Real Time High Resolution Digital Image Motion Estimation

James Carron, Min Wan, John T. Sheridan

Department of Electronic Engineering, University College Dublin

1. An Important Problem

Inferring physical motion from digital images is a very difficult task for computers to complete. If we can quantify the movement between two images, knowing the rest of the image system (lensing etc) we can quantify the physical movement that occured between the two images. With the explosion of automation this data is required for real time applications and current methods cannot process the high spatial resolution images. Modern Digital Image sensors are capable of capturing millions of pixels several times per second with each pixel containing up to 16 bits of information. For example storing an hour of uncompressed 4k 60fps footages requires approximately 360Gb, also making it difficult and expensive to store for later processing. Hence a new approach is needed.

2. Proposed Method

The proposed method uses the fact that image sensors average the light falling over each pixel and return this value. For example with a dark object moving in a positive x direction over the boundary between pixels. The pixels to the left of the object will get gradually brighter while the pixels to the right of the object will get gradually darker. Each image has a unique relationship between translation and the respective row or column total. This can be measured in simulation to calibrate the method and is related to the composition of the image.

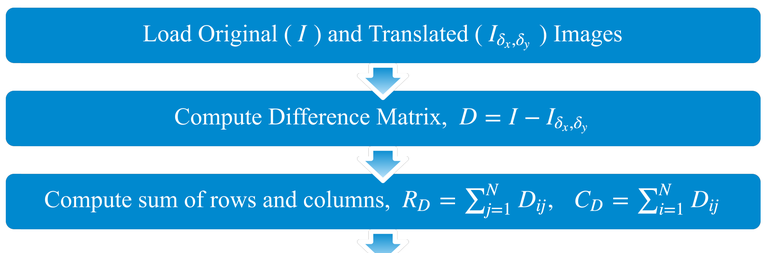
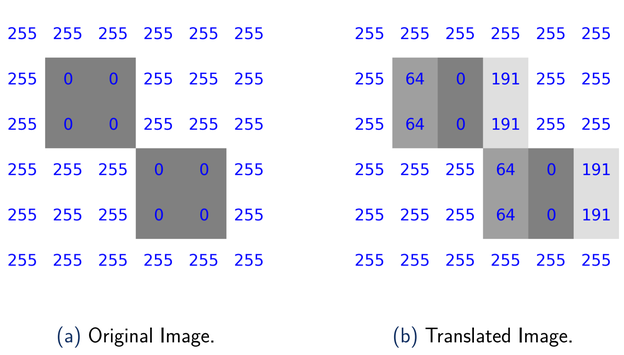
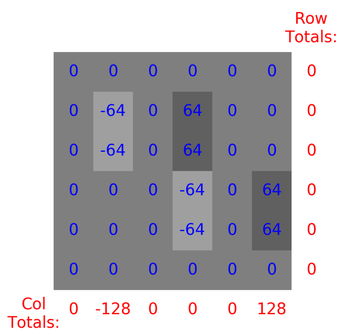


Figure: Proposed Method Flowchart.

T_{R_D} and T_{C_D} act as an indicators of motion in the x axis and y axis respectively.



(a) Original Image. (b) Translated Image.



(c) Difference Matrix.

Figure: Estimating a +0.25px in x Translation w/ pixel values.

For this image we can compute T_{R_D} and T_{C_D} to be 256 and 0 respectively. This indicates that there has been motion in the x direction and none in the y direction as we expect. This image has a t_x/T_{R_D} value of 1024 and hence $256/1024$ gives us our translation of 0.25px.

3. Current Methods

Correlation is the most widely used digital image motion estimation technique. It tests the similarity of two images using a convolution method, which involves multiplying the translated and test image pixel by pixel. It is defined mathematically as:

$$(f \star g)(\tau) \coloneqq \int_{-\infty}^{+\infty} f^*(t) g(t+\tau) dt \tag{1}$$

Time Domain Correlation operates by windowing the translated image and testing it at every possible translation on the original image. This produces a single value for how similar the test image is to the original image at each position.The point at which the images have the highest similarity quantifies the translation.

Frequency Domain Correlation utilises the shifted frequency domain property to reduce the computational expense of calculating the convolutions. Once calculated the translation deltas m_x and m_y can be separated out and converted back to the time domain.

