A study of sub-pixel motion estimation using phase correlation

V. Argyriou and T. Vlachos
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford GU2 7XH, United Kingdom
v.argyriou@surrey.ac.uk t.vlachos@surrey.ac.uk

Abstract

We propose a method for obtaining high-accuracy sub-pixel motion estimates using phase correlation. Our method is motivated by recently published analysis according to which the Fourier inverse of the normalized cross-power spectrum of pairs of images which have been mutually shifted by a fractional amount can be approximated by a two-dimensional sinc function. We use a modified version of such a function to obtain a sub-pixel estimate of motion by means of variable-separable fitting in the vicinity of the maximum peak of the phase correlation surface. We demonstrate that our method outperforms, in terms of sub-pixel accuracy, not only other surface fitting techniques but also the state-of-the-art in motion estimation using phase correlation including the technique that motivated our work in the first place. Furthermore our method performs particularly well in the presence of artificially induced additive white Gaussian noise and also offers better motion vector coherence in terms of zero-order entropy.

1 Introduction

Motion estimation is a key element of various video processing tasks such as standards conversion, frame-rate up-conversion, noise reduction, image stabilisation, mosaicing and artefact concealment in archived film sequences. More importantly it is a critical component of video compression systems allowing redundancy reduction in the temporal domain. International standards for video communications such as MPEG$-1/2/4$ and H.$261/3/4$ employ the well-established hybrid two-component architecture, which relies on motion estimation and compensation as well as on the lossy compression of the motion-compensated prediction error. While motion estimation algorithms are non-normative elements in video compression standards many practical encoder implementations use block matching in the data domain.

Recently there has been a lot of interest in motion estimation techniques operating in the frequency domain. These are commonly based on the principle of cyclic correlation and offer well-documented advantages in terms of computational efficiency due to the employment of fast algorithms. Perhaps the best-known method in this class is phase correlation [1] which has become one of the motion estimation methods of choice for

a wide range of professional studio and broadcasting applications [2]. In addition to computational efficiency, phase correlation offers key advantages in terms of its strong response to edges and salient picture features, its immunity to illumination changes and moving shadows and its ability to measure large displacements.

A key performance issue in motion estimation is sub-pixel accuracy. It is self-evident that actual scene motion has arbitrary accuracy and is oblivious to the pixel grid structure resulting from spatial sampling at the image acquisition stage i.e. by CCD arrays or other A/D post-acquisition operations. Theoretical and experimental analyses, such as the work in [3], have established that sub-pixel accuracy has a significant impact on motion compensated prediction error performance for a wide range of natural moving scenes. As a consequence, recent standardisation efforts in video compression have embraced the principle of sub-pixel accuracy for motion estimation and motion compensated prediction. The most popular techniques for subpixel image registration are based on interpolation. Examples are intensity interpolation [4], correlation interpolation [5], [4], phase correlation interpolation [1], [4] and geometric methods [6], [7]. The accuracy of these methods is variable and is largely determined by the characteristics of the core interpolation scheme used in each case.

In this paper we propose a method for obtaining high-accuracy sub-pixel motion estimates using phase correlation. In Section 2 we review the state-of-the-art in sub-pixel motion estimation using phase correlation. In Section 3 we discuss the principle of curve fitting on the phase correlation surface and we introduce variable-separable fitting using a modified sinc function. In Section 4 we present experimental results while in Section 5 we draw conclusions arising from this paper.

2 Sub-pixel motion estimation using phase correlation

In many practical encoder implementations sub-pixel motion estimation is achieved by straightforward extensions to the baseline integer-pixel block-matching algorithm mainly through the use of bilinear interpolation. Interpolation in the data domain is also applicable to frequency domain motion estimation methods such as phase correlation. This can be implemented as a pre-processing step amounting to interpolative upsampling. However, this is not a particularly efficient strategy either in terms of storage requirements or computational complexity to the extent that some of the advantages associated with operating in the frequency domain may be partially cancelled out especially for higher levels of accuracy. For example interpolative upsampling for $1/8-$th pixel accuracy involves the storage and manipulation of complex arrays that are 64 times larger than the original images being correlated. Moreover such an approach cannot provide estimates of true floating-point accuracy, only approximations to the nearest negative power of two. Similar limitations apply to zero padding of phase arrays in the frequency domain, which can be envisaged as an alternative to interpolation in the data domain. For this reason these alternatives have not been considered any further in this paper. To circumvent the above difficulties associated with interpolation alternative approaches have been developed.

