Multimodal Image Registration with Applications to Image Fusion

Jamie P. Heather
Waterfall Solutions Ltd
32 London Road, Guildford, UK
jamie.heather@waterfallsolutions.co.uk

Dr Moira I. Smith
Waterfall Solutions Ltd
32 London Road, Guildford, UK
moira.smith@waterfallsolutions.co.uk

Abstract - This paper presents an algorithm for accurately aligning two images of the same scene captured simultaneously by sensors operating in different wavebands (e.g. TV and IR). Such a setup is common in image fusion systems where the sensors are physically aligned as closely as possible and yet significant image mis-alignment remains due to differences in field of view, lens distortion and other camera characteristics. Our proposed registration method involves numerically minimising a global objective function defined in terms of local normalised correlation measures. The algorithm is demonstrated on real multimodal imagery and applications to image fusion are considered. In particular we illustrate that fused image quality is closely related to the degree of registration accuracy achieved. To maintain this accuracy in real systems it is often necessary to continuously update the transform over time. Thus we extend our registration approach to execute in real time on live imagery, providing optimal fused imagery in the presence of relative sensor motion and parallax effects.

Keywords: Image registration, image fusion.

1 Introduction

Image registration is the process of spatially aligning two or more images of the same scene, possibly recorded at different moments in time. This broad definition encompasses a multitude of image alignment problems in the fields of medicine, defence, remote sensing, computer vision and pattern recognition. In each case the fundamental problem is the same; to find a mapping

\[(x, y) \rightarrow (u, v), \quad u = u(x, y), \quad v = v(x, y)\]  

between the pixels \((x, y)\) in one image and the pixels \((u, v)\) in another. The complexity of the solution will depend on the application under consideration. In the simplest case, a straightforward geometric translation or rotation may be sufficient to accurately align the two images. More advanced global approaches include affine, polynomial and projective transformations. Where a global transformation is not appropriate, a piecewise, elastic membrane or optical flow technique may be applied instead. Brown [1] presents an overview of these various transformations and the established methods for obtaining a registration solution. She also emphasises that before trying to solve an image registration problem it is vital to first appreciate the cause of the mis-alignment and then select a transformation which appropriately models it. Sources of mis-alignment can be grouped into two categories:

- **Spatial mis-alignments** commonly arise due to differences in sensor position/viewing angle and also due to differences between the sensors themselves (e.g. field of view, resolution, lens distortion).
- **Temporal mis-alignments** occur when there is some relative motion between the sensor and the objects in the scene and the images are captured at different points in time.

Figure 1 below illustrates these sources of spatial and temporal alignment for a two-camera setup.

The possibilities above give rise to a large number of very different image registration problems. In this paper we consider the specific problem of aligning two images captured simultaneously from (approximately) the same viewpoint by rigidly mounted sensors operating in different regions of the electromagnetic spectrum (Figure 2). This setup is common in image fusion systems which aim to combine the complementary features from two or more wavebands into a single image with extended information content. The fused image offers significant benefit over working with the raw sensor outputs including increased situational awareness and improved target detection/identification accuracy. A variety of simple and advanced image fusion algorithms have been...
developed over the last two decades [2] but they all operate at the pixel level and make the fundamental assumption that the source imagery is properly spatially aligned. The accuracy of the registration process is therefore critical to overall system performance.

Figure 2. The raw outputs from a TV sensor (top) and an IR sensor (bottom). A significant difference in scale is immediately obvious and closer inspection of the IR image reveals a barrel distortion effect not present in the TV image.

Image fusion systems usually employ one of two sensor configurations (Figure 3). In the first illustration the sensors are closely mounted side-by-side and boresighted for a particular distance (often infinity). In terms of the registration problem this simple setup has an immediate disadvantage: parallax effects. The separation between the optical paths of the sensors means that it is not possible to find a fixed transformation which will always map one source image onto the other. Instead the transformation has a complex dependency on distance to objects in the scene. In practice, for closely mounted sensors boresighted at infinity, the parallax effect becomes negligible (i.e. sub-pixel) when the distance exceeds a certain value (usually a few hundred metres).

