi_k(\bar{r})|^2 = \sum_{i=1}^{M} |\gamma_i(\bar{r})|^2 \delta^2(\bar{r} - \bar{d}_i(k)) + 2\Re\left[(\gamma_b(\bar{r}) + \gamma_{\mathrm{mod}}(\bar{r},k)) \sum_{i=1}^{M} \gamma_i(\bar{r}) \delta(\bar{r} - \bar{d}_i(k)) \right] + |\gamma_b(\bar{r}) + \gamma_{\mathrm{mod}}(\bar{r},k)|^2. \quad (7)$$

From (7) we see that the peaks of $|\varphi_k(\bar{r})|^2$ now originate from terms containing delta functions around $\bar{d}_i(k)$, and also from terms of squared delta functions $\left(\sum_{i=1}^{M} |\gamma_i(\bar{r})|^2 \delta^2(\bar{r} - \bar{d}_i(k))\right)$. The remaining terms $|\gamma_b(\bar{r}) + \gamma_{\mathrm{mod}}(\bar{r},k)|^2$ in (7) do not significantly degrade the accuracy of the peaks in $\varphi_k(\bar{r})$, as they affect all coordinates $\bar{r}$, but do not have an impulsive nature around any specific $\bar{r}$. This fact, combined with the presence of delta functions and squared delta functions around each $\bar{d}_i(k)$ after the nonlinear processing, leads us to expect an enhancement of the peaks around the true displacements. Additionally, the robustness of FT phase-based methods for registration and displacement estimation has been documented in the relevant literature [9], [15], [26], [36], and is also verified by our experimental results, where the displacements are extracted with accuracy in experiments involving real video sequences. Fig. 1 shows the 3-D surface $\phi_{20}(\bar{r})$ for a real video of a car translating in a parking lot (Section VI-C), from frames $1 - 20$. Despite the presence of the background, which creates a dominant peak ($\bar{d}_b = (0,0)$ in Fig. 1) the peak corresponding to the car's translation can be clearly extracted. It is equal to

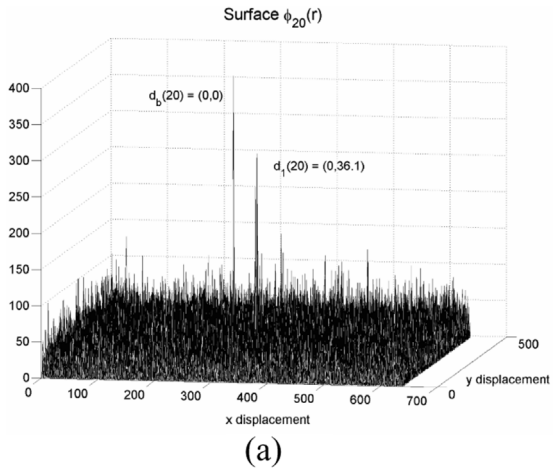


Fig. 1. Surface $\phi_{20}(\bar{r})$ for Parking Lot sequence, frames 1 and 20. Two peaks are dominant, one for the static background one for the car.

$\bar{d}_1(20) = (0, 36.1)$ (Fig. 1), corresponding to a horizontal inter-frame displacement of $d_1 = 1.9$, which is close to the true displacement, of 1.95 (Section VI-C). It should be noted that the center of the coordinate system is considered to be in the center of the surface $\phi_k(\bar{r})$, and the translations are estimated with respect to it.

When the proposed method examines the FT of frames 1 and k, it finds the total displacement $\bar{d}_i(k)$ between those two frames (for each object i), and estimates the corresponding inter-frame displacements as $\bar{d}_i(k)/(k-1)$. If the actual inter-frame displacement between frames $(1,2),(2,3),\ldots,(k-1,k)$ is not constant, i.e., not equal to $\bar{d}_i(k)/(k-1)$, further processing is necessary to extract its correct, time varying values. In practice, we can estimate time-varying object translations either (a) by extracting the translations between frames that are close to each other, so the inter-frame translations are approximately constant, or (b) by finding the translations between two frames, e.g., 1 and k, and then estimating the translations between frames 1, $\lfloor k/2 \rfloor$ and $\lfloor k/2 \rfloor + 1, k$, until the translation estimates in these smaller intervals become equal to each other.[1] This method was used, for example, to determine the inter-frame displacements corresponding to Fig. 1 and to the video of Section VI-C, which in this case are constant.

The proposed translation estimation algorithm is computationally efficient because of the numerous algorithms available for the fast estimation of the FT, namely the FFT, or its variants [21], [22]. For the estimation of the translation between N frames of size $L = N_1 \times N_2$, we need to estimate the FFT of all frames. This has computational complexity of the order $\mathcal{O}(N \cdot L \log_2 L)$, and takes 0.05 seconds per frame, for size 190×420 frames in Matlab, on a 3.4 GHz Pentium IV, while a C implementation would be even faster. For (7), we calculate the squared magnitude of the FFT ratios, which adds L multiplications and additions, and its inverse (IFFT) also adds $\mathcal{O}(L \log_2 L)$ computations, so the resulting complexity is $\mathcal{O}\left(L^2 \log_2 L\right)$. The motion estimation requires localizing the peaks of $\phi_k(\bar{r})$, which entails searching over all L values of this

[1] The notation $\lfloor k/2 \rfloor$ represents the nearest lower integer to $k/2$.

surface (which have already been estimated), and can become computationally costly for large frames. The computational cost of peak-picking [37] can be lowered, and its robustness increased, by searching over smaller windows (bins) in the surface $\phi_k(\bar{r})$ [38], by using methods of discrete stochastic optimization, as in [39], or by using joint time-frequency information [40]. Detailed investigation of these methods is beyond the scope of the current paper, however their use consists of an area of future work. Using peak-picking, for video frames of size 190×420, the translation estimation takes only 0.2 s in Matlab and would take even less time if implemented in C.

III. LS Segmentation for Translation in Frequency Space

In this section, we present a novel method that achieves the simultaneous extraction of a frame's background and of multiple, independently translating objects, using the frequency data. We represent the sequence as in Section II, where each frame's FT is given by (3), for each frequency $\bar{\omega}$ (there are in total $N_1 \cdot N_2$ frequencies) at frame k $(1 \leq k \leq N)$. By stacking the FTs of all N video frames, for any one frequency $\bar{\omega}$, we obtain the linear system

$$\mathbf{A} = \mathbf{GS} + \mathbf{E_{mod}} \tag{8}$$

where $\mathbf{A} = [A(\bar{\omega},1), A(\bar{\omega},2), \ldots, A(\bar{\omega},N)]^T$ are the $N \times 1$ frame FTs, and the $(M+1) \times 1$ vector $\mathbf{S} = [S_b(\bar{\omega}), S_1(\bar{\omega}), \ldots, S_M(\bar{\omega})]^T$ contains the background and the object FTs, at frequency $\bar{\omega}$. $\mathbf{E}_{\mathrm{mod}}$ is the corresponding $N \times 1$ vector of the modeling error, and $\mathbf{G}$ is an $N \times (M+1)$ matrix with the motion information, given by

$$\mathbf{G} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{-j\bar{\omega}^T \bar{d}_1(2)} & \cdots & e^{-j\bar{\omega}^T \bar{d}_M(2)} \\ \vdots & \ddots & \vdots & \\ 1 & e^{-j\bar{\omega}^T \bar{d}_1(N)} & \cdots & e^{-j\bar{\omega}^T \bar{d}_M(N)} \end{pmatrix} \tag{9}$$

for each frequency $\bar{\omega}$. The first line of matrix $\mathbf{G}$ contains ones because there is zero displacement from frame 1 to 1. The linear system of (8) is almost always over-determined, because the number of frames N that are available is usually higher than the number of moving objects M. Once the object translations $\bar{d}_i(k)$ are estimated, the only unknowns in (8) are $\mathbf{S}$ and $\mathbf{E_{mod}}$, at frequency $\bar{\omega}$. The term $\mathbf{E_{mod}}$ depends on the video being examined and cannot be estimated *a priori* (in order to be removed), as it depends on the background areas that are hidden or revealed by the moving objects.

Once the object translations have been estimated, we extract the background and moving objects by solving the linear system of (8) for each one of the $N_1 \times N_2$ frequencies $\bar{\omega}$ separately. We estimate the $N \times (M+1)$ matrix $\mathbf{G}$ of (9) for each $\bar{\omega}$ and for the extracted displacements $\bar{d}_i(k)$. We plug it into (8), along with the $N \times 1$ vector $\mathbf{A}$, containing the N values of each frame's FT at that $\bar{\omega}$, and solve the resulting linear system in a LS) sense [41]. This leads to the vector $\mathbf{S}$, which contains the estimates of the background and moving objects' FTs, for that frequency $\bar{\omega}$. By repeating this process for all $N_1 \times N_2$ frequencies, we obtain the 2-D FT of the background, S_b, and of the M objects S_i. The spatial representation of the background and the objects can be immediately obtained from the inverse FT of S_b and each S_i. It should be noted that this process is not computationally costly, as the LS system is solved using its singular value decomposition (SVD) (Section III-A), for which many computationally efficient methods exist [42]–[44].