Hoge presents in [8] a so-called Subspace Identification Extension method, which is based on the observation that a ”noise-free” phase correlation matrix (i.e. a matrix computed from shifted replicas of the same image) is a rank one, separable-variable matrix. For a ”noisy” phase correlation matrix (i.e. a matrix computed from consecutive frames

of a moving sequence), the sub-pixel motion estimation problem can be recast as finding the rank one approximation to that matrix. This can be achieved by using Singular Value Decomposition (SVD) followed by the identification of the left and right singular vectors. These vectors allow the construction of a set of normal equations, which can be solved to yield the required estimate.

In [9] Stone and co-workers advocate a Frequency-Domain Masking strategy. After obtaining an integer-precision alignment of the input images their method takes steps towards alias cancellation by eliminating certain spectral components of each of the two input images. Elimination is based on two criteria: (i) radial distance of a spectral component from the component located at the origin and (ii) magnitude of a spectral component in relation to a threshold. The latter is dynamically determined as follows. Spectral components are sorted by magnitude and are progressively eliminated starting with the lowest. The authors claim that there exists a range in which the accuracy of the computed motion estimate becomes stable and independent of the degree of progressive elimination. This stability range is indirectly used to determine the required threshold. A plane fitting operation on the frequencies that have survived the above two criteria yields the required motion estimates.

Finally work reported in [10] by Foroosh and Zerubia provides the main motivation for this paper. According to [10] images mutually shifted by a sub-pixel amount can be assumed as having been obtained by an integer pixel displacement on a higher resolution grid followed by subsampling. This assumption allows the analytic computation of the normalised cross-power spectrum as a polyphase decomposition of a filtered unit impulse. The authors demonstrate that the signal power of the resulting phase correlation surface is not concentrated in a single peak but is distributed to several coherent peaks adjacent to each other. The authors further show that this amounts to a Dirichlet kernel, which can be closely approximated by a *sinc* function. This approximation allows for the development of a closed-form solution for the sub-pixel shift estimate.

3 Fitting prototype functions on the phase correlation surface

Fitting prototype functions in the vicinity of the maximum peak of the phase correlation represents another important approach to the problem of sub-pixel motion estimation, which is particularly relevant to our work. To facilitate the discussion on different approaches to fitting we first briefly review phase correlation and introduce the necessary notation.

3.1 Phase correlation

The input to a phase correlation algorithm [11], [12], [1] consists of a pair of images or, more commonly for video compression applications, a pair of co-sited rectangular blocks f_t and f_{t+1} of identical dimensions belonging to consecutive frames or fields of a moving sequence sampled at t, $t+1$. The estimation of motion relies on the detection of the maximum of the cross-correlation function between f_t and f_{t+1}. Because all functions involved are discrete, cross-correlation is circular and for computational efficiency it can be carried out as a multiplication in the frequency domain using fast implementations.

The phase correlation surface is defined as:

$$c_{t,t+1}(k,l) = F^{-1}(\frac{F_t^* F_{t+1}}{|F_t^* F_{t+1}|}) \tag{1}$$

where F_t and F_{t+1} are respectively the two-dimensional discrete Fourier transforms of f_t and f_{t+1}, F^{-1} denotes the inverse Fourier transform and * denotes complex conjugate. The co-ordinates (k_m, l_m) of the maximum of the real-valued array $c_{t,t+1}$ can be used as an estimate of the horizontal and vertical components of motion between f_t and f_{t+1} as follows:

$$(k_m, l_m) = argmax Re\{c_{t,t+1}(k,l)\} \tag{2}$$

3.2 Fitting prototype functions

The above baseline method can only yield integer-precision motion estimates. To achieve sub-pixel accuracy a number of methods have been proposed in the literature and have already been discussed in Section 2 above. In addition to those methods, fitting prototype functions on the phase correlation surface is a particularly powerful approach and for that reason it is discussed to some detail below.