Parallax effects can be avoided (to a large extent) through the use of a common aperture. In this configuration (also illustrated) one or more beam-splitters are employed, thereby allowing the two sensors to share a common optical path. In practice some design compromises are usually necessary due to the physical constraints of sensor/beam splitter size (particularly in systems with more than two cameras) and consequently some small degree of parallax may remain. Common aperture systems can also be expensive and, as such, may not be the preferred ‘off-the-shelf’ low-cost solution.

Initially we will assume that parallax effects are negligible in our image fusion system (either because the objects in the scene are sufficiently distant or through the use of a common aperture). The additional assumption that the two sensors are synchronised allows us to seek a fixed transformation mapping the pixels in one image (referred to as the ‘reference’ image) onto those in the other image (referred to as the ‘input’ image). Any image mis-alignments are now due to differences in sensor characteristics (e.g. field of view, pixel resolution, optical distortion) and can generally be represented using an $N$th degree polynomial transformation

\[ u(x, y) = \sum_{i=0}^{N} \sum_{j=0}^{N} a_{i,j} x^i y^j, \quad v(x, y) = \sum_{i=0}^{N} \sum_{j=0}^{N} b_{i,j} x^i y^j \quad (2) \]

where the coefficients $a_{i,j}$ and $b_{i,j}$ are constants to be determined. When $N = 1$ equation (2) reduces to the popular affine transformation which is capable of representing translations, rotations and shears. Adding higher order terms to the equation allows compensation of complex lens distortions such as pin cushion or barrel distortion effects. A second or third degree polynomial is sufficient for most practical applications.

At this point we note that the projective transform, as defined by the equations

\[ u(x, y) = \frac{c_1 x + c_2 y + c_3}{c_4 x + c_5 y + 1}, \quad v(x, y) = \frac{d_1 x + d_2 y + d_3}{d_4 x + d_5 y + 1} \quad (3) \]
(where \( c_1, c_2, d_1, ..., d_2 \) are constants), is often argued to be a superior representation of image mis-alignment in image fusion systems. This is because the underlying model assumes the scene is being viewed from different viewpoints and hence is able to compensate for parallax effects. However, the model also assumes the scene is planar (i.e. completely flat) which is generally only a valid approximation in airborne applications, where parallax effects are less noticeable anyway. Moreover, even when the scene is effectively planar the transform coefficients must be re-calculated whenever the cameras are moved to a new viewing angle (Figure 4). We therefore advocate that for most practical image fusion applications the projective transform offers no benefit over the more flexible polynomial transform (although in section 5 we consider the real-time update of transform coefficients).

**Figure 4.** For ‘flat’ scenes the projective transform is able to compensate for parallax effects but the transform coefficients must be re-calculated if the viewpoint changes.

Determining optimal global transform coefficients for a multimodal registration problem is no trivial task due to the complex relationship between the wavebands. The intensity of a particular image pixel is determined not only by the camera response but also by a number of physical properties (e.g. materials in the scene, atmospheric conditions and background radiation) and hence, even after accurate spatial alignment, multimodal imagery often remains uncorrelated. Consequently, many established registration techniques (e.g. from the mature field of computer vision [3]) cannot be applied directly to multimodal imagery.

In this paper we propose an algorithm for automatically determining optimal transform coefficients for aligning multimodal imagery (under the assumptions described above). Our approach is based on previous work by Irani and Anandan [4] which split the registration problem into two stages:

- Identification of a suitable image representation based on a multi-scale analysis of spatial structure. This representation is (relatively) invariant to raw image intensity and hence is ideal for assessing alignment of multimodal imagery.
- Formulation of a new automatic alignment technique utilising normalised correlation as a local similarity measure.

