16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

DECAYING EXTENSION BASED PHASE CORRELATION FOR ROBUST OBJECT LOCALIZATION IN FULL SEARCH SPACE

Javed Ahmed and M.Noman Jafri

Image Processing Center, NUST Military College of Signals, Rawalpindi, 46000, Pakistan
email: {javed,mnjafri}@mcs.edu.pk

ABSTRACT

Phase correlation is an efficient tool for precisely localizing an object of interest in an image. However, its performance is severely deteriorated, if: (1) the images contain significant mismatch between the intensity levels of the pixels at the opposite boundaries, (2) the object in the search image is slightly rotated or scaled, or (3) the search image is significantly noisy or blurred. Some of these problems have been addressed by previous techniques, but at the cost of an spatial constraint that the object is fairly inside some central region in the search space. Therefore, we propose an efficient and effective preprocessing technique, that extends the search image and the template with new pixels having smoothly decaying values. It is demonstrated that the proposed method outperforms two recent techniques in localizing an object of interest in real images, especially when the object lies away from the central region in the search space.

1. INTRODUCTION

Object localization in an image is a critical step in the image registration and the visual tracking applications. It can be carried out efficiently by phase correlating a template (i.e. image of the whole object or its salient part) with the search image, and finding the position of the highest peak (i.e. maximum value) in the correlation response. The phase correlation (PC) is significantly robust to illumination variation and offers a normalized response with a sharp peak at the best match location [5, 7, 8, 10, 6]. The PC between an $M \times N$ search image, s, and a $P \times Q$ template, t, is computed as:

$$c = \mathfrak{F}^{-1}\left(\frac{\mathfrak{F}(s)}{|\mathfrak{F}(s)|} \cdot \frac{\mathfrak{F}(t)^*}{|\mathfrak{F}(t)|}\right) \qquad (1)$$

where $\mathfrak{F}(\cdot)$ and $\mathfrak{F}^{-1}(\cdot)$ are the 2-D DFT (discrete Fourier transform) and the 2-D inverse DFT functions, respectively, the $|\cdot|$ operator computes the magnitude of every complex number in its input matrix, the asterisk (*) is the complex-conjugate operation, and all the division and multiplication operations are performed element-by-element. The normalization of the DFTs of s and t is performed to equate the magnitude of all the complex numbers in the frequency domain to 1, because the information of the translation of the object in the image is contained only in the phase of the cross power spectrum. This normalization is usually called *whitening* of the signals [5].

The DFT considers a *discrete* image as cyclically repeated in every direction. For example, three cycles of a *house* image[1] in every direction are shown in Fig. 1(middle left). This phenomenon causes (1) the well-known *wrap-around effect* in the PC response [1, 8, 4], and (2) the discontinuities at the boundaries of the image accompanied with the corresponding spurious high frequency components in the frequency domain as depicted by the horizontal and vertical lines at the top and left sides of the power spectrum shown in Fig. 1(bottom left). The wrap-around effect in the PC response is eliminated by zero-padding the images under consideration up to the size $(M+P-1) \times (N+Q-1)$, as shown in Fig. 1(top right). However, the cyclically repeated zero-padded image contains even more discontinuities, as shown in Fig. 1(middle right), and the corresponding severe high-frequency artifacts, as shown in Fig. 1(bottom right). These artifacts increase the height of the false peaks and decrease that of the true peak in the PC response, resulting in incorrect localization of the object [10], e.g. a window (Fig. 2).