Fitting prototype functions in the vicinity of the maximum peak of the phase correlation surface located at (k_m, l_m) (as computed by using (2) above) is an elegant solution that circumvents many of the problems associated with interpolation while being robust and computationally efficient. This approach was followed in [2] where quadratic (parabolic) fitting was used and also in [13] where both quadratic and Gaussian fittings were considered.

Curve fitting is applied in a variable-separable fashion, vertically as well as horizontally, in the neighbourhood of the maximum of the phase correlation surface. Using the notation in (2) above, prototype functions are fitted to the triplets:

$$\{c_{t,t+1}(k_m - 1, l_m), c_{t,t+1}(k_m, l_m), c_{t,t+1}(k_m + 1, l_m)\} \tag{3}$$

and

$$\{c_{t,t+1}(k_m, l_m - 1), c_{t,t+1}(k_m, l_m), c_{t,t+1}(k_m, l_m + 1)\} \tag{4}$$

i.e. the maximum peak of the phase correlation surface and its two neighbouring values on either side, vertically and horizontally.

The location of the maximum of the fitted function provides the required sub-pixel motion estimate (dx, dy). For example fitting a parabolic function horizontally to the data triplet (3) yields a closed-form solution for the horizontal component of the motion estimate dx as follows:

$$dx = \frac{c_{t,t+1}(k_m + 1, l_m) - c_{t,t+1}(k_m - 1, l_m)}{2(c_{t,t+1}(k_m, l_m) - c_{t,t+1}(k_m + 1, l_m) - c_{t,t+1}(k_m - 1, l_m))} \tag{5}$$

Similarly fitting a Gaussian function horizontally yields:

$$dx = \frac{\log(c_{t,t+1}(k_m + 1, l_m)) - \log(c_{t,t+1}(k_m - 1, l_m))}{2(\log(c_{t,t+1}(k_m, l_m)) - \log(c_{t,t+1}(k_m + 1, l_m)) - \log(c_{t,t+1}(k_m - 1, l_m)))} \tag{6}$$

The fractional part dy of the vertical component can be obtained in a similar way using (4) instead of (3).

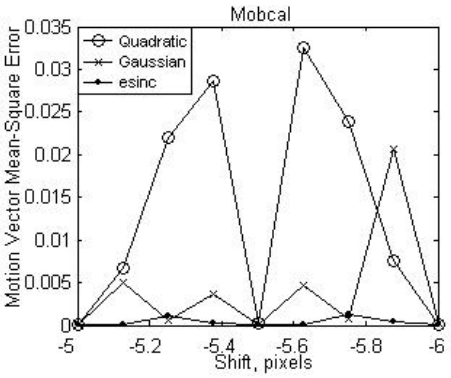

Figure 1: MSE performance comparison for artificially-induced motion.

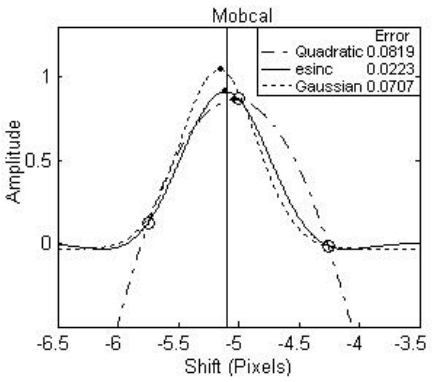

Figure 2: Example of function fitting for artificially-induced motion.

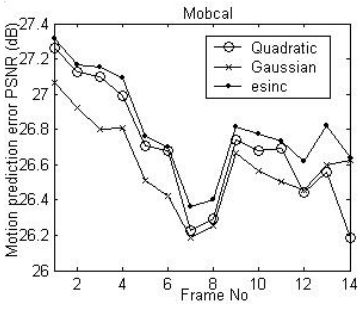

Figure 3: PSNR vs frame number performance comparison with alternative fitted functions for test sequence 'Mobcal'.

