# Applied Multi-Dimensional Fusion

*(A. Mahmood, P.M. Tudor)*[♠], *W. Oxford*[♦], *R. Hansford*[♣]
*(J.D B. Nelson, N.G. Kingsbury)*[✳], *(A. Katartzis, M. Petrou, N. Mitianoudis, T. Stathaki)*[✭], *(A. Achim, D. Bull, N. Canagarajah S. Nikolov, A. Łoza, N. Cvejic)*[✧]

## Abstract

The purpose of the AMDF Cluster is to investigate the benefits that data fusion, and related techniques, may bring to future military Intelligence Surveillance Target Acquisition and Reconnaissance (ISTAR) systems. In the course of this work it is intended to showcase the practical application of some of the best multi-dimensional fusion research in the United Kingdom. This paper highlights work done in the area of: Multi-spectral synthetic data generation, super-resolution techniques to enhance low resolution images, joint fusion and blind image restoration, multi-resolution target detection and identification and a task based look at assessment metrics for fusion. The paper also delves into the future aspirations of the work to look further at the use of hyper-spectral data and hyper-spectral fusion. The paper presents a wide work base in multi-dimensional fusion that is being brought together through the use of common synthetic data, posing real life problems faced in the theatre. Work done to date has produced practical pertinent research products with direct applicability to the problems posed.

## Introduction

Within the context of the Data and Information Fusion Defence Technology Centre (DIF-DTC) multi-dimensional fusion refers to the fusion of data and information that span several bands, in more than one dimension and in more than one mode of sensing.

The Applied Multi-Dimensional Fusion (AMDF) project has been developed to showcase key research in: Single and multi-modal data fusion, image enhancement, feature detection, tracking, and fusion metrics. The aim is to hone this research into practical applicable products through the technical expertise of commercial partners and the scientific excellence of academic partners.

In order to demonstrate the applicability of academic multi-dimensional fusion research to a military customer the project has constructed research activities around an urban surveillance, target acquisition and tracking scenario. For this purpose, the commercial partners have produced a simulated scenario that represents and highlights common issues within this challenging environment.

The scenario focuses on the detection of a known target moving through complex terrain (an urban environment) using 'video' imagery in both visible and thermal bands. It is representative of the support of an intelligence led operation where multiple air and ground based surveillance assets may be used to detect and confirm a known target within an Area of Interest derived from existing intelligence.

The scenario uses hypothetical linked assets of a type that might be used to provide the capability described in the near future. Hidden within the detail of the scenarios are many 'real-world' issues; truncated meta data, cumulative errors in sensor location and attitude determination, and changing environmental conditions.

This paper presents highlights of the work done in the area of dual-band video fusion: Effects of resolution, restoration and reconstruction of blurred data, and multi-sensor fusion on target detection, identification and tracking. The paper presents issues in a sequential manner though the research work is done in parallel streams. The paper starts with issues of multi-sensor scenario generation presenting a précis of scientific research conducted in the area of: Super resolution, fusion, tracking, and then concludes with a discussion on metrics and the utility of metrics to video fusion. It also presents the scope of future work in the area of utilisation of hyperspectral data.

## Dual Band Video Scenario Generation

GD-UK and QinetiQ produced synthetic visible and thermal video data at High Definition as shown in Figure 1 (in accordance with the NATO standard STANAG 4609 "Digital Motion Imagery"). These data are also down-sampled to conventional Standard Definition resolution to match current generation equipments.

[♠] General Dynamics United Kingdom Limited, Hastings – corresponding author asher.mahmood@generaldynamics.uk.com
[♦] Waterfall Solutions Limited, Guildford
[♣] QinetiQ, Farnborough
[✳] Cambridge University, Cambridge
[✭] Imperial College, London
[✧] University of Bristol, Bristol

The visible simulation for the scenario is developed by GD using the NewTek Lightwave 3D computer animation package; QinetiQ provides the corresponding long-wave IR imagery using its Cameosim hyper-spectral simulation system. This necessitates importing scene geometry and motion data, provided by GD, and assigning appropriate materials, and hence spectral and thermal properties, to the objects in the scene.

Lightwave's scene and model files are the foundation for rendering all the different modalities, visible, long wave IR, and hyperspectral band. Lightwave capability is more akin to developing visible band data, and Cameosim is a very effective multispectral tool. Although Cameosim is an effective tool it lacks facility to import data directly from Lightwave into its own proprietary format necessitating conversion. The conversion is performed using a two step process; the data files were first converted to the OpenFlight[1] (*.flt) format that is further converted into the Cameosim format. The second task involves significantly more work than the first. Unlike 'visible' ray tracers, like the one used by Lightwave, which use RGB image textures to determine the colours of objects; Cameosim requires the use of spectral signature data, where the amount of light reflected back at each frequency is defined. These signatures allow Cameosim to render objects in a more realistic fashion and at wavelengths beyond the visible band. The down side is that collecting these data is considerably harder than creating RGB (red, green, blue) textures.