	Original Image	Translated Image
Time Domain	$x(n_1, n_2)$	$x(n_1 - m_1, n_2) - m_2$
Frequency Domain	$x(w_1, w_2)$	$e^{-jw_1m_1}e^{-jw_2m_2}x(w_1, w_2)$

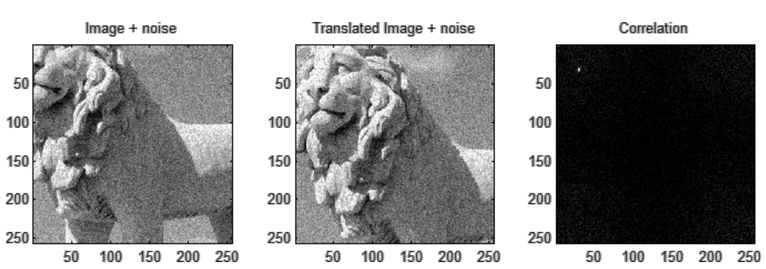
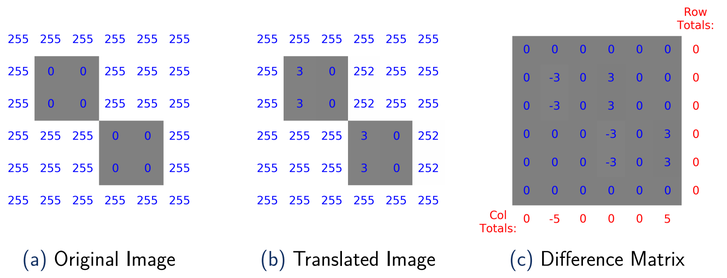


Figure: Correlation Example [1]

4. Fundamental Advantages

Spatial Resolution: As defined by Mas et al[2] the minimum detectable digital image translation is that whereby a single pixel changes value by the minimum amount. This is determined by the number of bits used to represent each pixel. For example for a 8bit image there are 256 values to choose from. The proposed method can detect this smallest change and hence has a minimum detectable translation of $1/2^b$. ie $1/256$ of a pixel for an 8 bit image.



(a) Original Image (b) Translated Image (c) Difference Matrix

Figure: A $1/100$ px +x translation is detectable with this method, $T_{C_D} = 10$

Fundamentally Correlation Methods cannot detect subpixel motion. Therefore they have to supersample the image (estimate what the values are between pixels) to generate a higher spatial resolution image which it then operates on to infer a sub pixel translation.

Computational Expense: To compare how difficult the algorithms are to compute, formulae are derived to calculate the number of operations needed to evaluate the motion translation between images. Reduced computational expense allows increased temporal resolution and real time processing on lower power devices. For this application data loading and other operations have negligible effect on the computation time.

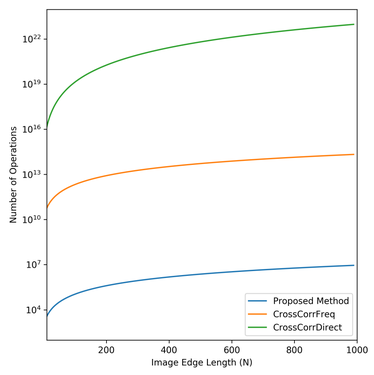


Figure: Varying size of 8bit images.

Method	Number of Operations
Time Domain Correlation	$\lceil b^{1.465} \rceil (2^b N)^4$
Frequency Domain Correlation	$\lceil b^{1.465} \rceil (4(2^b N)^2(2 \times log_2(2^b N) + 1))$
Proposed Method	$(3N^2 + 2N)(log_2(b))$

Table: Comparing each Algorithms Computational Expense.

In this case we define the images as two $N \times N$ pixel greyscale b-bit images. We can note that the computational expense versus the image size of the Direct Cross Correlation Method is approximately quartic, the Frequency Domain Cross Correlation method is approximately cubic and the Proposed method is approximately quadratic. For example using 1080p Images the proposed method is theoretically 10^6 times faster then Frequency Domain Correlation. The proposed method is also less sensitive to increasing bit depth.