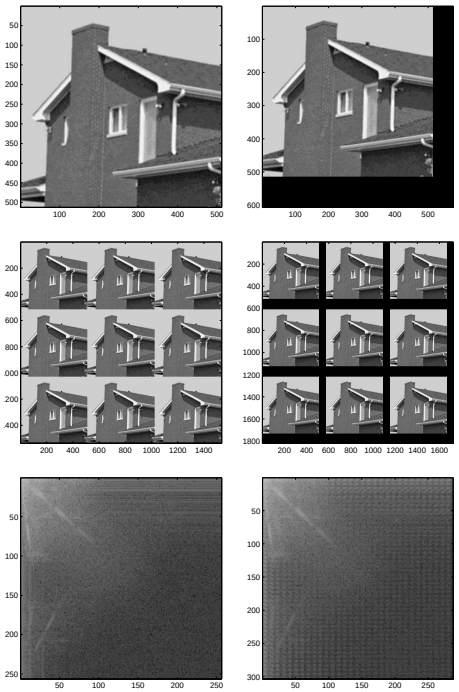

Figure 1: (*top left*) Original *house* image. (*top right*) zero-padded image. (*middle left*) Cyclically repeated image. (*middle right*) Cyclically repeated zero-padded image. (*bottom left*) Upper-left quadrant of the spectrum of the original image. (*bottom right*) Upper-left quadrant of the spectrum of the zero-padded image. In the spectrums, the top-left element is the DC component and the bottom right element is the highest frequency component.

[1] www.imageprocessingplace.com/root_files_V3/image_databases.htm.

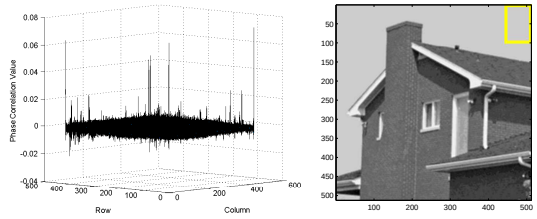

Figure 2: (*left*) Surface showing various false peaks in the response of the PC between the zero-padded *house* image and the zero-padded template of the window of the house. (*right*) The top-left vertex of the overlaid rectangle corresponds to the highest peak at (1, 449) having magnitude only 0.047. Thus, the standard PC could not localize the window.

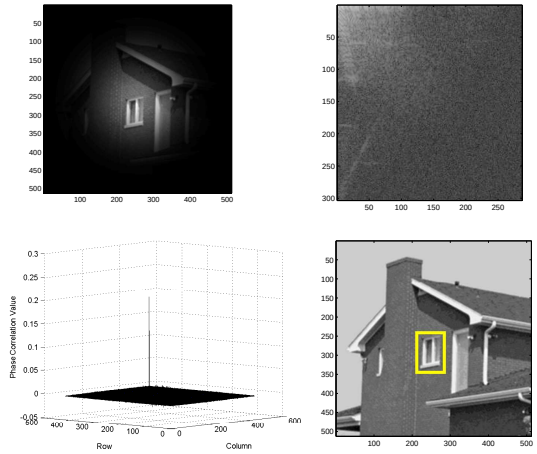

Figure 3: (*top left*) Blackman modulated *house* image. (*top right*) Top-left quarter of the spectrum of the zero-padded Blackman modulated image. (*bottom left*) Surface showing the response of the PC between the modulated *house* image and template of the window of the house. (*bottom right*) the top-left vertex of the overlaid rectangle corresponds to the highest peak at (246, 216) having magnitude 0.21.

In order to eliminate the discontinuities, Stone et al. [10] proposed to modulate (i.e. multiply element-by-element) the original images with a Blackman window [9] before zero-padding. The Blackman window smoothly attenuates the value of the pixel depending on its position relative to the image center, as shown in Fig. 3(top left). Thus, the artifacts in the high-frequency regions in the spectrum are eliminated [Fig. 3(top right)], the false peaks in the PC response are drastically attenuated [Fig. 3(bottom left)], and the object (i.e. window of the house) is accurately localized [Fig. 3(bottom right)]. However, as a side effect, the Blackman window hides all the non-central regions of the image [Fig. 3(top left)] and distorts the low-frequency components in the spectrum [Fig. 3(top right)]. Thus, the template of an object lying away from the central region (e.g. top portion of the house) does not match with the object as good as it does with some clutter at the central region. As a result, the PC response contains multiple false peaks at the central region having magnitude higher than that of the true peak at the actual location [Fig. 4(left)]. Thus, the actual object (i.e. top portion of the house) is not localized [Fig. 4(right)].