3.3 Fitting a modifed *sinc* function

The choice of fitted functions such as parabolic or Gaussian as above seems rather ad hoc and probably justifiable only on the grounds of the mathematical convenience associated with the derivation of a closed form solution for the sub-pixel accuracy motion estimates such as (5) and (6) above. In fact any even and unimodal function would be a sensible choice for such a purpose. In contrast, according to work reported by Foroosh and Zerubia in [10], when the input images are assumed to have been obtained by an integer pixel displacement on a higher resolution grid followed by subsampling, the shape of a phase correlation surface can be analytically shown to correspond to a Dirichlet kernel.

Let $f_t(x,y) = f(x,y)$ and $f_{t+1}(x,y) = f(x-x_0, y-y_0)$ be two successive frames from a moving sequence, which are shifted replicas of each other where (x_0, y_0) are the parameters of the shift. The cross-power spectrum of the downsampled frames by factors of M and N along x and y axes respectively is given by

$$C(u,v) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \frac{F(\frac{u+2\pi m}{M}, \frac{v+2\pi n}{N})}{\sum_{m'=0}^{M-1}\sum_{n'=0}^{N-1} F(\frac{u+2\pi m'}{M}, \frac{v+2\pi n'}{N})} exp(-i(\frac{u+2\pi m}{M}x_0, \frac{v+2\pi n}{N}y_0)) \quad (7)$$

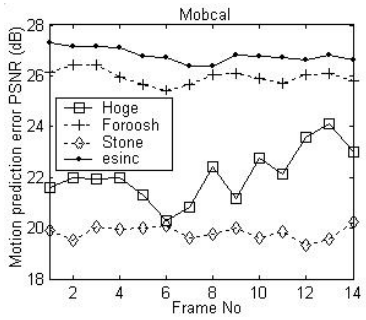

Figure 4: PSNR vs frame number performance comparison with state-of-the-art sub-pixel motion estimation algorithms using phase correlation for test sequence 'Mobcal'.

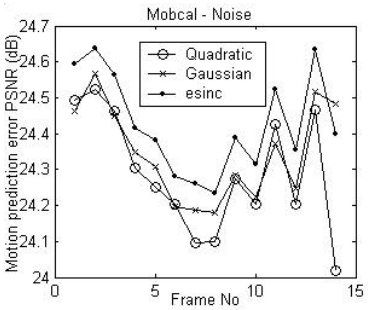

Figure 5: PSNR vs frame number performance comparison with alternative fitted functions for test sequence 'Mobcal' in the presence of artificially-induced additive Gaussian noise.

The discrete inverse Fourier transform of (7) yields a Dirichlet function

$$c(x,y) = \frac{1}{WH} \frac{sin(\pi(Mx - x_0))}{sin(\pi(Mx - x_0)/W)} \frac{sin(\pi(Ny - y_0))}{sin(\pi(Ny - y_0)/H)} \tag{8}$$

where W and H are the frame width and height before downsampling. The authors have further shown that in this case the Dirichlet kernel can be closely approximated by a *sinc* function and they have provided a detailed error analysis demonstrating the validity of this approximation. In our work we take into consideration the validity of the analysis in [10] and combine the findings of that paper on the one hand with the natural advantages of fitting prototype functions on the phase correlation surface on the other. This has led naturally to the consideration of the *sinc* as a candidate prototype function.