To render thermal imagery Cameosim needs the 'thermal properties' of each object to be defined, as well as the spectral signatures. This consists of defining one or more layers of different materials by defining appropriate data (density, thermal conductivity etc) for each substance. As an example, a typical cavity wall would consist of a layer of 'bricks' followed by a layer of 'insulation' then a layer of 'breeze blocks'. Together with the spectral signature, these two sets of data form 'a material', which can then be assigned to objects directly, or combined together to form textures in much the same way as the RGB textures used by 'visible' ray tracers.

One of the most technically challenging problems in generating the scene was rendering the smoke from the fire. A wide range of methods were considered, ranging from a full particle simulation to a simple post-process effect. The aspiration was to make smoke match the smoke in the visible image realistically, whilst keeping the setup and rendering times to a minimum. The method chosen was to create the smoke cloud using a set of small billboard-style discs, each having a semi-transparent texture. These discs were arranged in rough layers which were moved around to emulate the smoke drifting across the scene.

The process of converting Lightwave data to the OpenFlight format and then further converting it to the Cameosim format introduced errors which were not detected until the imagery had been rendered and both versions (visible and IR) were compared. These errors were compounded (and obscured) by differences in the routes taken by the moving objects in the two rendered scenes. After an investigation it was discovered that the two packages use different algorithms for interpolating ('tweening') the motion information. This was resolved by extracting frame-by-frame position information for every moving object in the scene from the Lightwave and then these data files were imported into Cameosim.

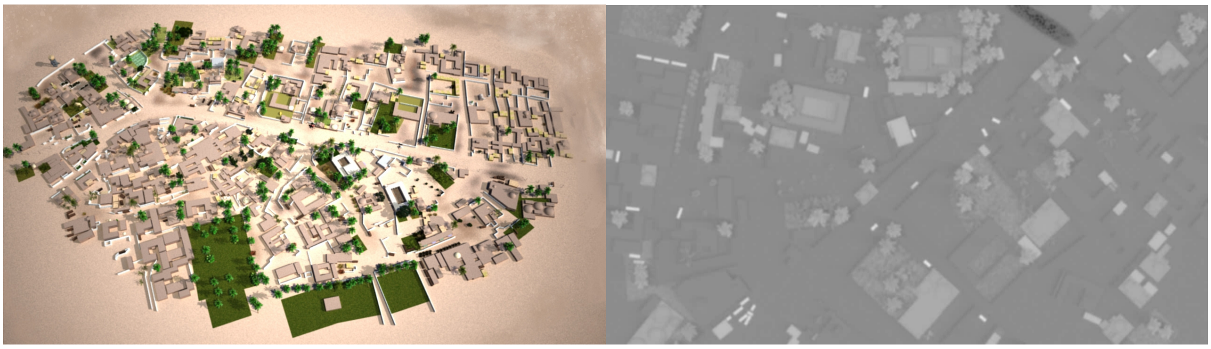


*Figure 1: Visual and LWIR Data*

## Super resolution

Super-resolution (SR) image reconstruction is a multiframe fusion process capable of reconstructing a high resolution (HR) image from several low resolution (LR) images of the same scene. It extends classical single frame image restoration methods by simultaneously utilizing information from multiple observed images to achieve restoration at resolutions higher than that of the original data.

Imperial College is presenting a new approach that circumvents to some degree some of the limitations previously associated with these techniques and makes it possible for it to be used in realistic scenarios with more complex geometric distortions (e.g. affine distortions). The SR reconstruction is formulated as a Bayesian optimization problem using a discontinuity adaptive robust kernel that characterizes the image's prior distribution. In addition, the initialization of the optimization is performed using an adapted Normalized Convolution (NC) technique [20] that incorporates the uncertainty due to mis-registration.

Imperial College has shown both qualitative and quantitative results on real video sequences and demonstrated the advantages of the proposed method compared with conventional methodologies. The general strategy that characterizes a multiframe SR process comprises three major processing steps:

a) *LR image acquisition*: acquisition of a sequence of LR images from the same scene with arbitrary geometric distortion between the images;
b) *Image registration / motion compensation*: estimation of the registration of the LR frames with each other with sub-pixel accuracy;
c) *HR image construction*: construction of an HR image from the co-registered LR images.