1) Estimating the Number of Moving Objects: The video sequence is represented in the spatial domain as a sum of the background, the M moving objects, and a term representing the modeling error. Our method does not require prior knowledge of the number of moving objects, as their number can be extracted simultaneously with the displacements, by simply counting the number of peaks of (7). Since there are multiple objects in the scene, and the background is present, the peaks at $\bar{r} = \bar{d}_i(k)$ are surrounded by side lobes. In practice this does not introduce inaccuracies, because the motion induced peaks correspond to delta functions, which are significantly more pronounced than the noise around each location $\bar{r} = \bar{d}_i(k)$.

For temporally constant inter-frame translations, the number of moving objects can be extracted à priori, before their translations are estimated. In that case, the displacement from frame 1 to k is $\bar{d}_i(k) = (k-1)\bar{d}_i$, where $\bar{d}_i$ is the (constant) inter-frame displacement for object i. Then, $\mathbf{G}$ of (9) becomes Vandermonde [42] for each frequency $\bar{\omega}$, i.e., the elements in each of its rows form a geometric progression, as follows:

$$\mathbf{G} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{-j\bar{\omega}^T \bar{d}_1} & \cdots & e^{-j\bar{\omega}^T \bar{d}_M} \\ \vdots & \ddots & \vdots & \\ 1 & e^{-j(N-1)\bar{\omega}^T \bar{d}_1} & \cdots & e^{-j(N-1)\bar{\omega}^T \bar{d}_M} \end{pmatrix}. \tag{10}$$

The autocorrelation matrix for the video frames' FT's (at a fixed frequency $\bar{\omega}$) $\mathbf{A} = \mathbf{GS}$ can be expressed by $\mathbf{R_A} = \mathbf{GR_S G^H}$, where $\mathbf{R_S}$ is the correlation matrix of $\mathbf{S}$, i.e., the background and objects' FT. It can be shown [31] that the rank of $\mathbf{R}_A$ is equal to the rank of $\mathbf{G}$. For constant inter-frame translations, $\mathbf{G}$ has $M+1$ independent columns, so its rank gives the number of independently moving objects [30]. The singular values of the noiseless data (i.e., $\mathbf{R}_A = \mathbf{GR_S G^H}$) correlation matrix $\mathbf{R_A}$ are $M+1$: $\{\sigma_0^2, \sigma_1^2, \ldots, \sigma_M^2\}$. When random noise is present, in the literature it is assumed to be Additive, White and Gaussian, for simplicity and without loss of generality. The cases of nonwhite Additive Gaussian noise and even general, nonGaussian noise, have also been studied [35] and can be dealt with. In our case, the noise is not random, as it is introduced by the modeling error E_{mod}, which is a deterministic, data-dependent quantity (see analysis below (1)). E_{mod} introduces an error ε_i^2 to each of the singular values σ_i^2 $(0 \leq i \leq M)$, arising when areas of the background are hidden or revealed by the moving objects, so it takes up a smaller part of each video frame than the moving objects, which are of interest. Consequently, its effect on the singular values in (11) does not become detrimental, i.e., the first $M+1$ singular values remain significantly higher than the other $N-M-1$ ones. The SVD of $\mathbf{R_A}$ is

$$\mathbf{R_A} = \mathbf{U_A} \begin{pmatrix} \sigma_0^2 + \varepsilon_0^2 & & & & & \\ & \sigma_1^2 + \varepsilon_1^2 & & & & \\ & & \ddots & & \mathcal{O} & \\ & & & \sigma_M^2 + \varepsilon_M^2 & & \\ & & & & \varepsilon_{M+1}^2 & \\ & & \mathcal{O} & & & \ddots \\ & & & & & & \varepsilon_N^2 \end{pmatrix} \mathbf{V_A^H} \tag{11}$$

where $\mathbf{U_A}$ and $\mathbf{V_A}$ are the eigenvector matrices corresponding to $\mathbf{A}$. The previous analysis focuses on the structure of the autocorrelation matrix for a fixed frequency $\bar{\omega}$. This is because the rank of $\mathbf{R_A}$ is determined only by the object displacements, i.e., it remains the same for any $\bar{\omega}$ (except the trivial cases of $\bar{\omega} = 0$ and $\bar{\omega}^T \bar{d}_i(k) = 2\pi\kappa$, $\kappa \in \mathcal{Z}$). In practice, we estimate the number of moving objects from $\mathbf{R_A}$ for all frequencies $\bar{\omega}$, and keep the estimate of M that is returned by the majority of $\mathbf{R_A}$'s. In order to find the number of moving objects (for each $\mathbf{R_A}$, i.e., for each $\bar{\omega}$), we count the number of singular values in the 75_{th} percentile (i.e., the highest 25% of the singular values). In our experiments, this proved to be a reliable estimate of the number of moving objects. Prior knowledge of the number of moving objects helps verify the number of peaks found in (7), but can also be a parameter of interest in itself if, for example, we are only interested in how many moving objects are present in a scene, and not in the full characterization of their motions.

A. Regularization for Least Squares Motion Segmentation

The LS solution for (8) is the solution that minimizes the mean squared error $\|\mathbf{GS} - \mathbf{A}\|^2$ between the noiseless data $\mathbf{GS}$, and the real data $\mathbf{A}$, so it essentially minimizes the energy of $\|\mathbf{E_{mod}}\|^2$. The LS solution is given by the Singular Value Decomposition (SVD) of $\mathbf{G}$, i.e., $\mathbf{G} = \mathbf{U\Sigma V^H}$

$$\hat{\mathbf{S}} = (\mathbf{G^H G})^{-1}\mathbf{G^H} A = \mathbf{V\Sigma^{-1}U^H A} \tag{12}$$

where $\mathbf{\Sigma^{-1}}$ is a diagonal $(M+1)\times(M+1)$ matrix with values $1/\sigma_i$ for $\sigma_i \neq 0$, and 0 for $\sigma_i = 0$. This is an inverse problem, which is inherently ill-posed, i.e., its solutions are very sensitive to noise and can easily become unstable. This can be seen if we replace $\mathbf{A}$ in (12)

$$\begin{aligned}\hat{\mathbf{S}} &= (\mathbf{G^H G})^{-1}\mathbf{G^H}(\mathbf{GS} + \mathbf{E_{mod}}) \\ &= \mathbf{S} + (\mathbf{G^H G})^{-1}\mathbf{G^H E_{mod}}.\end{aligned} \tag{13}$$

Since we have $(\mathbf{G^H G})^{-1}\mathbf{G^H} = \mathbf{V\Sigma^{-1}U^H}$, (13) becomes

$$\hat{\mathbf{S}} = \mathbf{S} + \mathbf{V\Sigma^{-1}U^H E_{mod}}. \tag{14}$$

For the noiseless case, this would correspond to the correct solution. However, $\mathbf{\Sigma^{-1}}$ is diagonal with values $1/\sigma_i$. From (14) we see that the smaller singular values σ_i, corresponding to the high-frequency components of the video, greatly enhance the term $\mathbf{E_{mod}}$. This appears in the solution for the object vector $\hat{\mathbf{S}}$ in the form of large oscillations, rendering it useless [31], [45].

A well known technique that deals with this instability is the Tikhonov regularization algorithm [46]. Essentially, Tikhonov regularization imposes a constraint on the magnitude of the solution $\mathbf{S}$, to eliminate solutions with magnitudes that go to infinity. Thus, instead of minimizing $\|\mathbf{GS} - \mathbf{A}\|^2$, we minimize $\|\mathbf{GS}-\mathbf{A}\|^2 + \lambda\|\mathbf{S}\|^2$, where λ is a positive constant that controls the size of the solution vector[2]. The LS solution then becomes

$$\mathbf{S} = (\mathbf{G^H G} + \lambda\mathbf{I})^{-1}\mathbf{G^H A} = \sum_{\mathbf{i=1}}^{\mathbf{M+1}} \frac{\sigma_\mathbf{i}}{\sigma_\mathbf{i}^2 + \lambda}\bar{\mathbf{v}}_\mathbf{i}\bar{\mathbf{u}}_\mathbf{i}^\mathbf{H}\mathbf{A} \tag{15}$$

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

so the effect of $\sigma_i \simeq 0$ is dampened by the regularization parameter λ. Note that for $\lambda = 0$, (15) reduces to the solution for the ideal (noise-free) case

$$\mathbf{S} = (\mathbf{G^H G})^{-1}\mathbf{G^H A} = \sum_{\mathbf{i=1}}^{\mathbf{M+1}} \frac{1}{\sigma_\mathbf{i}}\bar{\mathbf{v}}_\mathbf{i}\bar{\mathbf{u}}_\mathbf{i}^\mathbf{H}\mathbf{A} \tag{16}$$

From (15) it is evident that the regularized solution is more stable: when a singular value $\sigma_i \to 0$, the solution does not go to infinity, because of the regularization parameter λ. However, large values of λ also reduce the accuracy of the LS solutions, since then, the regularized singular values $\sigma_i/\sigma_i^2 + \lambda$ will deviate more from their true values $1/\sigma_i$. The LS estimates become smaller because of λ, so in our problem, namely that of object extraction for motion segmentation, the resulting object and background estimates are darker than in reality. Thus, there is a tradeoff in the choice of this regularization parameter. Furthermore, its ideal value cannot be determined à priori, since 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 Section VI), and found that a value around 1 gave consistently good object estimates. We also observed that the LS estimates were robust to small deviations of λ around 1.