While usage of the *sinc* function was observed to have consistently improved motion estimation accuracy, our experiments have further shown that even better performance is achievable by using a modified *sinc* function. This modification consists of applying exponential weighting to a conventional *sinc* function. We call this function *esinc* and we define it as follows:

$$esinc(x) = \exp(-x^2) \frac{\sin \pi x}{\pi x} \tag{9}$$

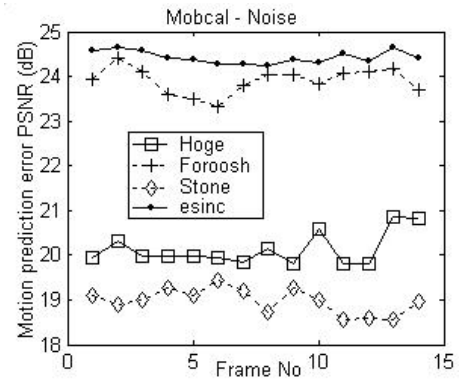

Figure 6: PSNR vs frame number performance comparison with state-of-the-art sub-pixel motion estimation algorithms using phase correlation for test sequence 'Mobcal' in the presence of artificially-induced additive Gaussian noise.

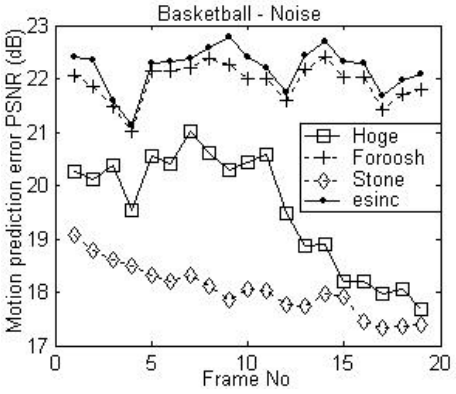

Figure 7: PSNR vs frame number performance comparison with state-of-the-art sub-pixel motion estimation algorithms using phase correlation for test sequence 'Basketball' in the presence of artificially-induced additive Gaussian noise.

One of the obvious consequences of the above exponential weighting is that the ratio of the magnitude of the central lobe to the magnitude of the side lobes is slightly increased. Also the central lobe becomes slightly narrower compared to the central lobe of the unweighted *sinc* function. These subtle shape changes seem to yield a function which is better tuned to approximate phase correlation surfaces obtained by using real video data. We parameterize *esinc* with regard to magnitude, scale and shift changes i.e. we consider a function of the form $Aesinc(B(x-C))$ and compute parameters A, B and C that provide the best fit to the data triplets in (3) and (4). It should be noted that in contrast to the parabolic and Gaussian cases, the functional form of *esinc* does not allow for the development of a closed form solution. Therefore computation of parameters A, B and C is achieved by least squares fitting.

Let us consider the one-dimensional case (x-axis) and denote by $c(x_m)$ the maximum peak of the correlation surface and by $c(x_{m-1})$and $c(x_{m+1})$ the nearest neighbouring values on either side of $c(x_m)$. The unknown parameters A_x, B_x and C_x can be approximated by solving numerically the following least-squares minimisation problem:

$$(A_x,B_x,C_x) = argmin \sum_{x_i=x_{m-1},x_m,x_{m+1}} [c(x_i) - A_x exp(-(B_x(x_i-C_x))^2)\frac{sin(\pi(x_i-C_x))}{\pi(x_i-C_x)}]^2 \tag{10}$$

The computation of parameters A_y, B_y and C_y along the y-axis can be obtained in a similar way. Given the very low dimensionality of the fitting task this is not envisaged to present a problem in relation to implementation. One consequence of least squares fitting is that the fitted *esinc* function may not include one or more of the original data points in (3) and (4). However this did not prove to have an adverse effect on the accuracy of the obtained sub-pixel estimates. Finally it should be noted that in our experiments we have constrained the parameter that controls scale changes so that at most one negative side lobe on either side of the central lobe of the fitted function occurs inside the data window

in which fitting is carried out.

4 Experimental results

In our experiments we used the well-known broadcast resolution (720 x 576 pixels, 50 fields per second) MPEG test sequences 'Mobcal' and 'Basketball'. Only the luminance component was considered and to avoid complications due to interlacing, only even-parity field data were retained.

MSE	*Mobcal*	*Basketball*	*Mobcal noise*	*Basketball noise*
esinc	146.222	333.040	252.774	416.383
Quadratic	150.378	338.831	261.093	424.205
Gaussian	153.461	338.992	257.593	420.563

Table 1. Time-averaged MSE performance comparison with alternative fitted functions.