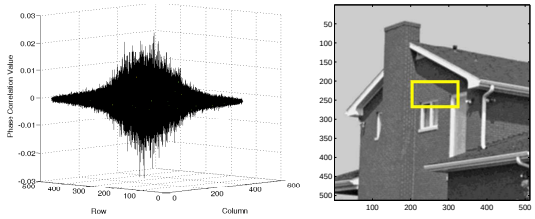

Figure 4: (*left*) Surface showing various false peaks in the response of the PC between the zero-padded Blackman modulated *house* image and the zero-padded template of the top portion of the house. (*right*) The top-left vertex of the overlaid yellow rectangle corresponds to the highest peak at (207, 209) having magnitude only 0.0248.

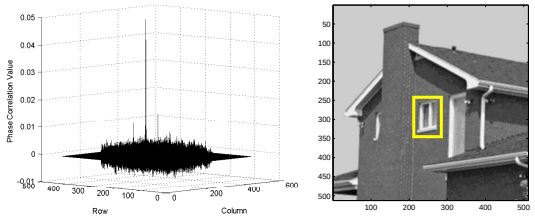

Figure 5: (*left*) Surface showing correct peak in the response of the projection operator based PC between the zero-padded original *house* image and the zero-padded original template of the window of the house. (*right*) The top-left vertex of the overlaid rectangle corresponds to the highest peak at (246, 216) having magnitude only 0.049.

Recently, Keller et al. [6] suggested performing post-processing instead of the pre-processing. That is, they phase correlated the zero-padded original images and used a projection operator that zeros the correlation result beyond a rectangular support region (typically 21×21 size) at the center. This technique is also successful in case the object is present within the central region in the search space, e.g. the window of the house (Fig. 5). However, if the object lies outside the support region at the center, it is not localized, e.g. the top portion of the house (Fig. 6), even when we used large support region of size half of the correlation surface.

In order to address the limitations of the two previous techniques, we propose an efficient and effective method that slightly extends the search image and the template with new pixels having gradually decaying values before zero-padding operation. The proposed pre-processing technique eliminates the mismatch between the boundary pixels in the image without degrading the actual scene, and improves the phase correlation to localize the object robustly even when it lies at the boundary of the search image.

The next section describes the process of decaying extension of an image in detail. Sect. 3 compares the proposed technique with other methods on cluttered, rotated, blurred, and noisy images. Finally, Sect. 4 concludes the paper.

2. DECAYING EXTENSION OF AN IMAGE

We propose to extend an image using three steps as follows.

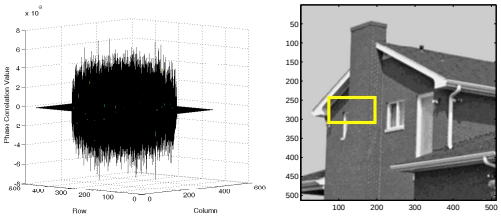

Figure 6: (*left*) Surface showing various false peaks at the central region and zeroed true peak at the boundary region in the response of the projection operator based PC between the zero-padded original *house* image and the zero-padded original template of the top portion of the house. (*right*) The top-left vertex of the overlaid yellow rectangle corresponds to the highest peak at (248, 80) having magnitude only 0.007.

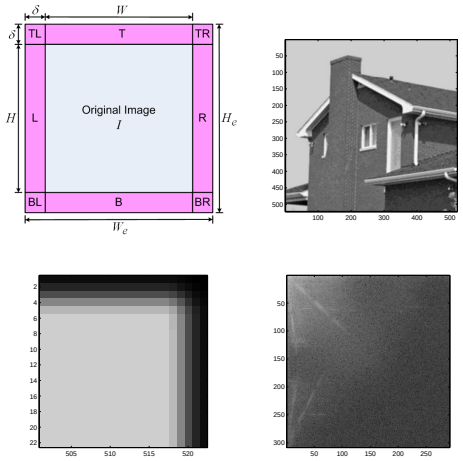

Figure 7: (*top left*) Structure of an extended image, I_e, where T = Top, L = Left, R = Right, and B = Bottom. (*top right*) House image extended using $\delta = 5$. (*bottom left*) Zoomed in top-right corner of the extended image showing new pixels with smoothly decaying values. (*bottom right*) Top-left quadrant of the spectrum of the zero-padded extended image.