B. Effect of Modeling Error

As noted in Section II, the model used to describe each video frame cannot capture the background masking effect caused by the moving objects. This is because the modeling error is directly related to the unknowns in our problem, namely the object segmentation. However, in our experiments (Section VI) we observed that the LS solutions for object segmentation in the frequency domain contain artifacts with a clearly repetitive, periodic pattern. In this section, we show that this periodic component in the LS solution is introduced by the modeling error E_{mod}. For simplicity of analysis, we consider the example of only two objects in frame 1, that have translated by $\bar{d}_i$, $i = \{1, 2\}$, from frame 1 to 2

$$\begin{aligned}A(\bar{\omega}, 2) = S_b(\bar{\omega}) + S_1(\bar{\omega})e^{-j\bar{\omega}^T\bar{d}_1} \\ + S_2(\bar{\omega})e^{-j\bar{\omega}^T\bar{d}_2} + E_{\text{mod}}(\bar{\omega}, 2).\end{aligned} \tag{17}$$

The $N_1 \times N_2$ LS solutions $\hat{S}_b(\bar{\omega})$, $\hat{S}_1(\bar{\omega})$, $\hat{S}_2(\bar{\omega})$ include the effects of $E_{\text{mod}}(\bar{\omega}, 2)$, and are "exact solutions" when there is no modeling error, i.e., they correspond to

$$A(\bar{\omega}, 2) = \hat{S}_b(\bar{\omega}) + \hat{S}_1(\bar{\omega})e^{-j\bar{\omega}^T\bar{d}_1} + \hat{S}_2(\bar{\omega})e^{-j\bar{\omega}^T\bar{d}_2}. \tag{18}$$

If we set $E_{\text{mod}}(\bar{\omega}, 2) = u_b(\bar{\omega}) + u_1(\bar{\omega}) + u_2(\bar{\omega})$, where $u_i(\bar{\omega}) = |u_i(\bar{\omega})|e^{-j\vartheta_i(\bar{\omega})}$, $i = \{b, 1, 2\}$, (17) becomes

$$\begin{aligned}A(\bar{\omega}, 2) &= S_b(\bar{\omega}) + S_1(\bar{\omega})e^{-j\bar{\omega}^T\bar{d}_1} + S_2(\bar{\omega})e^{-j\bar{\omega}^T\bar{d}_2} \\ &\quad + [u_b(\bar{\omega}) + u_1(\bar{\omega}) + u_2(\bar{\omega})] \\ &= \Big(S_b(\bar{\omega}) + u_b(\bar{\omega})\Big) \\ &\quad + \Big(S_1(\bar{\omega}) + u_1(\bar{\omega})e^{j\bar{\omega}^T\bar{d}_1}\Big)e^{-j\bar{\omega}^T\bar{d}_1} \\ &\quad + \Big(S_2(\bar{\omega}) + u_2(\bar{\omega})e^{j\bar{\omega}^T\bar{d}_2}\Big)e^{-j\bar{\omega}^T\bar{d}_2}.\end{aligned} \tag{19}$$

Thus, each LS estimate is related to the desired solution by

$$\hat{S}_i(\bar{\omega}) = S_i(\bar{\omega}) + u_i(\bar{\omega})e^{j\bar{\omega}^T\bar{d}_i}, \quad i = \{b,1,2\}. \tag{20}$$

The error introduced by E_{mod} is $u_i(\bar{\omega}) = |u_i(\bar{\omega})|e^{-j\vartheta_i(\bar{\omega})}e^{j\bar{\omega}^T\bar{d}_i}$, $i = \{b,1,2\}$. The quantity $\vartheta_i(\bar{\omega})$ changes at every frequency $\bar{\omega}$, depending on the unknown modeling error term $u_i(\bar{\omega})$, whereas the term $\bar{\omega}^T\bar{d}_i$ defines a fixed repeated pattern, namely a plane in $\bar{\omega}$-space, specified by the normal vector $\bar{d}_i$. Since this term appears in the FT phase as $e^{j\bar{\omega}^T\bar{d}_i}$, it corresponds to a harmonic, i.e., a sinusoidal component, with constant frequency $\bar{\omega}^T\bar{d}_i$. Indeed, after some algebra, we find that the magnitude of $\hat{S}_i(\bar{\omega})$ is

$$|\hat{S}_i(\bar{\omega})|^2 = |S_i(\bar{\omega})|^2 + |u_i(\bar{\omega})|^2 + 2\Re\left[S_i^*(\bar{\omega})u_i(\bar{\omega})e^{j\bar{\omega}^T\bar{d}_i}\right] \tag{21}$$

$i = \{b,1,2\}$, where the term $2\Re\left[S_i^*(\bar{\omega})u_i(\bar{\omega})e^{j\bar{\omega}^T\bar{d}_i}\right]$ introduces sinusoidal artifacts in the LS solutions, with frequencies determined by the object translations. This agrees with our experimental observations in Section VI, where the LS solutions also contain sinusoidal artifacts that are parallel to the object translations.

IV. Integration With Spatial Estimates

The LS object estimates presented above differ from their true values due to the effect of the background (Section II) and the approximation error introduced by the regularization (Section III-A). To reduce these errors, we fuse the results from the frequency domain with complementary spatial information. It should be noted that the methods that follow are based on statistical properties of the low (random) measurement noise in the video frames. This noise had not been taken into account during the motion estimation stage, as it is very low. However, its statistical properties are useful for the design of the motion segmentation algorithm in a principled, non-ad-hoc manner, that can be applied to any video sequence.

A. Correlation of LS Solution and Original

We propose to fuse the frequency domain solutions to (8) with the spatial data by correlating each LS solution $s_i(x,y)$ $(1 \le i \le M)$, in the spatial domain, with the original frame $a(x,y,1)$. The normalized cross-correlation $c_b(x,y)$ of $s_i(x,y)$ with $a(x,y,1)$, computed over a square neighborhood $\mathcal{N}_b(x,y)$ around pixel (x,y), obtains high values at pixels that belong to object i, and low values elsewhere. As explained in Section II [below (1)], the presence of random measurement noise has been ignored in the other stages of the method, as it is very low. However, this source of randomness can be used to understand the distribution of the correlation coefficients $c_b(x,y)$. Specifically, the low values of random measurement noise are used to interpret each coefficient $c_b(x,y)$ as a sum of random variables, whose individual variances are small (due to the low noise levels). Since each individual variance is small compared to the sum of the variances, the Lindeberg conditions[3] are satisfied [47]. Then, the Central Limit Theorem holds [47], [48], and the distribution of these normalized sums converges

[3]The Lindeberg conditions are the conditions on the variances of a sum of mutually independent random variables that need to be satisfied so the Central limit theorem can be applied. They state that the individual variances σ_i^2 are small compared to $\sum_{i=1}^n \sigma_i^2$. Analytical discussion is provided in [47].

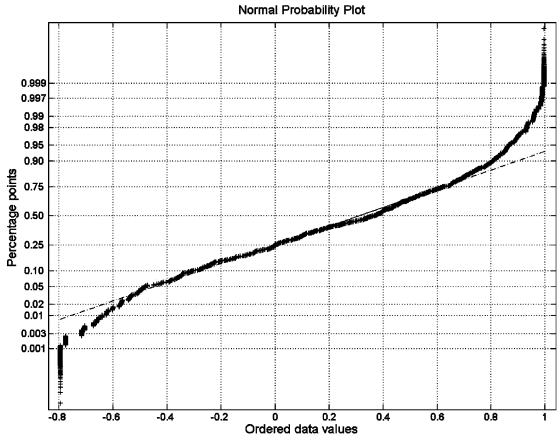


Fig. 2. Normal probability plot of correlation coefficients from the correlation of the LS solution of the moving object in the helicopter sequence with the original frame over 10×10 blocks. The horizontal axis represents the 79800 data values, from which their mean has been subtracted, and the vertical axis the percentage points.

to the Normal distribution, as long as we have a large number of correlation values (a condition which is satisfied in practice, since the correlation values are of the same order as the number of frame pixels). Thus, the coefficients $c_b(x,y)$, estimated over all (x,y), lead to a "correlation map" $C(x,y)$, whose values are considered to approximate a Normal distribution. We make the assumption that the coefficients are ergodic (in the mean and variance), so the distribution's mean is approximated by the sample mean $\mu_b = (1/N)\sum_{x,y} c_b(x,y)$, and its variance by the sample variance $\sigma_b^2 = (1/N)\sum_{x,y}[c_b(x,y) - \mu_b]^2$.

We verified this experimentally by estimating $c_b(x,y)$ between the 190×420 pixels of the first frame and the LS solution for the Helicopter sequence in Section VI-B [Fig. 4(a) and (d)]. The normality of the resulting 79800 coefficients is examined via their Normal Probability Plot (NPP) [49]. The NPP plots the ordered sample values j $(1 \le j \le N)$, after subtracting their mean value from them, against the corresponding "percentage points" $p_j = \Phi^{-1}((j-3/8)/(N+1/4))$ of the normal distribution (Φ is the cumulative distribution function of the normal distribution), which are linearly related for data that is normally distributed [50]. The NPP shown in Fig. 2 indicates that the correlation coefficients for the LS solution and the original frame of the Helicopter sequence indeed approximate a Normal distribution. Similar NPPs resulted for the correlation maps of the LS results with the video frames for the other experiments as well, verifying our assumptions about the normality of their distribution.