MSE	*Mobcal*	*Basketball*	*Mobcal noise*	*Basketball noise*
esinc	146.222	333.040	252.774	416.383
Foroosh	178.208	353.179	285.569	441.45
Stone	731.050	920.134	889.18	1083.8
Hoge	445.557	562.100	681.971	783.827

Table 2. Time-averaged MSE performance comparison with state-of-the-art sub-pixel motion estimation algorithms using phase correlation.

4.1 Artificially-induced motion

As a first simple test, estimation accuracy was assessed when global motion was artificially induced and hence known a priori. The estimation area was limited to the central 512x256 pixels of the retained even-parity fields. Sub-pixel motion was simulated using bi-linear interpolation to displace test image 'Mobcal' at desired positions. Various positive as well as negative sub-pixel shifts were used whose accuracy was limited to $1/8$ of a pixel. The mean-square error of the estimation for a sample range of horizontal sub-pixel displacements is plotted in Fig. 1 for each of the three fitting schemes namely quadratic, Gaussian and the *esinc*. Similar results were obtained for vertical sub-pixel displacements and for other displacement ranges. An example of fitting quadratic, Gaussian and the *esinc* functions along with the corresponding absolute error values is shown in Fig. 2. In this case the true value of the displacement is -5.125 pixels and is indicated in the figure by a vertical line.

4.2 Local motion

Next we turn our attention to block-based estimation of local motion occurring in co-sited blocks in consecutive even-parity fields of the test sequences. This is arguably a more relevant performance assessment framework due to the proliferation of block-based motion estimation schemes in hybrid, two-component video compression architectures.

A common design issue is determining the optimum block size. Our experimental evidence suggested that good results could be achieved for a block size of 32x32 pixels. This agrees with the findings of other workers in the area [2] and was used for the remainder of our experiments. The central 672x256 pixel area was retained and then further partitioned

to 32x32-pixel blocks. Performance was assessed by applying motion compensation using the estimated motion parameters and computing the Peak-Signal-to-Noise Ratio (PSNR) of the motion compensated prediction error.

We present two sets of results. First performance comparisons are made using the three fitting schemes namely quadratic, Gaussian and the *esinc*. Results are shown as PSNR values plotted against frame number in Fig 3 for test sequence 'Mobcal'. Then *esinc* is separately compared against the state-of-the-art algorithms in [8], [9] and [10] reviewed in this paper and the results are shown in Fig 4. The corresponding MSE values averaged over all the processed frames (first 20 frames) are shown in the first two data columns of Tables 1 and 2 respectively. In those tables the lowest MSE values have been underlined. Our results show that the use of *esinc* offers a small but consistent advantage over the two competing fitting schemes. More importantly the use of *esinc* consistently outperforms the state-of-the-art methods by as much as 7 dB.

We tested the performance of the proposed scheme in relation to artificially induced additive white Gaussian noise of varying power. Results are shown in Figs 5, 6 and 7 for a PSNR of 30 dB. The corresponding MSE values averaged over all the processed frames are shown in the last two data columns of Tables 1 and 2 respectively. Again, the lowest MSE values have been underlined. Our results underline the superiority of the proposed scheme both in relation to the two fitting alternatives as well as the state-of-the-art methods. In the latter comparison, differences by as much as 4.5 dB can be observed in favour of the proposed scheme. Similar trends were observed for other levels of artificially induced noise in the range of 20-30 dB.

4.3 Motion vector coherence

The quality of block-based motion estimation was also assessed from a different angle by considering the coherence of the resulting motion vector field per processed field. While coherence may appear to be a rather vague term it is intuitively associated with properties such as smoothness, consistency and compressibility of the resulting motion vector field. Compressibility is especially relevant to video coding applications and in particular to hybrid, two-component schemes employing motion-compensated prediction. For this purpose, and in common with other studies in the field, we used as a simple criterion the zero-order two-dimensional entropy of the estimated motion vector field. The results expressed in bits per vector component are shown in Table 3 and further confirm that the proposed technique consistently outperforms the state-of-the-art methods for all cases.