2.1 Initializing the Extended Image

Let the original image be denoted by I having width, W, and height, H. We want to extend the image from every direction by δ pixels, as shown in Fig. 7(top left). Thus, the size of the extended image, I_e, becomes $W_e \times H_e$, where $W_e = W + 2\delta$ and $H_e = H + 2\delta$. If $\delta < 4$, the discontinuities are not eliminated adequately. On the other hand, if $\delta \gg 4$, the discontinuities are eliminated very smoothly, but the computation time taken by the phase correlation between the larger images is increased accordingly. In order to remain in the safe side, we set $\delta = 5$ in all of our experiments to have robust image matching without significantly increasing the computation time. We initialize every pixel in the extended regions with the value of its nearest boundary pixel in the original image. Thus, the initialized extended image is obtained, as:

$$I_e(x,y) = \begin{cases} I(0,0) \\ \quad \text{if } 0 \le x \le \delta - 1 \ \& \ 0 \le y \le \delta - 1, \\ I(x-\delta, y-\delta) \\ \quad \text{if } \delta \le x \le W + \delta - 1 \ \& \ \delta \le y \le H + \delta - 1, \\ I(0, H-1) \\ \quad \text{if } 0 \le x \le \delta - 1 \ \& \ H \le y \le H + \delta - 1, \\ I(W-1, 0) \\ \quad \text{if } W \le x \le W + \delta - 1 \ \& \ 0 \le y \le \delta - 1, \\ I(W-1, H-1) \\ \quad \text{if } W + \delta \le x \le W_e - 1 \ \& \ H + \delta \le y \le H_e - 1. \end{cases}$$

2.2 Generating the Decaying Weights

We use a Gaussian function (because of its smooth and symmetric behavior) to generate the decaying weights to be applied on the new pixels in the initialized extended image.

Consider a $\mathbb{R}^K$ column vector (where $K = 2\delta + 1$), denoted by g, containing the weights obtained by a Gaussian function, as: $g(k) = \exp\big[-(1/2)\{(k-\delta)^2/\sigma^2\}\big]$, where $k = 0, 1, \ldots, K-1$, and the standard deviation, σ, controls the spread of the Gaussian function. If the value of σ is too large, the function will go down too slow [Fig. 8(*left*)], and the boundary elements will be too high from zero, resulting in a sharp discontinuity during the zero-padding operation. On the contrary, if its value is too low, the function will go down abruptly to near-zero value [Fig. 8(*middle*)], resulting in a discontinuity even before the zero-padding operation. We want to have the function, that goes down smoothly and becomes near-zero at its boundaries. We achieve this objective by computing an appropriate value of σ automatically according to the size of the vector using the expression: $\sigma = 0.3\big[(K/2) - 1\big] + 0.8$, as in [3]. The appropriate value of σ for $\delta = 5$ (i.e. $K = 11$) and the corresponding Gaussian weights are shown in Fig. 8(*right*). Then, we split the vector g into its top and bottom halves (i.e. $\mathbb{R}^\delta$ vectors), as: $g_t(k) = g(k)$ and $g_b(k) = g(k + \delta + 1)$, where $k = 0, 1, \ldots, \delta - 1$.

Now, consider a $\mathbb{R}^{K \times K}$ matrix, G, containing the weights obtained by a 2D Gaussian function, as:

$$G(k_x, k_y) = \exp\left[-\frac{1}{2}\left\{\frac{(k_x - \delta)^2 + (k_y - \delta)^2}{\sigma^2}\right\}\right], \quad (2)$$

where k_x and $k_y = 0, 1, \ldots, \delta - 1$, and the value of σ is the same as computed previously. Then, we split the matrix G into its four quarters (i.e. $\mathbb{R}^{\delta \times \delta}$ matrices). The top-left, top-right, bottom-left, and bottom-right quarters are obtained as: $G_{tl}(k_x, k_y) = G(k_x, k_y)$, $G_{tr}(k_x, k_y) = G(k_x + \delta + 1, k_y + \delta + 1)$, $G_{bl}(k_x, k_y) = G(k_x, k_y + \delta + 1)$, $G_{br}(k_x, k_y) = G(k_x + \delta + 1, k_y)$, where k_x and $k_y = 0, 1, \ldots, \delta - 1$.