For normally distributed correlation coefficients, which have been standardized, $c_s(x,y) = (c_b(x,y) - \mu_b)/\sigma_b$, we consider that the frame pixels with a high correlation value belong to the object extracted from the LS solution

$$\text{Prob}((x,y) \in object) = P(c_s(x,y) > \eta) = Q(\eta) \tag{22}$$

where $Q(x) = 1/\sqrt{2\pi}\int_x^\infty e^{-t^2/2}dt$ gives the tail probability of a Normal distribution. The threshold for the correlation values that are in the highest $\alpha\%$ is then given by $\eta = Q^{-1}(\alpha)$. We make the assumption that the coefficients in the 90th percentile $(\alpha = 0.9)$ belong to the moving object, or equivalently, that

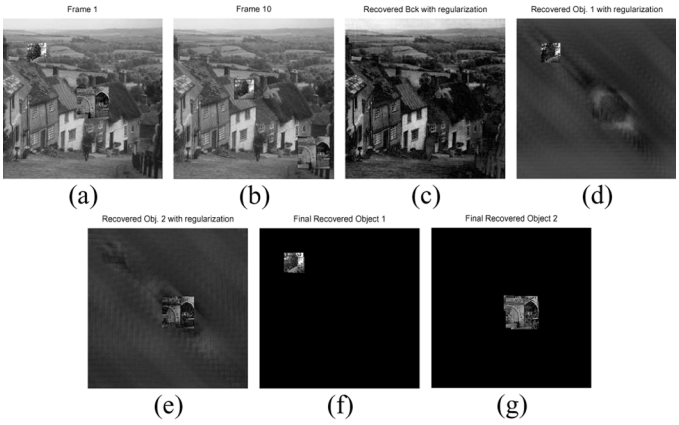


Fig. 3. Synthetic Sequence. (a) Frame 1. (b) Frame 10. LS solutions: (c) Background. (d) Object 1. (e) Object 2. (f) Finally recovered object 1. (g) Finally recovered object 2.

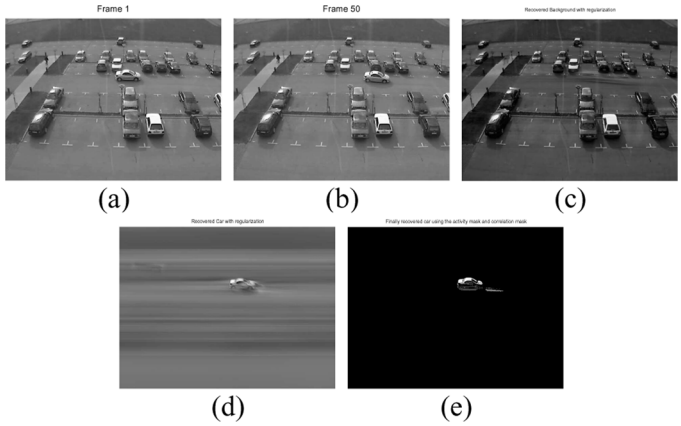


Fig. 5. Parking lot sequence. (a) Frame 1. (b) Frame 50. LS solutions: (c) Background. (d) Car. (e) Finally recovered car after fusion of frequency and spatial information.

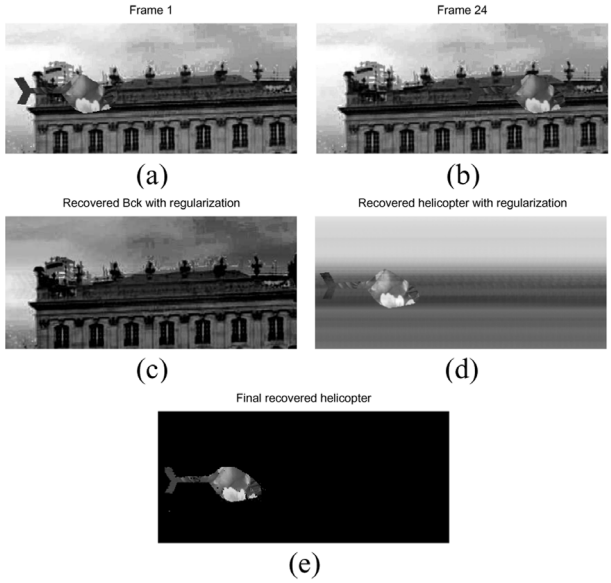


Fig. 4. Helicopter sequence. (a) Frame 1. (b) Frame 24. (c) LS solution for the background. (d) LS solution for the helicopter. (e) Finally recovered helicopter after fusion of frequency and spatial information.

$P(c_s(x,y) > \eta) = \alpha$, for $\alpha = 0.1$. Thus, we have a general, spatial domain method, that combines the frequency domain motion segmentation results with illumination information, in order to refine the extracted object areas.

B. Activity Areas

The correlation process described above combines the frequency-based LS solutions for the moving objects with the spatial data, in order to improve the object estimates. However, the correlation process may not be precise enough, as it may mistake some background pixels for object pixels, and is likely to suffer from block artifacts [Fig. 6(c)]. For this reason, we develop another spatial method that complements the previously derived segmentation results. Frame k is warped by applying the opposite of each motion estimate $\bar{d}_i(k)$ to the pixel positions (x,y), and estimating the luminance values on the warped pixel positions via cubic interpolation. As in Section IV-A, at this stage we take into account the (low) random measurement noise that is present in the video. Then, after warping, frame k becomes $a_W(\bar{r},k)$ and the areas in $a_W(\bar{r},k)$ and $a(\bar{r},1)$ that correspond to object i will differ only by measurement noise, whereas the parts of the frame that have been incorrectly warped will differ by a higher value, that depends on the sequence. A pixel that has been correctly displaced in the warped frame k will have luminance

$$a_W(\bar{r},k) = a(\bar{r},1) + v_{\text{noise}}(\bar{r},k) \tag{23}$$

where $v_{\text{noise}}(\bar{r},k)$ represents the random measurement noise, whereas an incorrectly displaced pixel is

$$a_W(\bar{r},k) = a(\bar{r},1) + w(\bar{r},k) + v_{\text{noise}}(\bar{r},k) \tag{24}$$

where $w(\bar{r},k)$ is an unknown value, introduced by the incorrect warping. The problem of determining if a pixel belongs to object i or not can now be formulated as a binary hypothesis test

$$\begin{aligned} H_0 &: d(\bar{r},k) = v_{\text{noise}}(\bar{r},k) \\ H_1 &: d(\bar{r},k) = w(\bar{r},k) + v_{\text{noise}}(\bar{r},k) \end{aligned} \tag{25}$$

where $d(\bar{r},k) = a_W(\bar{r},k) - a(\bar{r},1)$. Under H_0, $d(\bar{r},k)$ follows the measurement noise distribution, which is commonly modeled as Gaussian [32], but under H_1, its distribution changes significantly, since an unknown, pixel-dependent quantity $w(\bar{r},k)$ is added to $d(\bar{r},k)$. Thus, to determine whether $d(\bar{r},k)$ 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

$$\text{kurt}(y) = E\{y^4\} - 3(E\{y^2\})^2. \tag{26}$$

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 $\text{kurt}(d(\bar{r},k))$ show that the pixel has been incorrectly displaced, and zero (or near-zero) values show that the pixel was correctly displaced, so it belongs to object i.

In order to estimate the kurtosis at each pixel $\bar{r}$, we need to estimate the means of the $d(\bar{r},k)$ over all video frames $1 \le k \le N$. We make the assumption that the differences $d(\bar{r},k)$ form an ergodic process, so their ensemble means can be approximated

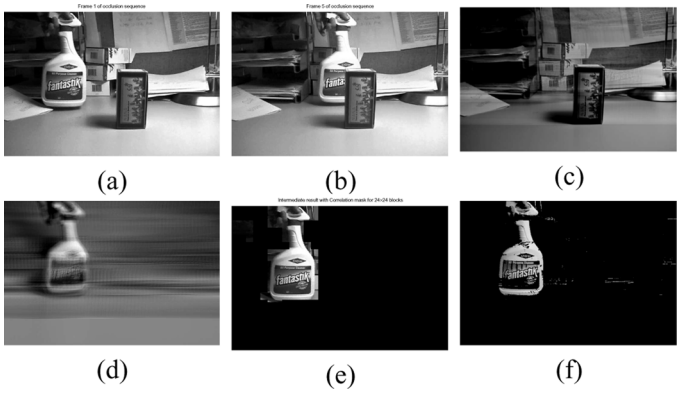


Fig. 6. Bottle sequence. (a) Frame 1. (b) Frame 5. LS solution: (c) background, (d) bottle. (e) Recovered bottle after correlation with spatial data. (f) Finally recovered bottle after fusion of spatial correlation results and activity areas.

by their time averages. Thus, in practice, we estimate the kurtosis of each pixel $\bar{r}$ by

$$\mathrm{kurt}(d(\bar{r})) = \frac{1}{N}\sum_{k=1}^{N} d(\bar{r},k)^4 - 3\left[\frac{1}{N}\sum_{k=1}^{N} d^2(\bar{r},k)\right]^2. \quad (27)$$

As our experiments showed, this leads to a reliable and accurate localization of the activity areas, which can be combined with the results of Section IV-A for the final motion segmentation.