This method has been demonstrated on real multimodal imagery and appears to work well. Our approach also utilises an invariant image representation and local correlation measures but they are formulated into a global objective function to be minimised. We then have a rigorously-defined optimisation problem which is solved to provide highly accurate registered images.

## 2 Approach

We present an overview of our automatic registration algorithm as an optimisation problem where an objective function is to be minimised. The objective function provides an assessment of ‘how good’ a particular choice of transform coefficients is. It is constructed by locally assessing the similarity between the reference image and the transformed image and then summing over all local regions. Standard correlation techniques can be used as similarity measures provided the multimodal imagery is first decomposed into an intensity invariant representation. However, the weighting of alignment achievement across the field of view (particularly for large fields of view) should also be considered. This region could, for example, be driven as a foveal patch in piloting applications.

### 2.1 Image correlation techniques

Image correlation techniques have been successfully applied in a wide range of applications for many years, and particularly in the field of computer vision. Perhaps the most common example is in stereo-vision, where two snapshots taken from different viewpoints are used to recover depth information. Here the sensors are usually identical and hence a straightforward mean-square-difference calculation

\[
\lambda_{AB} = \sum_{x,y} (A(x,y) - B(x,y))^2
\]

is suitable as a similarity measure which is minimised when the two images \( A \) and \( B \) are perfectly aligned. A related measure is the *normalised correlation*...
\[ \theta_{AB} = \frac{\sum_{x,y} A(x,y)B(x,y)}{\sqrt{\sum_{x,y} A(x,y)^2} \sqrt{\sum_{x,y} B(x,y)^2}} \]  

which lies in the range \([-1, 1]\) and is invariant to scalar multiplication. This function attains its maximum/minimum value when \(B = \alpha A\) for some positive/negative constant \(\alpha\) respectively, and is zero when the two images are uncorrelated. The final measure we introduce here is the statistical correlation

\[ \phi_{AB} = \frac{\sum_{x,y} (A(x,y) - \mu_A)(B(x,y) - \mu_B)}{\sqrt{\sum_{x,y} (A(x,y) - \mu_A)^2} \sqrt{\sum_{x,y} (B(x,y) - \mu_B)^2}} \]  

which also lies in the range \([-1, 1]\) and is invariant to both scalar multiplication and addition. It is therefore the most robust of the three measures.

### 2.2 Invariant image representation

The correlation techniques above can be made more robust by applying them to local image regions rather than to whole images. We can construct a global similarity measure by summing the correlation over all local regions and the registration task is then defined as an optimisation problem. However, this global similarity measure is still not sufficiently robust to be applied directly to multimodal imagery, where the wavebands are often uncorrelated (both globally and locally) and may exhibit disjoint features. Correlation techniques can be used though if we choose an image representation that emphasises common spatial structures (lines, corners, contours, etc) and suppresses low spatial frequency features which tend to be more modality-dependent.

Laplacian filtering is an obvious approach for extracting high frequency spatial detail. In practice though, this filter tends to extract ‘too much’ information for registration due to its rotational invariance. Irani and Anandan suggest the more selective approach of applying directional derivative filters to the raw imagery and then optimising the transformation coefficients for vertical, horizontal and diagonal alignment. To improve performance the filtered images are also squared to make the representation more invariant to contrast reversals. These energy images tend to be well-correlated (at least locally) when the original images are well-aligned and hence are a good invariant image representation. In terms of implementation we propose a variant of the method described above; instead of transforming the energy images and then testing alignment it is better to transform the raw images and then calculate their energy and alignment. This ensures we are comparing like-with-like.

Figure 5 demonstrates the benefit of this approach when testing for large image distortions.

![Figure 5](image-url)

Irani and Anandan also propose using image pyramids to aid the registration process. Pyramids are a familiar tool in image fusion for performing multi-scale image analysis [2] and are easily created by successively filtering and down-sampling the source images. A common example is the Gaussian pyramid which takes the name of its filter (Error! Reference source not found.).