2.3 Applying the Decaying Weights

We multiply the g_t vector element-by-element with every column in the **T** patch in the initialized extended image [see Fig. 7(top left)], g_b vector with every column in the **B** patch, g_t^T row-vector with every row in the **L** patch, g_b^T row-vector with every row in the **R** patch, G_{tl} matrix with the **TL** patch, G_{tr} matrix with the **TR** patch, G_{bl} matrix with the **BL** patch, and G_{br} matrix with the **BR** patch. Figure 7(top right) illustrates the extended version of the *house* image. It may be noted that all the content of the original image is intact in the extended image, and that only the pixels at the external regions are gradually getting more and more black as

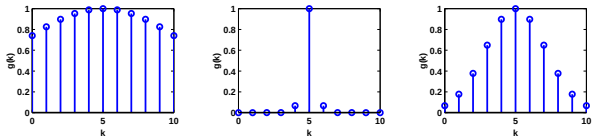

Figure 8: Effect of σ on the Gaussian weights, when $\delta = 5$ (i.e. $K = 11$). (*left*) $\sigma = 6.45$, (*middle*) $\sigma = 0.43$, and (*right*) $\sigma = 2.61$ (computed automatically).

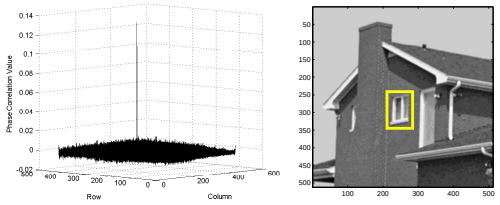

Figure 9: (*left*) Surface showing single dominant peak at the correct location in the response of the PC between the zero-padded extended *house* image and the zero-padded extended template of the window of the house. (*right*) The top-left vertex of the overlaid rectangle corresponds to the highest peak at (246, 216) having magnitude 0.134.

depicted by the zoomed-in top-right corner of the extended image in Fig. 7(bottom left). Furthermore, the resulting spectrum [Fig. 7(bottom right)] of the zero-padded extended image does not contain the high-frequency artifacts [unlike the spectrum of the zero-padded original image – see Fig. 1(bottom right)] or the distorted low frequency components [unlike the spectrum of the zero-padded Blackman modulated image – see Fig. 3(top right)].

3. EXPERIMENTAL RESULTS

When we phase correlated the zero-padded extended versions of the *house* image and the template of the window of the house, we obtained a single dominant peak in the PC response [Fig. 9(left)] corresponding to the true location of the window in the image [Fig. 9(right)]. Similarly, when we phase correlated the zero-padded extended versions of the same *house* image and the template of the top portion of the house, again we obtained a single dominant peak in the PC response [Fig. 10(left)] corresponding to the true location of the top portion of the house in the image [Fig. 10(right)]. We can see that the proposed technique has drastically attenuated all the false peaks in the PC response even when the object lies away from the central region in the search space. This is a significant improvement introduced by the proposed approach in the PC response as compared to the response obtained using the Blackman window (Fig. 4) or the projection operator (Fig. 6).

Furthermore, we compared the proposed approach with the two techniques also for localizing a moving target in the consecutive frames of numerous real videos. The proposed approach outperformed them in all the videos, especially when the object moved away from the central region of the video frame. For example, Fig. 11 shows a noisy, shaky, and blurred video of a ground vehicle recorded from a UAV (unmanned aerial vehicle). The ground vehicle is being continuously translated and slightly rotated, and there is a noticeable glare effect at the end of the video. The template – shown at the bottom-left of every frame – is kept constant to make the scenario more challenging. We can observe that only the proposed approach is successful in tracking the ground vehicle persistently in this video even when the vehicle is far from the central region of the frame. Similarly, Fig. 12 depicts the results of the three techniques on a cluttered and noisy video (OneShopOneWait2cor.avi) from CAVIAR database[2]. Since the PC is only for estimating the translation, it is not supposed to handle the significant variation in the appearance of the walking woman to be tracked in this video. Therefore, we smoothly update the template in each iteration of the techniques being compared, using α-tracker template updating method [2] (with $\alpha = 0.25$). The adaptive template is shown at the bottom-left of every frame. It can be observed again that the proposed method outperforms the other two techniques in tracking the walking woman robustly even when she goes away from the central region of the frame.