V. Rotational and Translational Motions

Frequency domain representations of images are used to recover more complex transformations, namely rotations combined with translations. Such methods have been developed mainly for image registration [25], [51], [52], so they are designed for the recovery of only one such transformation. We show how they can also be employed for the recovery of multiple inter-frame rotational and translational motions in video sequences. As in Section II, we shall design the motion estimation algorithm under the assumption that measurement noise is negligible, so it is not included in the sequel. In order to make the rotation and translation estimates more accurate, we remove the background from the video frames.

There exist numerous methods for background removal [53], [54], which can be used to obtain a more precise model of each frame. Methods that model the background as mixtures of Gaussians [55], [56] give very good results, but require a training stage and have a high computational cost. Simpler methods, such as median filtering, provide acceptable accuracy, at a very low computational cost and limited memory requirements. Specifically, the background pixels are estimated by the median of each pixel's luminance values over the N frames. When a moving object covers a background pixel $\bar{r}$ over some frames, the luminance of $\bar{r}$ is an outlier, compared to its value in the frames where it was not hidden by the moving object (i.e., its actual background value). The median of a pixel's luminance values over the N frames rejects the outliers and keeps only the background values. This requires that each background pixel is revealed during the video sequence. In the experiments presented in this paper, median filtering based background elimination gives very good results, since all parts of the background are uncovered in different frames during the video. In poor quality sequences, where the video is very jittery and the illumination changes significantly, "local" frame information can also be used, to obtain a more accurate model of the background and remove it more effectively.

A. One Object

Consider a video frame with one object, which undergoes a rotation θ around the center of the frame, and a subsequent translation of $\bar{d} = [d_x, d_y]$ between frames 1 and 2, with no background present. Its luminance $a(x,y,1)$ in the first frame becomes $a(x,y,2) = a(x\cos\theta + y\sin\theta - d_x, -x\sin\theta + y\cos\theta - d_y)$ after its rotation and translation. This object's displacement can also have been caused by a translation of $[x_0, y_0]$, followed by the object's rotation around its center, as follows:

$$\begin{aligned} a(x,y,2) &= a((x-x_0)\cos\theta + (y-y_0)\sin\theta \\ &\quad - (x-x_0)\sin\theta + (y-y_0)\cos\theta) \\ &= a(x\cos\theta + y\sin\theta - (x_0\cos\theta + y_0\sin\theta) \\ &\quad - x\sin\theta + y\cos\theta - (-x_0\sin\theta + y_0\cos\theta)). \end{aligned}$$

This is equivalent to the first formulation, with $[d_x, d_y] = [x_0\cos\theta + y_0\sin\theta, -x_0\sin\theta + y_0\cos\theta]$, so in the sequel we shall consider, without loss of generality, that the object is first rotated (around the center of the frame) and then translated. Its FT for frame 1 is $A(\omega_x, \omega_y, 1)$ and for frame 2

$$\begin{aligned} A(\omega_x,\omega_y,2) = A(\omega_x\cos\theta + \omega_y\sin\theta, -\omega_x\sin\theta + \omega_y\cos\theta) \\ \times e^{-j(\omega_x d_x + \omega_y d_y)}. \quad (28) \end{aligned}$$

From (28), it is evident that the translation only appears in the phase of the FT. Thus, by taking the magnitude of the FT, the effect of the translational motion is eliminated, and only the rotation information is retained. If the FT magnitude of the first frame in the polar domain is expressed as $|A(\rho,\psi,1)|$, then the magnitude of the second frame is

$$|A(\rho,\psi,2)| = |A(\rho,\psi-\theta,1)| \quad (29)$$

for polar coordinates $\omega_x = \rho\cos\psi$, $\omega_y = \rho\sin\psi$. Thus, in the magnitude of the FT in polar coordinates ((29)), the rotation θ becomes a simple translation of the ψ coordinate. Consequently, θ can be recovered from (29) by estimating the translation of ψ, in $|A(\rho,\psi,1)|$, to $\psi - \theta$ in $|A(\rho,\psi,2)|$. The translation is estimated using the method described in Section II. Note that this requires the transformation of the frame from: 1) the spatial domain to the Fourier domain; 2) the extraction of its FT magnitude; 3) the expression of the FT magnitude in polar coordinates; and 4) the FTation of the polar FT magnitude, to extract the rotation angle.

B. Multiple Objects

We consider a video containing multiple objects, that undergo a combination of rotational and translational motions. As explained in Section V-A, we consider without loss of generality, that the objects first undergo rotations around the center of the image frames, and then translations. Also, the background is removed in order to obtain reliable rotation and translation estimates. Then, the additive model $a_{\mathrm{RT}}(x,y,1) = \sum_{i=1}^{M} s_i(x,y)$

for frame 1 is accurate, and, unlike Section II, there is no modeling error in frames k, for $2 \le k \le N$, which can then be written as

$$a_{\mathrm{RT}}(x,y,k) = \sum_{i=1}^{M} s_i(x\cos(\theta_i(k)) + y\sin(\theta_i(k)) - d_i^x(k), \\ -x\sin(\theta_i(k)) + y\cos(\theta_i(k)) - d_i^y(k))$$

where $\theta_i(k)$ gives each object's rotation from frame 1 to k and $\bar{d}_i(k) = [d_i^x(k), d_i^y(k)]$ gives its translation. The FT of (30) is given by

$$A_{\mathrm{RT}}(\omega_x,\omega_y,k) = \sum_{i=1}^{M} S_i(\omega_x\cos(\theta_i(k)) + \omega_y\sin(\theta_i(k)) \\ - \omega_x\sin(\theta_i(k)) + \omega_y\cos(\theta_i(k))) \\ \times e^{-j\left(\omega_x d_i^x(k) + \omega_y d_i^y(k)\right)}.$$

In polar coordinates $\omega_x = \rho\cos\psi$, $\omega_y = \rho\sin\psi$, we have

$$A_{\mathrm{RT}}(\rho,\psi,k) = \sum_{i=1}^{M} S_i(\rho,\psi-\theta_i(k))e^{-jQ(\rho,\psi-\theta_i(k))},$$

where we let $Q(\rho,\psi-\theta_i(k)) = \alpha\cos(\psi-\theta_i(k)) + \beta\sin(\psi-\theta_i(k))$, for $\alpha = \rho\,[d_i^x(k)\cos(\theta_i(k)) + d_i^y(k)\sin(\theta_i(k))]$ and $\beta = \rho\,[d_i^y(k)\cos(\theta_i(k)) - d_i^x(k)\sin(\theta_i(k))]$ (see Appendix A). Setting $P_i(\rho,\psi) = S_i(\rho,\psi)e^{-jQ(\rho,\psi)}$, we obtain

$$A_{\mathrm{RT}}(\rho,\psi,k) = \sum_{i=1}^{M} P_i(\rho,\psi-\theta_i(k)). \tag{30}$$

After these transformations, the multiple rotations in frame k are represented by M "translations" $\theta_i(k)$ of the polar coordinate ψ, which can be easily estimated from its FT phase as in Section II.

Once the rotation angles $\theta_i(k)$ have been extracted, we also need to estimate the corresponding translations $\bar{d}_i(k)$. For each object i $(1 \le i \le M)$, frame k is "derotated" by the angle $\theta_i(k)$ under examination. In the resulting, derotated frame, object i has undergone pure translation, whereas the other objects have undergone a new rotation $\theta_j(k) - \theta_i(k)$ (for $i \ne j$), in addition to translation. After being "derotated" by $\theta_i(k)$ (equivalently, rotated by $-\theta_i(k)$), frame k becomes

$$a'_{\mathrm{RT}}(x,y,k) = s_i\left(x - T_i^x(k), y - T_i^y(k)\right) \\ + \sum_{j=1, j\ne i}^{M} s_j\left(x' - t_j^x(k), y' - t_j^y(k)\right) \tag{31}$$

where, after some algebra, we get

$$\begin{aligned} T_i^x(k) &= d_i^x(k)\cos(\theta_i(k)) - d_i^y(k)\sin(\theta_i(k)) \\ T_i^y(k) &= d_i^x(k)\sin(\theta_i(k)) + d_i^y(k)\cos(\theta_i(k)) \\ x' &= x\cos(\theta_j(k) - \theta_i(k)) + y\sin(\theta_j(k) - \theta_i(k)) \\ y' &= -x\sin(\theta_j(k) - \theta_i(k)) + y\cos(\theta_j(k) - \theta_i(k)) \\ t_j^x(k) &= d_j^x(k)\cos\theta_i(k) - d_j^y(k)\sin\theta_i(k) \\ t_j^y(k) &= d_j^x(k)\sin\theta_i(k) + d_j^y(k)\cos\theta_i(k). \end{aligned} \tag{32}$$

In (31) we see that all objects' translations are affected by the derotation. However, only object i undergoes a pure translation $[T_i^x(k), T_i^y(k)]$. The other objects undergo a translation after rotation by $\theta_j(k) - \theta_i(k)$, i.e., their translations occur on the rotated coordinates (x', y') and their motion is still rotational-translational. Consequently, the method of Section II only applies to the pure translation of object i and will give prominent peaks around $[T_i^x(k), T_i^y(k)]$. The initial $\bar{d}_i(k)$ can be easily estimated from $[T_i^x(k), T_i^y(k)]$, since $\theta_i(k)$ is known. The rotational-translational motions of the other objects may introduce some noise in this translation estimate, but are not expected to lead to significant peaks, as they do not correspond to pure translations. This is also verified in the experiments, where the translations of object i are estimated with accuracy after derotation.

