Spatial and Fourier Error Minimization for Motion Estimation and Segmentation

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

Abstract

We present a new approach to motion estimation by minimizing the squared error in both the spatial and frequency domains and we show that the spatially global nature of FT leads to a motion estimation error that is much lower than that obtained via spatial motion estimation. On the other hand, spatial analysis is useful for accurate segmentation. We describe a novel, hybrid approach combining the above two estimates of motion and segmentation. We examine the robustness of minimizing the error terms in both domains, both theoretically and experimentally. Experiments with real and synthetic sequences demonstrate the capabilities of the proposed algorithm.

1. Introduction

Traditionally, motion analysis is based directly on the spatial representation of video [1], [6]. In this work, we integrate information from both spatial and frequency domains. By using the frequency data, we avoid the inaccuracies and noise sensitivity of spatial methods [1]. The spatial information helps achieve reliable segmentation of the scene.

Existing work on the fusion of the frequency and spatial data for motion analysis is based on phase correlation and inverse imaging techniques [2]. In this paper, we follow a different approach: we minimize the Fourier and spatial domain errors for motion estimation and segmentation, respectively. We demonstrate, both in theory and experimentally, that the Fourier error is much more robust to additive noise than the spatial error, so it is useful for reliable motion estimation. However, spectral methods suffer from the "localization problem", i.e. their global nature does not indicate to which pixels the motion estimates correspond. Thus, the spatial information is needed to achieve object segmentation.

2. Error Minimization

We first present the spatial squared error to be minimized, with M moving objects $s_i(\bar{r})$ against a dark (0) background[1]. If the first $(N_1 \times N_2)$ video frame has luminance $f(\bar{r},1)$ at pixel $\bar{r} = (x,y)$ and M objects $s_i(\bar{r})$, we have $f(\bar{r},1) = \sum_{i=1}^{M} s_i(\bar{r})$. If each object is displaced by $\bar{r}_i(t)$ from frame 1 to t, frame t is $f(\bar{r},t) = \sum_{i=1}^{M} s_i(\bar{r}-\bar{r}_i(t))$. The squared error in the spatial domain between frame t and frame 1, with the objects displaced by the estimates $\bar{r}_k(t)$, is:

$$J_{spatial}(\bar{r}) = \| \sum_{i=1}^{M} [s_i(\bar{r}-\bar{r}_i(t)) - s_i(\bar{r}-\bar{r}_k(t))] \|^2. \quad (1)$$

Spatial methods minimize $J_{spatial}(\bar{r})$ over all pixels $\bar{r}$ to estimate the correct displacements: the error is minimized at $\bar{r}_k(t) = \bar{r}_i(t)$, $1 \leq i \leq M$. Thus, the motion is estimated at each pixel, and segmentation is achieved simultaneously.

The Fourier transform (FT) of frame 1 is $F(\bar{\omega},1) = \sum_{i=1}^{M} S_i(\bar{\omega})$, and of frame t $F(\bar{\omega},t) = \sum_{i=1}^{M} S_i(\bar{\omega}) e^{-j\bar{\omega}^T \bar{r}_i(t)}$ and the corresponding squared error is given by:

$$J_{freq}(\bar{\omega}) = \| \sum_{i=1}^{M} S_i(\bar{\omega}) [e^{-j\bar{\omega}^T \bar{r}_i(t)} - e^{-j\bar{\omega}^T \bar{r}_k(t)}] \|^2. \quad (2)$$

Similarly to Eq. 1, this error is minimized when $\bar{r}_k(t) = \bar{r}_i(t)$. According to Parseval's Theorem, the total energy

[1] In real applications, where the background is only occasionally occluded by moving objects and is exposed most of the time, it can be detected and zeroed be achieved by simple median filtering.

Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

of a signal (the error in this case) estimated in the spatial domain is equal to the energy of its FT [4]. This is easily proven mathematically, but it is also a statement of the physical principle of the conservation of energy. Thus, the spatial and frequency domain errors are equal, and their minimization gives the same result. However, as we show in the following section, and in the experiments (Sec. 5), these errors behave very differently in the presence of additive image noise, and no longer give the same results.