Incorporating a multi-resolution image representation into the registration algorithm greatly simplifies the search for the optimal transform coefficients; crude estimates are obtained by aligning the small images at the base of the pyramid and these are then steadily refined by repeating the alignment process on the higher levels. This approach quickly locates the global minimum of the objective function while avoiding local minima and is thus able to overcome fairly large image distortions.

In practice, the need for coarse-to-fine registration can be greatly reduced by providing a good initial guess for the transform coefficients, e.g. by manually specifying a few tie points and then applying least squares [1]. This level of human guidance is usually acceptable for image fusion applications where the registration process only needs to be performed once for a particular sensor configuration.

![Figure 6](image-url)

**Figure 6.** A Gaussian image pyramid with 4 levels of resolution.
However, image pyramids are still useful because often the information they contain in their lower levels is just as relevant for assessing alignment as the full size imagery. This is particularly true when aligning noisy imagery or warping a low resolution sensor onto a higher resolution sensor (Figure 7). Thus we propose that instead of performing coarse-to-fine registration a better strategy is to align all the levels in the image pyramid simultaneously.

![Image of TV sensor and IR sensor images](image)

**Figure 7.** Coarse-to-fine processing can sometimes lead the registration process away from an optimal solution. In the example above the IR image contains a fixed noise pattern which manifests itself in the first energy image (corresponding to high spatial frequencies) but is suppressed in the lower levels of the pyramid. Consequently a better end result is achieved by considering all resolution levels simultaneously when aligning the images.

### 3 Algorithm

We formulate our objective function as a sum of squares and then, given an initial guess for the transform coefficients, rapidly search for the local minimum. Provided the initial guess is sufficiently accurate the algorithm will terminate with the optimal set of transform coefficients. An advantage of working with global parametric transforms (as opposed to more advanced models of image distortion, such as optical flow [5][6]) is the relatively low dimensionality of the search space. Typically we must solve for a dozen or so unknown coefficients (Table 1) and when combined with our fast optimisation technique this makes real-time alignment of multimodal imagery a feasible proposition.

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Number of coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affine</td>
<td>6</td>
</tr>
<tr>
<td>Projective</td>
<td>12</td>
</tr>
<tr>
<td>Quadratic</td>
<td>12</td>
</tr>
<tr>
<td>Cubic</td>
<td>20</td>
</tr>
</tbody>
</table>

**Table 1.** Some common parametric transforms and the number of coefficients that the registration algorithm must determine.

The following notation is now introduced: \( R(x,y) \) and \( U(u,v) \) denote the reference image and the unregistered image respectively. We will map the pixels \((x, y)\) in \( R \) onto the pixels \((u, v)\) in \( U \) using a global transformation parametrised by the vector \( p \). Our transformed image \( T \) is then a function of both position \((x, y)\) and the chosen parameters \( p \).

\[
T(x, y; p) = U[u(x, y; p), v(x, y; p)] .
\]  

(7)

When expressed in this way it is clear that in practical terms the backward mapping from the reference image onto the unregistered image is more useful for constructing the transformed image than the forward mapping, even if this is a little counter-intuitive.

Our aim is to find the parameter vector \( p \) that gives the best possible spatial alignment between the images \( T \) and \( R \). In section 2.2 it was proposed that alignment should be tested not just on the full size images but also on their multi-scale decompositions. This can be achieved by computing the Gaussian pyramids of \( T \) and \( R \) and then summing the similarity measures across the different scales. Unfortunately, though this leads to an objective function with a complex dependency on \( p \) which makes optimisation difficult. A much simpler strategy is to calculate the Gaussian pyramids of the original images \( U \) and \( R \) and then appropriately transform the unregistered images at each level. Thus our method begins with the construction of the pyramid \( \{U_1, U_2, ..., U_L\} \) defined iteratively by

\[
U_l = \begin{cases} 
U & \text{for } l = 1 \\
\text{REDUCE}(U_{l-1}) & \text{for } l = 2, 3, ..., L 
\end{cases}
\]