Finally, to correctly correspond the estimated translations to the rotation angles $\theta_i(k)$, we warp and interpolate (with cubic interpolation) frame k by all possible combinations of the extracted rotational and translational motions. We then compare it with frame 1 by the process described in Section IV-B. The correctly warped areas correspond to object i and to the motions $\theta_i(k)$ and $\bar{d}_i(k)$. Note that, in this manner, we simultaneously extract the moving objects as well, i.e., we perform motion segmentation in the spatial domain.

VI. Experiments

We perform experiments with synthetic and real sequences, to demonstrate the translation and rotation estimation capabilities of our method. We also show the Fourier based motion segmentation, and the segmentation results after fusion of the spatial and frequency information. In the real sequences, the ground truth for the translations and rotations is generated via manual feature point matching.

A. Synthetic Sequence

We initially conduct experiments with a synthetic sequence, consisting of two textured squares translating in the x and y directions against a textured background [Fig. 3(a) and (b)] over ten frames. Since their translation is constant, we apply the method of Section III-A, and find three singular values that are over 13, whereas the rest are below 0.5, so we expect to extract two objects. This is verified in the motion estimation process, where there are two peaks corresponding to nonzero translations. The inter-frame translation for the larger square is (8,8) and for the smaller one (6.5,6.5). The FT method estimates their motions as (8.1,8.1) and (6.3,6.3) respectively. We then apply the LS object segmentation. Fig. 3(c) shows that the background is successfully separated from the objects, which are also successfully extracted [Fig. 3(d) and (e)]. Note that all the details in the objects' texture are retained, and their shape is extracted, with no boundary effects. The diagonal "lines" in the area around each extracted object are due to the modeling error, analyzed in Section III-B. The luminance of the recovered objects is also slightly darker than the original objects, due to the regularization process (Section III-A). The LS results are correlated with the first frame [Fig. 3(a)], and the correlation values are compared against a threshold of 0.842 (Section IV-A) in order to localize the moving objects. Finally, the method of Section IV-B is used to further refine the results of the spatial correlation, leading to the final results of Fig. 3(f) and (g).

B. Helicopter

In this experiment, a helicopter [Fig. 4(a) and (b)] translates with a constant velocity 8.8 to the right, over 24 frames. The SVD of its FT's correlation matrix shows there is only one translating object, as its first two singular values are equal to 3.18 and 0.61 (from the background and the helicopter), whereas the others are less than 0.01. The FT estimation method of Section II estimates the translation to be horizontal, to the right, and equal to 8.6, which is close to the ground truth. Fig. 4(c) and (d) show the Fourier domain LS segmentation, where the background and moving object are successfully separated. The shape and texture of the helicopter are retrieved accurately, demonstrating that the LS Fourier-based segmentation can recover objects regardless of their shape, size and texture. In this case, it extracts the irregularly shaped helicopter, and the texture on it, with great precision. There are few horizontal artifacts in the recovered object, because of the modeling error (Section II), and the recovered background and object are darker than the originals due to the regularization. Spatial information is then used to extract the moving object more accurately (Section IV). The LS recovered objects are correlated with the original frame, and the helicopter is localized by keeping the correlation values that are higher than the threshold of Section IV-A, which in this case is equal to 0.757. Activity masks are also used to refine the areas in the frame where the objects are located. This leads to an accurate representation of the moving object, in Fig. 4(e).

C. Parking Lot

A real sequence from a parking lot, with 50 frames of a car translating to the right, is examined [Fig. 5(a) and (b)]. The 75th percentile of the singular values of its correlation matrix contains two values, 4.24 and 1.71, so we expect to have one object. The FT method leads to an inter-frame translation estimate of 1.9 pixels, which is very close to the ground truth estimate of 1.95 pixels. The LS segmentation in Fig. 5(c) and (d) is also very accurate. The car shape and details such as its tires, windows and windshield have been extracted, even though spatial information has not been used yet. These results are correlated with the original frame, with a threshold of 0.73 for the correlation values. They are finally combined with the "activity area" derived from the spatial data, to give the final segmentation of the car, in Fig. 5(e).

D. Bottle With Occlusion

In this experiment, we examine a real sequence with 100 frames of a translating bottle, which is occluded in a third of the frames (over 30 frames) by a box of tea [Fig. 6(a) and (b)]. The highest singular values of the autocorrelation of its FT are 75 and 72, while the others are zero (or almost zero), so we expect only one moving object. The FT method of Section II then estimates its horizontal inter-frame translation to be 8 pixels to the right, with ground truth 9 pixels. The motion estimation is robust to the occlusion, because the bottle is not entirely hidden by the box, so in those frames, its visible part gave enough information for the inter-frame translation to be extracted. Naturally, when a moving object is entirely hidden in a frame, its translation cannot be extracted. Nevertheless, even in such cases, the missing motion information can be inferred from the motion estimates of the rest of the sequence, by applying continuity constraints. The motion estimate is then used to obtain the LS estimates shown in Fig. 6(c) and (d), where, despite the occlusion, the background and the bottle are successfully extracted and separated. This is because we are solving an over-determined linear system [(8)], so data that is lost in some frames because of object occlusion, is compensated for by the redundant information, from other video frames. This demonstrates our method's robustness to common difficulties encountered in real applications. Also, details of the bottle, like its irregular shape and its label, are visible even in its initial, LS estimate. The horizontal lines are caused by the modeling error, as explained in Section III-B. This solution is then combined with the spatial representation of the first frame, as detailed in Section IV-A, by estimating the cross-correlation between Fig. 6(a) and (d). In Fig. 6(e), we show the intermediate result of cross correlation for 24×24 blocks (with a threshold of 0.82). There are blocking artifacts, as expected (Section IV-B), but this result still successfully localizes the moving bottle, and gives a good approximation of its shape. After fusion with the activity areas, an accurate final object estimate is obtained, as shown in Fig. 6(f). There are some small errors in the final segmentation, after fusion with the results of Section IV-B, which are due to the varying illumination in the video sequence (the bottle is translating towards a lamp, so its luminance changes in the later frames).

E. Cars

We examine a sequence with two cars undergoing time-varying translations [Fig. 7(a) and (b)] over 58 frames. The translations are estimated accurately from the FT phase, with deviations that remain below 10% of the ground truth values (Fig. 7(c)). These estimates are used for the LS segmentation of the background and the moving objects, shown in Fig. 7(d)–(f). In Fig. 7(d), the extracted background appears blurry in some areas where the cars were, because it had been occluded by them. In fact, the car on the right did not move during the first frames [Fig. 7(c)], so that part of the background appears most blurred in the LS solution of Fig. 7(d). We then correlate each LS object estimate with the first frame, and the areas with the highest correlation correspond to each object area. The threshold for the correlation values is 0.79 for the first car and 0.68 for the second. Finally, the frames are also warped by the translation estimates and compared with the first frame, as described in Section IV-B. These spatial results are combined, giving the final, accurate object segmentation of Fig. 7(g) and (h). It should be noted that, in this experiment, some blocking artifacts that were not removed by the activity area method are still present at the boundaries of the cars.

F. Traffic

Experiments are conducted with a real traffic sequence, consisting of two cars that are turning [Fig. 8(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. As Fig. 9(a) and (b) show, the rotations are estimated quite accurately using the method of Section V. The frames are derotated by the estimated rotation angles to extract the corresponding translations of each object,

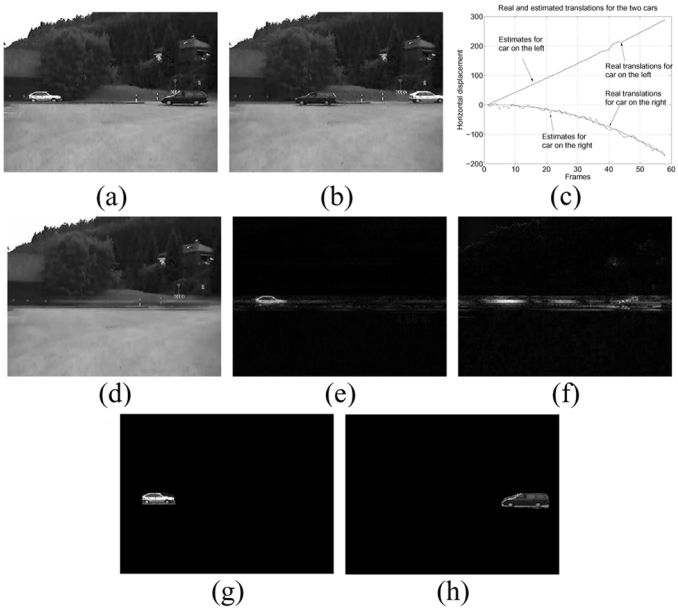


Fig. 7. Cars Sequence: (a) Frame 1. (b) Frame 58. (c) Time varying translation estimates as functions of time. LS Segmentation: (d) Background. (e) Object 1. (f) Object 2. Final Segmentation: (g) Object 1. (h) Object 2.