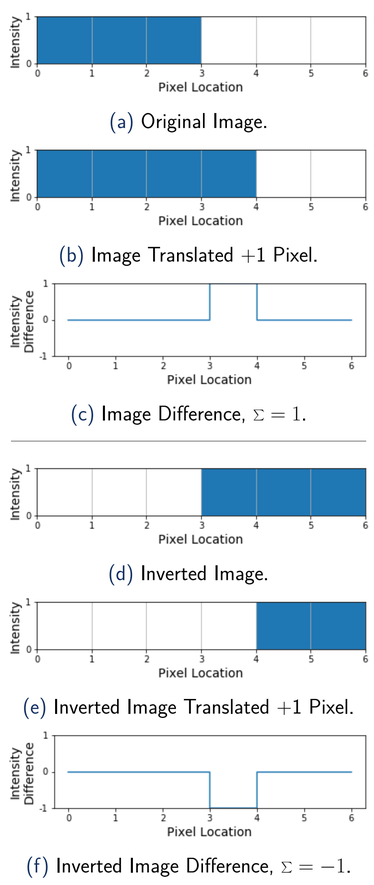
Robustness to Noise As noise is normally a random process with a mean of zero (Gaussian, Poisson) an additive or subtractive process is able average the noise out across an image. Meanwhile a multiplicative process (like convolution) will exacerbate the effect of noise. With increasing image size the proposed method is more resistant to noise meanwhile the convolution method becomes more sensitive to noise.

Error Detection As an image moves in a direction (for example positive x) the pixels behind it change and the ones in front of it change by an inversely proportional amount. Hence these two areas in the difference image cancel out. Deviation from zero of this averaged value can be used to quantify the effect of noise and edge effects on the image and quantify the amount of error in a motion measurement.

5. Fundamental Limitations

Translation Direction: Consider a 1D case where we have each pixel take either a blank (0) or dark (1) value and a simple step function is used to model an idealised edge.

$$\text{Pixel Value} = \begin{Bmatrix} 1 & \forall x \leq 3 \\ 0 & otherwise \end{Bmatrix}$$



(a) Original Image.

(b) Image Translated +1 Pixel.

(c) Image Difference, $\Sigma = 1$.

(d) Inverted Image.

(e) Inverted Image Translated +1 Pixel.

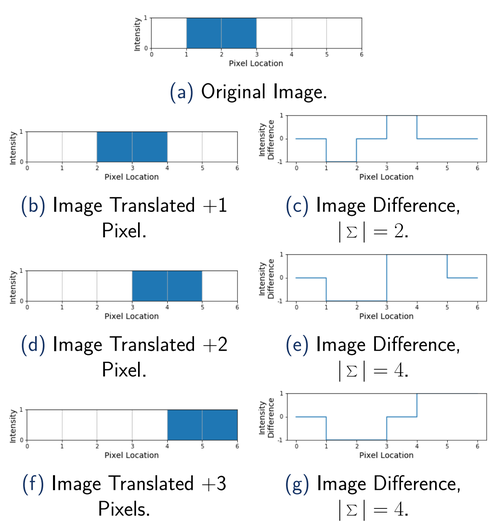
(f) Inverted Image Difference, $\Sigma = -1$.

Figure: Translation Direction Limitation.

We can show that the sign of the value output from the proposed method is not dependent on the direction of translation and the composition of the image. This is because the subtraction operation used to create the difference matrix is non commutative. This means that values (images) connected by operators will give different results if the order of the values is changed.
Consider the Original and Translated Images two matrices A and B respectively. Inverting the image pixel intensity is confused by the proposed method as inverting the translation direction can be described as follows. The original operation can be considered as $A - B$, however this is the same as $(-(B - A))$ where the translation going in the opposite direction with inverted values.

5. Fundamental Limitations Continued

Max Translation: Here it can be seen that 1 pixel translation can be interpreted as an increase of 2 of the sum of the absolute difference value. However once the translation becomes larger then the continuous block size, the sum of the absolute difference value plateaus as the continuous blocks are no longer overlapping. Hence the proposed methods max translation is heavily reliant on the composition of the image.



(a) Original Image.

(b) Image Translated +1 Pixel.

(c) Image Difference, $|\Sigma| = 2$.

(d) Image Translated +2 Pixel.

(e) Image Difference, $|\Sigma| = 4$.