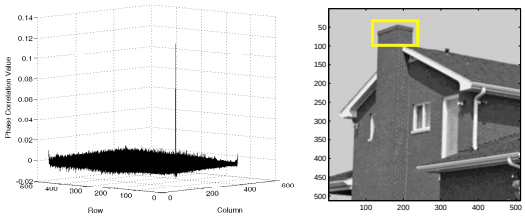

Figure 10: (*left*) Surface showing single dominant peak at the correct location in the response of the PC between the zero-padded extended *house* image and zero-padded extended template of the top portion of the house. (*right*) The top-left vertex of the overlaid rectangle corresponds to the dominant peak at (39, 123) having magnitude 0.123.

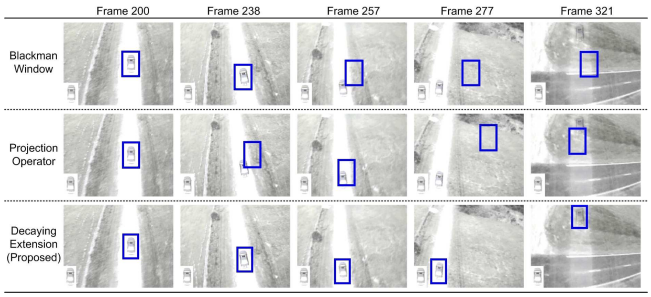

Figure 11: Localizing a ground vehicle in the frames of a noisy, shaky, and blurred UAV video containing slight rotations and significant glare effect, when the template (shown at the bottom left of every frame) is not updated. Only the proposed method localizes the vehicle in all the frames.

Inherently, the PC is not invariant to scale, rotation, etc. However, it should be a little tolerant to these situations. We performed another set of experiments to quantitatively find out the tolerances of the three PC schemes against noise, scale change, blur, and rotation of the search image. In these experiments, we used twelve standard real images each of size 512×512: *cameraman*, *house*, *jetplane*,

[2] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

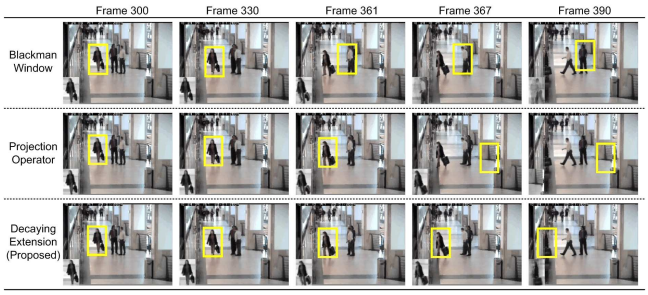

Figure 12: Localizing a woman in a noisy and cluttered CAVIAR video, when the template (shown at the bottom-left of every frame) is iteratively updated. Only the proposed method localizes the woman in all the frames.

Table 1: Average tolerances of different PC techniques

	Black.	**Proj.**	**Proposed**
Gaussian Noise (σ_n)	28.6	101.0	106.1
Scale Change (%)	$\pm$3.6	$\pm$8.6	$\pm$8.7
Blur (Avg. Filt. Size)	2×2	10×10	11×11
Rotation ($^\circ$)	$\pm$2.5	$\pm$3.2	$\pm$2.6

lake, *lena*, *livingroom*, *mandrill*, *peppers*, *pirate*, *walkbridge*, *woman_blonde*, and *woman_darkhair* (see footnote at the first page of this paper). From every image, we extracted a 75×75 template with its top-left position at $(316, 256)$, *because the other two techniques could not detect the object at all when it was far from the central region*. Once the template was prepared, we degraded every test image under consideration with: (1) zero-mean Gaussian noise by increasing its standard deviation parameter (σ_n), (2) change in scale using bi-cubic interpolation method, (3) blur effect by increasing the size of an average filter, or (4) degree of rotation using bi-cubic interpolation method. We went on increasing the intensity of degradation, until the particular PC technique failed to localize the object correctly. Then, we averaged the limiting intensity of the corresponding degrading effect applied on all the 12 images, as listed in Table 1. The more the tolerance value, the more robust the PC technique. The table shows that the proposed approach is the most tolerant to the noise, the scale change, and the blur effect, and is the second most tolerant to the rotation of the test image.