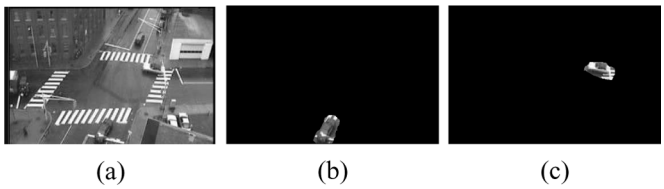


Fig. 8. Real traffic sequence. (a) Frame 10. (b) Reconstructed bottom right car. (c) Reconstructed top right car.

as described in Section V-B. As shown in Fig. 9(c)–(f), the estimated translations in the horizontal and vertical directions are also close to their true values. This sequence has rotational-translational motions, so only spatial data can be used for the segmentation. We apply the method of Section IV-B to warp frame 10 by the motions between frames 1 and 10, and compare the warped frame with the first one. As Fig. 8(b)–(c) show, this leads to the accurate recovery of the bottom right and top right cars.

G. Taxi

We examine a sequence, where a taxi is undergoing a rotation and a car is translating [Fig. 10(a) and (b)] over 20 frames. The log-polar Fourier method of Section V gives an angle of rotation of 9.8° between frames 20 and 1 for the taxi, which is close to our hand-generated ground truth of 9.5°. We then derotate frame 20 (after removing the background), and estimate the taxi's translation to be (2,3), which also is very close to our ground truth of (2.5,2.5). By applying the FT translation estimation method of Section II, we extract the displacement of the car on the left (from frame 1–20) to be (53,12) with ground truth (54,10). Consequently, we see that the proposed method indeed leads to accurate translation and rotation estimates, for a real sequence, that can arise in many practical applications.

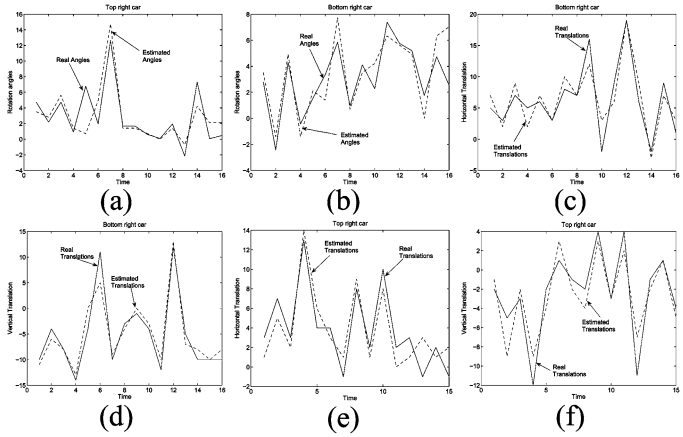


Fig. 9. Real and estimated rotation angles as functions of time for (a) bottom right car and (b) top right car. Real and estimated translations as functions of time for bottom right car (c) horizontal translation and (d) vertical translation. For top right car (e) horizontal translation and (f) vertical translation.

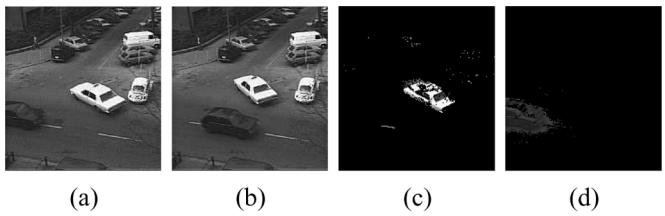


Fig. 10. Taxi sequence. (a) Frame 1. (b) Frame 20. (c) Segmented taxi. (d) Segmented car on the left. Only spatial data has been used for the segmentation because the sequence contains rotation and translations.

Since there is rotational motion in the sequence, we use only spatial data for the object segmentation. We warp frame 20 by each of the motions under examination, and then compare the warped frame with the first one, as described in Section IV-B. This gives the final segmentation results of Fig. 10(c) and (d). Some of the background around the cars has also been recovered, because the color of these cars is very similar to that of the road, making it difficult to discern the two. Nevertheless, the object rotations and translations have been estimated correctly, and as a consequence, the objects are successfully localized.

VII. Conclusion

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 method avoids problems of spatial methods, such as sensitivity to global illumination changes, inaccuracies at object boundaries and motion discontinuities. The use of the FT can be implemented via the FFT, so it is computationally efficient. For the case of purely translational motions, a novel formulation is presented for the object extraction, via the LS solution of a linear system in the Fourier domain. A significant advantage of this approach is that the results are independent of the objects' shapes, texture, motion discontinuities, and are robust to local inaccuracies, such as occlusion, because of the redundancy in the linear system. Due to modeling error and the regularization process used, the LS segmentation results contain some noise, so they are further refined

by using spatial information in a non-ad-hoc manner. The FT motion estimation method has been generalized to more complex motions, involving combinations of rotations and translations. Experiments with both synthetic and real sequences show that the proposed approach can indeed estimate both the translations and rotations reliably. In this case purely spatial methods are used for the object segmentation, which lead to accurate object extraction, as seen in the experimental results. Future directions of research involve the joint use of spatial and frequency data for the analysis of even more complex motions, including random variations, as well as the examination and analysis of the motion of nonrigid bodies.

APPENDIX
DERIVATION OF $Q(\rho, \theta)$

For the case of multiple rotations and translations, we cannot immediately extract the rotation angles from the FT phase in polar coordinates (as in Section V-A). This is because we have a summation of the moving object FTs, so the translation induced phase changes $e^{-j\left(\omega_x d_i^x(k)+\omega_y d_i^y(k)\right)}$ do not disappear when we take the magnitude of the frame's FT. The FT of frame k, after each object i undergoes a rotation of $\theta_i(k)$ and a translation of $\bar{d}_i(k)$, is given by

$$A_{\mathrm{RT}}(\omega_x,\omega_y,k)=\sum_{i=1}^{M} S_i(\omega_x\cos(\theta_i(k))+\omega_y\sin(\theta_i(k)),$$
$$-\omega_x\sin(\theta_i(k))+\omega_y\cos(\theta_i(k)))e^{-j\left(\omega_x d_i^x(k)+\omega_y d_i^y(k)\right)}. \quad (33)$$

The exponential terms need to be expressed as a function of $\psi-\theta_i(k)$, so we let

$$\begin{aligned}&\omega_x d_i^x(k)+\omega_y d_i^y(k)\\&\quad=\rho\left[\cos\psi d_i^x(k)+\sin\psi d_i^y(k)\right]\\&\quad=\alpha\cos(\psi-\theta_i(k))+\beta\sin(\psi-\theta_i(k))\end{aligned} \quad (34)$$

where $\omega_x=\rho\cos\psi$ and $\omega_y=\rho\sin\psi$. We then have

$$\begin{aligned}&\alpha\cos(\psi-\theta_i(k))+\beta\sin(\psi-\theta_i(k))\\&\quad=\alpha\cos\psi\cos\theta_i(k)+\alpha\sin\psi\sin\theta_i(k)\\&\qquad+\beta\sin\psi\cos\theta_i(k)-\beta\sin\theta_i(k)\cos\psi\\&\quad=\left[\alpha\cos\theta_i(k)-\beta\sin\theta_i(k)\right]\cos\psi\\&\qquad+\left[\alpha\sin\theta_i(k)+\beta\sin\psi\cos\theta_i(k)\right]\sin\psi.\end{aligned} \quad (35)$$

From (34) and (35), we have $\rho d_i^x(k)=\alpha\cos\theta_i(k)-\beta\sin\theta_i(k)$, $\rho d_i^y(k)=\alpha\sin\theta_i(k)+\beta\cos\theta_i(k)$, from which we obtain

$$\begin{aligned}\alpha&=\rho\left[d_i^x(k)\cos\theta_i(k)+d_i^y(k)\sin\theta_i(k)\right]\\\beta&=\rho\left[d_i^y(k)\cos\theta_i(k)-d_i^x(k)\sin\theta_i(k)\right].\end{aligned} \quad (36)$$

REFERENCES

[1] G. Chuang and L. Ming-Chieh, "Semiautomatic segmentation and tracking of semantic video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 7, pp. 572–584, Sep. 1998.

[2] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Semantic video analysis for adaptive content delivery and automatic description," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1200–1209, Oct. 2005.

[3] T. Sikora, "MPEG digital video-coding standards," *IEEE Signal Process. Mag.*, vol. 14, no. 5, pp. 82–100, Sep. 1997.

[4] T. Sikora, "The MPEG-7 visual standard for content description-an overview," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 696–702, Jun. 2001.

[5] B. Horn and B. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, pp. 185–203, 1981.

[6] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt, "Performance of optical flow techniques," in *Proc. 1992 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. , pp. 236–242.

[7] D. Fleet and A. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comput. Vis.*, vol. 5, no. 1, pp. 77–104, 1990.

[8] D. Saint-Felix, A. Mohammad-Djafari, and G. Demoment, "Motion detection using 3D-FFT spectrum," in *Proc. 1993 IEEE Int. Con. Acoust., Speech, Signal Process.*, Apr. 1993, vol. 5, pp. 213–216.

[9] E. D. Castro and C. Morandi, "Registration of translated and rotated images using finite Fourier transforms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, no. 9, pp. 700–703, Sep. 1987.