(8)

where REDUCE denotes the filter/decimate operation [2]. Each image in this pyramid is then warped by applying a rescaled version of the original transformation

\[
T_l(x, y; p) = U_l[u(x, y; g_{l}(p)), v(x, y; g_{l}(p))]
\]

(9)

where the function \( g_{l} \) describes the effect of the image down-sampling that takes place between the 1st and the \( l \)th levels of the pyramid on the original transformation parameters. In most cases \( g_{l} \) is easily determined and takes a fairly simple form.
Following the construction of the transformed pyramid \( \{ T_1, T_2, \ldots, T_L \} \) the next step is to convert to an ‘invariant’ image representation. For this we introduce the directional derivative operator \( \nabla_k \) defined by

\[
\nabla_k = \alpha_k \frac{\partial}{\partial x} + \beta_k \frac{\partial}{\partial y}
\]

(10)

where \( \alpha_k \) and \( \beta_k \) are constants satisfying \( \alpha_k^2 + \beta_k^2 = 1 \). We use the four directional derivatives defined in Table 2 to assess the horizontal, vertical and diagonal alignment of each transformed image against the corresponding reference image. In practice a numerical approximation to the \( \nabla_k \) operator must be employed due to the discrete nature of the images. This is implemented as a convolution operation with a small discrete kernel (e.g. a Sobel filter). To make the representation invariant to contrast reversals we will also square the filtered images (as discussed in section 2.2).

<table>
<thead>
<tr>
<th>( k )</th>
<th>Description</th>
<th>( \alpha_k )</th>
<th>( \beta_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Horizontal derivative</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Vertical derivative</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>First diagonal derivative</td>
<td>1/√2</td>
<td>1/√2</td>
</tr>
<tr>
<td>4</td>
<td>Second diagonal derivative</td>
<td>1/√2</td>
<td>-1/√2</td>
</tr>
</tbody>
</table>

Table 2. Four directional derivatives and their corresponding values of \( \alpha_k \) and \( \beta_k \) in definition (10).

At each level of resolution we assess local similarity between the corresponding squared derivative images to determine the degree of alignment. Clearly there are many possible ways of dividing the images into local regions; partitioning into disjoint blocks or sliding neighbourhood processing are two common approaches. Whatever method is chosen let us denote the family of local regions for the \( \ell \)th resolution level as

\[
\Omega_\ell = \{ \omega_{m1}, \omega_{m2}, \ldots, \omega_{mM} \}
\]

(11)

where \( \omega_{m} \) is the set of all pixels \((x, y)\) that make up the \( m \)th local region and \( M \) is the number of local regions at this resolution. Defining an image similarity measure then allows the formulation of our global objective function which is minimized over all possible choices of \( p \). Solving using our fast optimization technique tends to bring rapid convergence and usually only a few iterations are necessary to achieve very good image alignment.

### 4 Results

The algorithm outlined above has been successfully applied with a variety of multimodal imagery with excellent results. In particular, the algorithm has been used to align the TV and IR sensors in a real-time system developed by Waterfall Solutions and QinetiQ which utilises a proprietary high-performance multi-resolution fusion technique. A cubic transform was used to transform the IR image onto the TV image (both measuring 768×576 pixels), compensating for the significant difference in scale and the barrel distortion effect between the two. Iterations were initialised by specifying a few tie points by hand and calculating a crude affine transformation. The results are shown in Figure 8. Accurate image alignment can be observed by forming the composite of the transformed IR and the reference TV image (Figure 9).

![Figure 8](image1.png)

**Figure 8.** Registration of TV and IR images captured by an image fusion system. A grid has been overlayed on the IR image to help visualise the transform.