We also compared the computation time taken by each technique in MATLAB, using a 512×512 search image and varying the size of a square template from 25×25 to 250×250. In every case, the proposed technique took about as much time as the standard PC or the projection operator based PC, and was about 2.5 times faster than the Blackman window based PC, as shown in Fig. 13. For example, in case of 75×75 template, the standard PC took 0.542 s, the projection operator based PC 0.554 s, the proposed PC 0.561 s, and the Blackman window based PC 1.545 s.

4. CONCLUSION

The cyclically repeated image assumption of DFT and the zero-padding operation cause discontinuities in the spatial domain and high-frequency artifacts in the frequency domain, especially when the pixels at the opposite boundaries of the images are significantly mismatched. This phenomenon results in various false peaks in the PC response. The proposed decaying extension technique eliminates the discontinuities and the high frequency artifacts, and attenuates the false peaks, resulting in robust object localization. The major benefit of the proposed technique over the previous methods is that it can localize the object lying *anywhere* in the *whole* search space. The comparison results show that the proposed technique is also: (1) the most tolerant to the Gaussian noise, the scale change, and the blur effect, (2) the second most tolerant to the rotation, and (3) about 2.5 times faster than the Blackman window based PC.

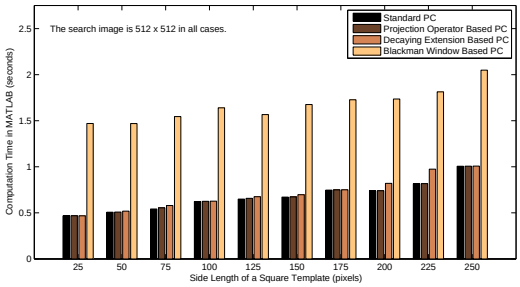

Figure 13: Computation time comparison in MATLAB.

REFERENCES

[1] J. Ahmed, M.N. Jafri, M. Shah, and M. Akbar. Real-time edge-enhanced dynamic correlation and predictive open-loop car-following control for robust tracking. *Machine Vis. and App.*, 19(1):1–25, Jan. 2008.

[2] S. Blackman and R. Popoli. *Design and Analysis of Modern Tracking Systems*. Artech House, Boston, 2nd edition, 1999.

[3] G. Bradski, A. Kaehler, and V. Pisarevsky. Learning-based computer vision with open source computer vision library. *Intel Tech. J.*, 9(3):118–131, May 2005.

[4] D. Donnelly and B. Rust. The fast fourier transform for experimentalists, part ii: Convolutions. *Computing in Science and Engineering*, 7(4):92–95, 2005.

[5] H. Foroosh, J.B. Zerubia, and M. Berthod. Extension of phase correlation to sub-pixel registration. *IEEE Trans. on Image Processing*, 11(3):188–200, Mar. 2002.

[6] Y. Keller, A. Averbuch, and O. Miller. Robust phase correlation. In *Proc. 17th IEEE International Conference on Pattern Recognition*, 2004.

[7] D. Kuglin and D.C. Hines. The phase correlation image alignment method. In *Proc. IEEE Conference on Cybernetics and Society*, pages 163–165, Sep. 1975.

[8] R. Manduchi and G.A. Mian. Accuracy analysis for correlation-based image registration algorithms. In *Proc. ISCAS-93: Int'l Symposium on Circuits and Systems*, volume 1, pages 834–837, Sep. 1993.

[9] A.V. Oppenheim, R.W. Schafer, and J.R. Buck. *Discrete-Time Signal Proc.* Prentice Hall, 2nd edition, 1999.

[10] H.S. Stone, B. Tao, and M. McGuire. Analysis of image registration noise due to rotationally dependent aliasing. *Journal of Visual Communication and Image Representation*, 14(3):114–135, 2003.