3. Effect of Noise on Error Minimization

The noisy frames at time t are $f_n(\bar{r},t) = \sum_{i=1}^{M} s_i(\bar{r} - \bar{r}_i(t),1) + n(\bar{r})$, with additive noise $n(\bar{r})$. We assume that the noise follows a zero-mean Gaussian probability density $n(\bar{r}) \sim \mathcal{N}(0,\sigma_n^2)$. This is the most common model for measurement noise in videos [3]. Then, the spatial error of Eq. (1) becomes:

$$\begin{aligned} & J_{spatial,n}(\bar{r}) = \|f(\bar{r},t) - \hat{f}(\bar{r},t) - n(\bar{r})\|^2 \\ = \; & \|\sum_{i=1}^{M}[s_i(\bar{r}-\bar{r}_i(t)) - s_i(\bar{r}-\bar{r}_k(t))]\|^2 + \|n(\bar{r})\|^2 \\ - \; & 2n(\bar{r})(\sum_{i=1}^{M}[s_i(\bar{r}-\bar{r}_i(t)) - s_i(\bar{r}-\bar{r}_k(t))]). \end{aligned} \quad (3)$$

Since the additive noise is a random process, its values change for every "realization" of the video. Thus, we examine the mean of the spatial error, with respect to the noise $n(\bar{r})$:

$$\begin{aligned} & E_n[J_{spatial,n}(\bar{r})] = \\ & E_n\Big[\|\sum_{i=1}^{M}[s_i(\bar{r}-\bar{r}_i(t)) - s_i(\bar{r}-\bar{r}_k(t))]\|^2 \\ + \; & \|n(\bar{r})\|^2 - 2n(\bar{r})(\sum_{i=1}^{M}[s_i(\bar{r}-\bar{r}_i(t)) - s_i(\bar{r}-\bar{r}_k(t))])\Big] \\ = \; & J_{spatial}(\bar{r}) + \sigma_n^2, \end{aligned} \quad (4)$$

for zero mean noise, i.e. $E_n[n(\bar{r})] = 0$, with variance σ_n^2 at each pixel $\bar{r}$. Since the error is "contaminated" by this additive noise at each pixel (Eq. (4)), the noise influence is significant, and is expected to degrade the motion estimation process as its variance σ_n^2 increases. In the frequency domain, Eq. (2) becomes:

$$\begin{aligned} & J_{freq,n}(\bar{\omega}) = \|F(\bar{\omega},t) - \hat{F}_n(\bar{\omega},t)\|^2 \\ = \; & \|\sum_{i=1}^{M} S_i(\bar{\omega})[e^{-j\bar{\omega}^T\bar{r}_i(t)} - e^{-j\bar{\omega}^T\bar{r}_k(t)}] - N(\bar{\omega})\|^2 \\ = \; & J_{freq}(\bar{\omega}) + \|N(\bar{\omega})\|^2 \\ - \; & 2N(\bar{\omega})(\sum_{i=1}^{M} S_i(\bar{\omega})[e^{-j\bar{\omega}^T\bar{r}_i(t)} - e^{-j\bar{\omega}^T\bar{r}_k(t)}]), \end{aligned} \quad (5)$$

where the noise FT $N(\bar{\omega}) = FT[n(\bar{r})]$ follows a zero-mean Gaussian distribution with $E_N[N(\bar{\omega})] = 0$ and $E_N[N^2(\bar{\omega})] = \sigma_n^2\delta(\bar{\omega})$. Then the average frequency domain squared error is given by:

$$E_n[J_{freq,n}(\bar{\omega})] = J_{freq}(\bar{\omega}) + \sigma_n^2\delta(\bar{\omega}). \quad (6)$$

This shows that the noise influence will be present only for $\bar{\omega} = 0$, since it is proportional to $\delta(\bar{\omega})$. Thus, it is expected to affect the mean squared error to a much lesser extent than in the spatial domain. This is verified in our experiments, where we see that the noise affects the spatial error significantly, whereas the frequency domain error remains much more stable, for increasing noise variance.

Thus, the motion estimates acquired from the Fourier space squared error are more robust than from spatial squared error, making the Fourier domain information more useful for video motion estimation.

4. Spatial Fusion for Object Segmentation

Because of the localization problem (Sec. 1), spatial data needs to be used for the motion segmentation. Once the object displacements $\bar{r}_i$ are estimated, all frame pixels are displaced by them, giving $d_i(\bar{r}) = s_i(\bar{r} - \bar{r}_i(t)) - s_i(\bar{r} - \bar{r}_k(t)) - n(\bar{r})$ $(1 \le i,k \le M)$:

$$\begin{aligned} d_i(\bar{r}) &= n(\bar{r}), \quad i = k \\ d_i(\bar{r}) &= m_i(\bar{r}) + n(\bar{r}), \quad i \neq k, \end{aligned} \quad (7)$$