[10] J. Magarey and N. Kingsbury, "Motion estimating using a complex-valued wavelet transform," *IEEE Trans. Signal Process.*, vol. 46, pp. 1069–1084, Apr. 1998.

[11] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Imaging Understand. Workshop*, 1981, pp. 121–130.

[12] J. Domingo, G. Ayala, and E. Dias, "A method for multiple rigid-object motion segmentation based on detection and consistent matching of relevant points in image sequences," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Apr. 1997, vol. 4, pp. 3021–3024.

[13] G. D. Borshukov, G. Bozdagi, Y. Altunbasak, and A. M. Tekalp, "Motion segmentation by multistage affine classification," *IEEE Trans. Image Process.*, vol. 6, no. 11, pp. 1591–1594, Nov. 1997.

[14] D. Zhong and S. Chang, "An integrated approach for content-based video object segmentation and retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 1259–1268, Dec. 1999.

[15] W. Chen, G. B. Giannakis, and N. Nandhakumar, "A harmonic retrieval framework for discontinuous motion estimation," *IEEE Trans. Image Process.*, vol. 7, no. 9, pp. 1242–1257, Sep. 1998.

[16] M. J. Black and P. Anandan, "A framework for the robust estimation of optical flow," in *Proc. IEEE 4th Int. Conf. Comput. Vis.*, May 1993, pp. 231–236.

[17] Y. Weiss and E. H. Adelson, "A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models," in *Proc. IEEE CVPR*, Jun. , pp. 321–326.

[18] L. JacobsonH and Wechsler, "Derivation of optical flow using a spatiotemporal-frequency approach," *Comput. Vis., Graph. Image Process.*, vol. 38, pp. 29–65, 1987.

[19] T. S. Huang and R. Y. Tsai, *Image Sequence Analysis: Motion Estimation*. Berlin, Germany: Springer-Verlag, 1981.

[20] P. Tsai, M. Shah, K. Keiter, and T. Kasparis, "Cyclic motion detection for motion based recognition," *Pattern Recognit.*, vol. 27, no. 12, pp. 1591–1603, 1994.

[21] R. E. Blahut, *Fast Algorithms for Digital Signal Processing*. Reading, MA: Addison-Wesley, 1984.

[22] P. Duhamel and M. Vetterli, "Fast Fourier transforms: A tutorial review," *Signal Process.*, vol. 19, pp. 259–299, 1990.

[23] G. Cunningham and W. Williams, "Fast implementation of time-frequency distributions," in *Proc. IEEE Int. Symp. Time-Frequency Time-Scale Anal.*, Oct. 1992, pp. 241–244.

[24] L. Cohen, "Time-frequency distributions-a review," *Proc. IEEE*, vol. 77, pp. 941–981, Jul. 1989.

[25] B. S. Reddy and B. N. Chatterji, "An FFT-based technique for translation, rotation, and scale-invariant image registration," *IEEE Trans. Image Process.*, vol. 5, no. 8, pp. 1266–1271, Aug. 1996.

[26] H. Foroosh, J. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 188–200, Mar. 2002.

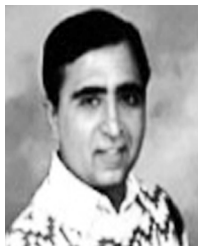
[27] C. E. Erdem, G. Z. Karabulut, E. Yanmaz, and E. Anarim, "Motion estimation in the frequency domain using fuzzy c-planes clustering," *IEEE Trans. Image Process.*, vol. 10, no. 12, pp. 1873–1879, Dec. 2001.

[28] A. Briassouli and N. Ahuja, "Fusion of frequency and spatial domain information for motion analysis," in *Proc. 17th Int. Con. Pattern Recognit.*, Aug. 2004, vol. 2, pp. 175–178.

[29] A. Briassouli and N. Ahuja, "Integrated spatial and frequency domain motion segmentation and estimation," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 244–250.
[30] J.-J. Fuchs, "Estimating the number of sinusoids in additive white noise," *IEEE Trans. Audio*, vol. 36, no. 12, pp. 1846–1853, Dec., 1988.
[31] S. M. Kay, *Modern Spectral Estimation, Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[32] R. Gonzalez and R. Woods, *Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 2002.
[33] S. I. Olsen, "Estimation of noise in images: An evaluation," *CVGIP: Graph. Models Image Process.*, vol. 55, pp. 319–323, 1993.
[34] R. Boie and I. Cox, "An analysis of camera noise," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 6, pp. 671–671, Jun. .
[35] X. Zhang and Y. Li, "Harmonic Retrieval in Mixed Gaussian and Non-Gaussian ARMA Noises," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 42, no. 12, pp. 3539–3543, Dec. 1994.
[36] D. Fleet and A. Jepson, "Stability of phase information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 12, pp. 1253–1268, Dec. 1993.
[37] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
[38] M. R. Every and J. E. Szym, "Separation of synchronous pitched notes by spectral filtering of harmonics," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 5, pp. 18451856–18451856, Oct. 2006.
[39] C. Athaudage and V. Krishnamurthy, "A low complexity timing and frequency synchronization algorithm for ofdm systems," in *Proc. Global Telecommun. Conf.*, Nov. 2002, vol. 1, pp. 244–248.
[40] P. Depalle and T. Hlie, "Extraction of spectral peak parameters using a short-time Fourier transform modeling and no sidelobe windows," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 1997, p. 4.
[41] E. E. Osborne, "On least squares solutions of linear equations," *J. ACM*, vol. 8, pp. 628–636, 1961.
[42] G. Golub and C. V. Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins, 1989.
[43] S. V. Huffel and J. Vandewalle, "The total least squares problem: Computational aspects and analysis," *SIAM Rev.*, vol. 35, no. 4, pp. 660–662, Dec. 1993.
[44] D. Sima, S. V. Hu, and G. Golub, "Regularized total least squares based on quadratic eigenvalue problem solvers," *BIT*, vol. 44, no. 4, pp. 793–812, Dec. 2004.
[45] M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*. Philadelphia, PA, : IOP, 1998.
[46] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. New York: Wiley, 1977.
[47] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York: Wiley, 1966, vol. 1.
[48] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, 1987.
[49] M. B. Wilk and R. Gnanadisikan, "Probability plotting methods for the analysis of data," *Biometrika*, no. 55, pp. 1–17, 1968.
[50] H. A. David, *Order Statistics*. New York: Wiley, 1970.
[51] G. Wolberg and S. Zokai, "Robust image registration using log-polar transform," in *Proc. 2000 Int. Conf. Image Process.*, Sep. 2000, vol. 1, pp. 493–496.
[52] Q. Chen, M. Defrise, and F. Deconinck, "Symmetric phase-only matched filtering of Fourier-mellin transforms for image registration and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 12, pp. 1156–1168, Dec. 1994.
[53] M. Piccardi, "Background subtraction techniques: A review," in *Proc. IEEE Conf. Syst., Man Cybern.*, 2004, pp. 3099–3104.
[54] R. Pless, J. Larson, S. Siebers, and B. Westover, "Evaluation of local models of dynamic backgrounds," in *Proc. CVPR 2003*, Jun. , pp. 1063–1069.
[55] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattren Recognit.*, Jun. 1999, pp. 246–252.
[56] D. H. A. Elgammal and L. Davis, "Nonparametric model for background subtraction," in *Proc. Eur. Conf. Comput. Vis.*, Jun. 2000, pp. 751–767.

Alexia Briassouli (M'07) was born in 1976 in Athens, Greece. She received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, in 1999, the M.S. degree in signal and image processing systems from the University of Patras, Patras, Greece, in 2000 and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 2005.

She is currently a Visiting Lecturer at the Department of Computer Engineering and Telecommunications, University of Thessaly, Greece. Since 2006, she is also a Research Assistant at the Institute of Telematics and Informatics (ITI/CERTH), Thessaloniki, Greece. From 2000 to 2001, she was a Research Assistant at ITI/CERTH. Her current research interests lie in the areas of digital image and video processing, and statistical signal processing, and applications in the areas of communications, multimedia systems, and the semantic web.



Narendra Ahuja (F'98) received th Ph.D. degree from the University of Maryland, Baltimore, in 1979.

He is a Professor in the Department of Electrical and Computer Engineering, and the Coordinated Science Laboratory (CSL) at the University of Illinois at Urbana-Champaign (UIUC), and a full-time faculty member in the Beckman Institute Artificial Intelligence Group. His fields of professional interest are next generation cameras, 3-D computer vision, video analysis, image analysis, pattern recognition, human computer interaction, image processing, image synthesis, and robotics.

Dr. Ahujs's honors include Donald Biggar Willet Professorship, UIUC College of Engineering (1999); On Incomplete List of Teachers Ranked Excellent by Their Students (2002), UIUC Campus Award for Guiding Undergraduate Research-Honorable Mention (1999); Emanuel R. Piore Award, IEEE (1999); SPIE Technology Achievement Award (1998); Fellow, ACM (1996); Fellow, AAAS; Fellow, AAAI; Fellow, SPIE; Fellow, IAPR; Beckman Associate, UIUC Center for Advanced Study (1990–1991, 1998); UIUC University Scholar Award (1985); and the Presidential Young Investigator Award (1984).