ENTROPY	*Mobcal*	*Basketball*	*Mobcal noise*	*Basketball noise*
esinc	3.3043553	2.7951233	3.4380093	3.2708717
Foroosh	3.6538312	3.6654390	3.7991231	4.0524807
Stone	6.1250425	6.2682382	6.3103670	6.5689116
Hoge	3.8271115	3.1946844	4.3081350	4.4066632

Table 3. Time-averaged zero-order entropy performance comparison with state-of-the-art sub-pixel motion estimation algorithms using phase correlation.

5 Conclusions

In this paper we presented a function fitting approach towards obtaining high-accuracy sub-pixel motion estimates using phase correlation. Our approach was based on fitting a suitably modified sinc type of function to data located in the neighbourhood of the maximum peak of the phase correlation surface. Fitting was performed in a variable-separable way, horizontally and vertically and the location of the maximum of the fitted function determined the required sub-pixel motion estimate. Our results have shown that the proposed method outperforms other popular function fitting alternatives as well as the state-of-the-art in sub-pixel motion estimation using phase correlation. Performance gains were observed both in terms of measured motion-compensated prediction error as well as zero-order entropy of the resulting motion vector fields. The latter was used as an alternative indicator of motion estimation quality related to vector field coherence. Performance gains were observed consistently for a variety of scenarios including artificially-induced global motion, actual inter-frame motion estimated in a block-based fashion as well as artificially-induced additive white Gaussian noise.

6 References

[1] J. J. Pearson, D. C. Hines, S. Goldman, and C. D. Kuglin, "Video rate image correlation processor", Proc. SPIE , Vol. 119, Application of Digital Image Processing, 1977.
[2] G. A. Thomas, "Television motion measurement for DATV and other applications", BBC Res. Dept. Rep., No. 1987/11.
[3] B. Girod, "Motion-compensating prediction with fractional-peel accuracy", IEEE Trans. Comm., vol. 41, no. 4, pp. 604–612, Apr. 1993.
[4] Q. Tian and M. N. Huhns, "Algorithms or subpixel registration," Comput. Vis. Graph. Image Process., vol. 35, pp. 220223, 1986.
[5] V. N. Dvorchenko, "Bounds on (deterministic) correlation functions with applications to registration," IEEE Trans. PAMI., vol. PAMI-5, no. 2, pp. 206213, 1983.
[6] C. A. Bernstein, L. Kanal, D. Lavin, and E. Olson, "A geometric approach to subpixel registration accuracy," Comput. Vis. Graph. Image Proc, vol. 40, pp. 334360, 1987.
[7] A. Goshtasby, G. C. Stockman, and C. V. Page, "A region-based approach to digital image registration with subpixel accuracy," IEEE Trans. Geosci. Remote Sensing, vol. 24, no. 3, pp. 390399, 1986.
[8] W. S. Hoge, "A subspace identification extension to the phase correlation method", IEEE Trans. Med. Imag., vol. 22, no. 2, pp. 277-280, Feb. 2003.
[9] H. S. Stone, M. T. Orchard, E-C. Chang and S. A. Matrices, "A fast direct Fourier-based algorithm for subpixel registration of images", IEEE Trans. Geo. and Rem. Sensing, vol. 39, no. 10, pp. 2235-2243, Oct. 2001.
[10] H. Foroosh, J. Zerubia and M. Berthod, "Extension of phase correlation to sub-pixel registration" , IEEE Trans. Image Processing, vol. 11, no. 3, pp. 188-200, 2002.
[11] C. D. Kuglin and D. C. Hines, "The phase correlation image alignment method," in Proc. Int. Conf. Cybernetics Society, 1975, pp. 163165.
[12] A. Papoulis, Signal Analysis, New York: McGraw-Hill, 1977.
[13] I. E. Abdou, "Practical approach to the registration of multiple frames of video images", Proc. SPIE Conf. Vis. Comm. Image Proc., Vol. 3653, pp. 371-382, San Jose CA, Jan. 1999.