where $m_i(\bar{r}) = s_i(\bar{r} - \bar{r}_i(t)) - s_i(\bar{r} - \bar{r}_k(t))$ for $i \neq k$. When a pixel is correctly displaced, $d_i(\bar{r}) = n(\bar{r})$, which is Gaussian noise $\mathcal{N}(0,\sigma_n^2)$. If it is incorrectly displaced, it follows a non-Gaussian distribution, since $m_i(\bar{r})$ is an unknown, frame-dependent quantity. Thus, to determine the area of object i, we only need to examine the Gaussianity of $d_i(\bar{r})$. This is easily done by estimating the kurtosis $kurt(y) = E\{y^4\} - 3(E\{y^2\})^2$, which is zero when the random variable y is Gaussian [5].

5. Experiments

5.1. Synthetic Sequence

We initially perform experiments with a synthetic sequence of two moving objects (Fig. 1), with different horizontal displacements. By obtaining local minima of $J_{spatial}$ and J_{freq}, we obtain estimates of their motions, as they should be minimized at $\bar{r} = \bar{r}_1$ and $\bar{r} = \bar{r}_2$. Indeed, Fig. 2 shows that the errors are minimized at these two locations, and that the two error curves coincide in the absence of noise, as expected from Parseval's Theorem.

We then examine this method in the presence of additive noise (Fig. 1(b)). The effect of noise is very detrimental on

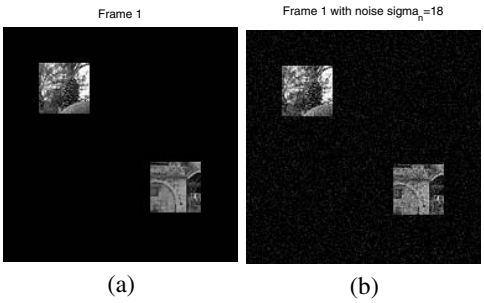


Figure 1. Synth. sequence frame 1: (a) Noiseless. (b) Noise with $\sigma_n^2 = 18$.

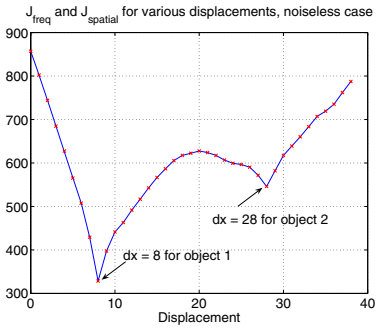


Figure 2. Synth. Sequence: Squared FT and spatial Errors. Noiseless case.

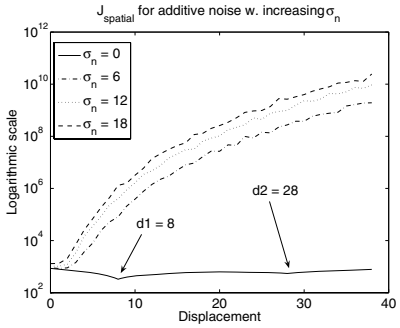


Figure 3. Synth. Seq. $J_{spatial}$ for $0 \leq \sigma_n^2 \leq 19$.

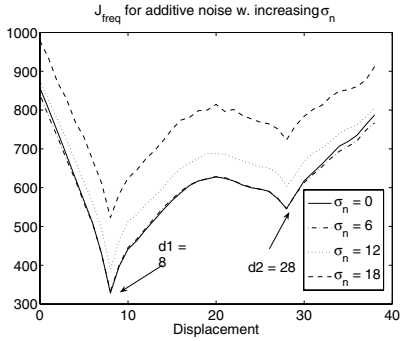


Figure 4. Synth. Seq. J_{freq} for $0 \leq \sigma_n^2 \leq 19$.

the minimization process of $J_{spatial}$ (Fig. 3), as the motion estimates become erroneous even for small amounts of additive noise, The error becomes so high, that we need to plot it on a logarithmic scale. On the other hand, J_{freq} displays remarkable robustness to the same additive noise (Fig. 4).

Accordingly, as argued in the previous sections, we use J_{freq} for the motion estimation.The spatial localization of the motion estimates is performed by using the spatial data. The estimated displacements are used to warp the entire frame at time t, and the warped frame is compared to the first one (Sec. 4), giving accurate object segmentation (Fig. 5).