In terms of image fusion the excellent performance of the registration algorithm can be seen in the high quality image generated by our proprietary fusion algorithm (Figure 10). The relationship between registration accuracy and fused image quality is discussed further in the following section.
Real-Time Fusion Systems

In many image fusion systems the sensors are synchronised and rigidly mounted in place and hence a fixed transform is usually considered sufficient to align the source imagery. After boresighting the sensors as closely as possible in the laboratory an appropriate set of imagery must be captured to enable accurate registration to be performed. This should consist of two (or more) good quality images containing a large number of common features, distributed throughout the image, that can aid the registration process. The algorithm described in the preceding sections is ideal for determining an accurate set of transform coefficients given such imagery. These coefficients can then be hard-coded into the fusion system.

In practice there are a number of factors that can cause the source images in a fusion system to become misaligned. If the system is mounted on a moving platform (e.g. an aircraft) then operational phenomena such as accelerations, vibrations and even temperature changes can cause the sensors to gradually ‘drift’ from their original alignment over time. On a much shorter time scale sensor vibration can cause individual image frames to become mis-aligned. In both cases the extent of the mis-alignment is often fairly small but nevertheless can lead to a noticeable reduction in fused image quality (Figure 11). A more significant issue though in many fusion systems is parallax. It is important to appreciate that the initial registration coefficients will only remain valid if the distance between the sensors and the objects in the scene does not change from the initial setup. This requirement is clearly unrealistic and often violated when a system is put into service, leading to large misalignments and poor quality fused imagery.

In view of the limitations described above it is highly desirable in many real-time fusion systems to continuously update the transform coefficients over time rather than use a fixed transformation. This helps ensure that fused image quality is maintained at all times (although the extent to which the system can compensate for parallax will depend on the underlying transform: the polynomial transform is a relatively poor model of this effect; the projective transform is a better choice when the scene is planar – see section 1). As it stands the algorithm described in section 3 is too complex to execute as part of a real-time system (at least with currently affordable computing technology) but such an implementation does become feasible after a straight-forward modification of the algorithm; we propose to perform one iteration of the optimization method for each time-step in the real-time system.

The modified method begins as before: after acquiring the latest source images they are decomposed into Gaussian pyramids and a new set of candidate transform coefficients is calculated and accepted/rejected according to the observed convergence as before. At this point though instead of looping around for another attempt at refining the coefficients we use the current best guess to register the images and then the fusion algorithm is applied. The process is repeated at the following time-step to further refine the transform coefficients for the next set of acquired images. The rapid convergence of the optimisation method ensures that given an accurate initial guess at the transform coefficients our latest estimate is never far from the optimal solution. The extended registration algorithm has been applied to short image sequences with promising results. In terms of implementation, two points are observed:

- It is possible for one or more of the source sensors to degrade over time (e.g. due to noise or saturation) or for the different wavebands to exhibit a lack of common features. Thus for a robust implementation some restraint must be placed on the registration process to prevent it from going too far ‘off-course’. The magnitude of the objective function can be used as the basis for this controlling logic.
- The registration algorithm described above naturally combines with a pyramidal fusion scheme; the image pyramids generated at each timestep may be re-used at the fusion stage, reducing the number of calculations that must be performed there.
6 Discussion

This paper has reported on some findings from recent research conducted by Waterfall Solutions into improved methods for registering multi-modal imagery. A number of issues associated with registering images of different modality have been highlighted, and particular emphasis has been placed on exploitation within image fusion systems for visible and infrared imagery. A new approach for achieving off-line, highly accurate image registration has been discussed and pictorial results provided. Consideration has also been given to the suitability of the algorithm to real-time applications and the very encouraging progress made in this area has been reported. Further research into a number of important aspects of this work continues, including: registration requirements definition; faster methods for improved real-time performance; metrics for establishing registration accuracy achieved; and human factors impact of image registration on fused imagery.

Acknowledgements

The authors would like to thank QinetiQ for the use of their source imagery in this paper.

References