(f) Image Translated +3 Pixels.

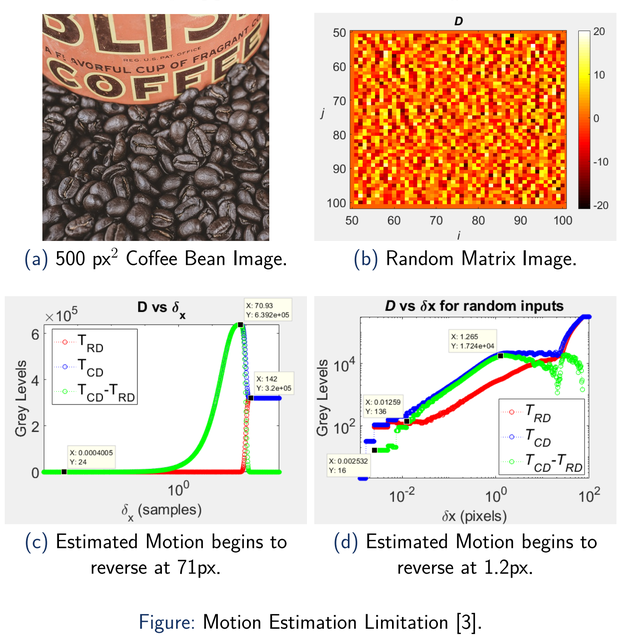
(g) Image Difference, $|\Sigma| = 4$.

Figure: Motion Estimation failing.

In Plane translation: We can only measure in plane motion (x and y translations relative to the sensor plane) with this method. For example z translations (towards and away from the sensors plane) and rotations about any axis can not be quantified directly with this method.

6. Experimental Validation

Max Translation: Testing the method on real images verifies the maximum translation theory applies for real world images.



(a) 500 px^2 Coffee Bean Image.

(b) Random Matrix Image.

(c) Estimated Motion begins to reverse at 71px.

(d) Estimated Motion begins to reverse at 1.2px.

Figure: Motion Estimation Limitation [3].

With the Coffee Bean Image the method begins to breakdown at 71 pixels of translation which corresponds neatly to the average of x length of the coffee beans which are the continuous shapes which dominate the image. For the Random Matrix Image we can see the motion begins to break down at $\approx$ 1px of translation, this again corresponds to the dominant continuous size which is the pixels themselves.

Computational Expense: Running each algorithm on the same hardware using the same images and recording the average duration for them to run allows us to verify the relationship between image size and computational expense.

Image Size	Proposed Method Duration (secs)	Correlation Method Duration (secs)	Speed up
150x150	0.1	15.4	154 x
600x600	3.2	2153.8.3	673 x
800x800	7.1	36011.3	5072 x

Table: Comparing each Algorithms Computational Expense.

7. Conclusion

The proposed method has greater spatial resolution with several magnitudes reduced computational expense compared to current methods. It has several drawbacks however as it can only quantify in plane translations, it can only determine the magnitude of a movement in any direction and has a image composition maximum detectable translation limit. It is therefore well suited to real time high speed operation where the movement will usually be subpixel and fast compute is necessary.

8. Next Steps

A method for combining the proposed method with a conventional correlation algorithm to mitigate some of the disadvantages will be investigated and tested. Combining these would allow much increased temporal and spatial resolution at much decreased computational expense. The proposed method could be used to track sub pixel movement up until a threshold translation amount is detected whereupon the proposed methods estimation can be verified with the correlation method and the original image updated.
The computation space needed for each method will also be investigated to determine if the proposed method can run in high speed memory of modern processors (high level cache) which would enable a massive speed up and to determine if it can fit in the tiny volatile memory of low power devices often used in the field.

[1] "WikiPhase Correlation.Png". en. In: *Wikipedia* ().

[2] D. Mas, B. Ferrer, J. T. Sheridan, et al. "Resolution Limits to Object Tracking with Subpixel Accuracy". EN. In: *Optics Letters* 37.23 (Dec. 2012), pp. 4877–4879. ISSN: 1539-4794. DOI: 10.1364/OL.37.004877.

[3] C. Duignan. "High Speed High Resolution Digital Image Motion Estimation". en. In: (), p. 102.