The segmentation results can also be used to decrease the values of the squared errors. In this experiment, the two moving object areas are separated throughout the sequence, and the corresponding FT and spatial errors are estimated again. In the noiseless case, we find that the resulting spatial and FT are equal to zero. When noise is present, the spatial errors are equal to $2\sigma_n^2$ at each pixel and the FT errors are proportional to $\sigma_n^2 \delta(\bar{\omega})$ (in reality we have an approximation to the spike $\delta(\bar{\omega})$, i.e. it doesn't have infinite height and zero width, as in theory [4]).

5.2. Real Sequence: Tennis Sequence

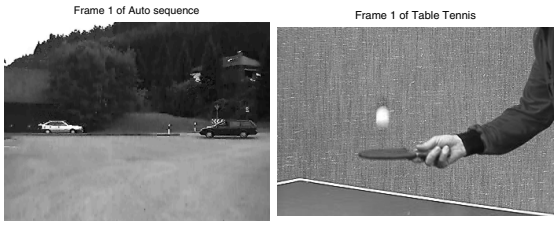
We applied our method on the well-known Table Tennis sequence (Fig. 6(b)). The spatial and frequency space errors are equal in the absence of additive noise. When noise is present ($0 \leq \sigma_n^2 \leq 10$), both display minima at the correct displacement between frames 1 and 5, $d = 55$, but the $J_{spatial}$ increases more with noise and for higher noise values its minima are difficult to detect, whereas J_{freq} has pronounced minima even for the noisiest case.

5.3. Real Sequence: Auto Sequence

Experiments with a real sequence consisting of two cars moving towards each other are also performed (Fig. 6(a)). In the noiseless case $J_{spatial} = J_{freq}$, and they are minimized for the correct displacements, $d_1 = -90$ and $d_2 = 20$ pixels. In Fig. 7 we see the spatial error on a logarithmic scale: it is less robust to noise than the FT error (Fig. 8). Fig. 9 shows that using the spatial data (Sec. 4) we achieve correct object extraction.

(a) (b)

Figure 5. Synth. Sequence: Recovered objects with spatial fusion: (a) Object 1. (b) Object 2.



(a) (b)

Figure 6. Frame 1: (a)Auto. (b)Table Tennis.

6. Conclusions

We presented a novel, hybrid error minimization approach for motion estimation and segmentation. We show that the FT error is very robust to noise and should be used for motion estimation purposes, whereas the spatial error can be used to perform the object segmentation. Experiments for motion estimation and object segmentation, in noiseless and noisy setups, demonstrate the effectiveness and robustness of our proposed algorithm, even when the sequence is severely degraded by noise.

References

[1] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, Dec. 1994.

[2] A. Briassouli and N. Ahuja. Fusion of frequency and spatial domain information for motion analysis. In *ICPR 2004, Proceedings of the 17th International Conference on Pattern Recognition*, volume 2, pages 175–178, Aug. 2004.

[3] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Boston, MA, 2001.

[4] A. Oppenheim and R. Schafer. *Digital Signal Processing*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1975.

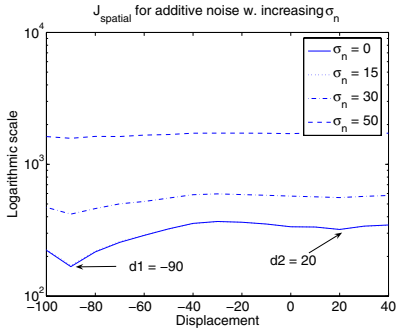


Figure 7. Auto: $J_{spatial}$ for $0 \leq \sigma_n^2 \leq 29$.

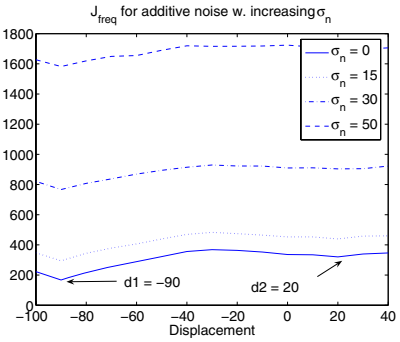
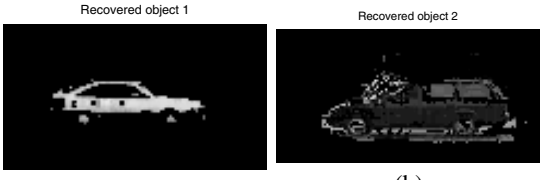


Figure 8. Auto: J_{freq} for $0 \leq \sigma_n^2 \leq 29$.

[5] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, 2nd edition, 1987.

[6] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3:623–638, Sept 1994.



(a) (b)

Figure 9. Segmentation: (a) Car 1. (b) Car 2.