To start with, a brief look at the general formulation of the SR problem will be of use. First, an observation model relating the LR frames to the HR image should be formulated. The observed LR frames are assumed to have been produced by a degradation process that involves geometric warping $T(r_k)$ that depends on a parameter vector $r_k$, blurring $B$, and uniform downsampling process $D$ performed on the sought HR image $z$ (see Fig. 2). Moreover, each LR frame is typically corrupted by an additive Gaussian noise field $n_k$ which is uncorrelated between the different LR frames. Thus, the $k^{\text{th}}$ LR frame may be written as:

$$y_k = DBT(r_k)z_k + n_k = W(r_k)z_k + n_k \ \forall k = 1,2,3,...,K$$

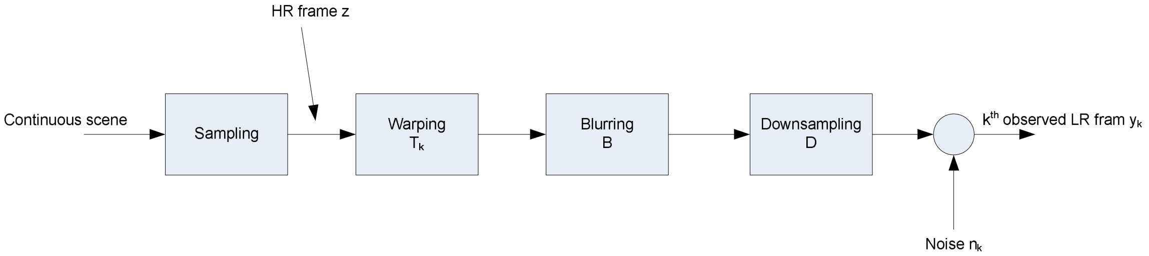


*Figure 2: Block diagram of the degradation process relating each HR frame with its LR counterpart*

A joint estimation of both the unknown HR image **z** and registration parameters $r \equiv \left[r_1^T, r_2^T, ..., r_K^T\right]^T$ is performed using the Bayesian framework. In particular, the estimates for both **z** and **r** are given by:

$$(\hat{z}, \hat{r}) = \arg_{z,r} \max P(z, r \mid y) = \arg_{z,r} \max P(y \mid z, r)P(z)$$

This is equivalent to the minimization of a posterior energy function $U(z, r \mid y)$

$$(\hat{z}, \hat{r}) = \arg_{z,r} \min U(z, r \mid y) = \arg_{z,r} \min\{-\log P(y \mid z, r) - \log P(z)\}$$

Considering that the elements of the noise field are independent and identically distributed (iid) Gaussian samples with variance $\sigma_n^2$, the data likelihood term can be expressed as:

$$P(y \mid z, r) \propto \exp\left\{-\frac{1}{\sigma_n}\sum_k \left\|y_k - W(r_k)z\right\|^2\right\}$$

On the other hand, the prior distribution $P$(**z**) incorporates discontinuity adaptive smoothness constraints in the final solution and is characterized by a Gibbs distribution of the following form:

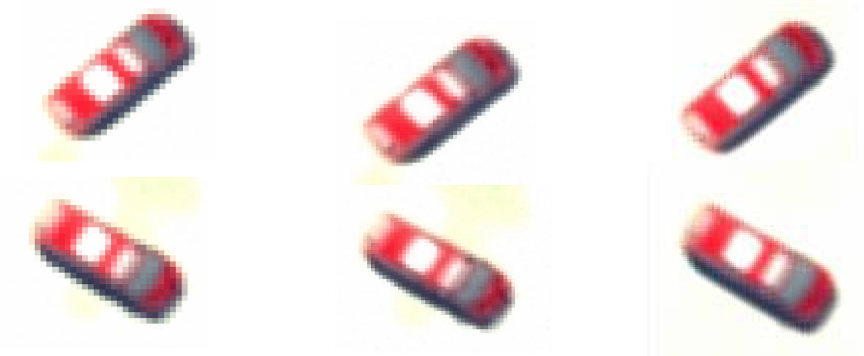
$$P(z) \propto \exp\left\{-\sum_s \rho\left(\sum_t Q_{st}(z_t)\right)\right\}$$

Where, $Q$ represents the Laplacian operator and $\rho(x)$ is a Lorentzian robust error norm, which considers predominant discontinuities in the signal as outliers (large accidental values).

It is evident that the resulted energy function $U(z, r \mid y)$ is non-convex, with several local minima. Its minimization is performed using a deterministic *continuation* method, called *Graduated Non-Convexity* approach of [23], via the construction of successive convex approximations of $U(z, r \mid y)$ and their minimization with a gradient descent approach.

This deterministic optimization method requires a good first approximation $z^0$ of the reference HR frame. This fact is generally neglected by the existing SR methodologies, which mainly resort to simple interpolation techniques using only the LR reference frame. A fast and efficient way of obtaining $z^0$ is the method of Normalized Convolution (NC), by initially registering all samples from the available LR frames in a HR reference grid. NC is a technique for local signal reconstruction, using a *certainty* map that describes our confidence in the data that constitute the unknown signal and an *applicability function* that localizes the polynomial fit. Common practise suggests that missing data in the irregularly sampled image have a certainty equal to zero, while the observed samples have a certainty equal to one. An alternative approach is proposed, which accounts for errors related to sub-optimal registration. In particular, a non-binary set of certainties is used, where samples of the reference frame get a certainty value of one, whereas samples from neighbouring frames get a positive value equal to $\varepsilon < 1$, which reflects the accuracy of the registration method. On the other hand, the applicability function corresponds to an isotropic Gaussian kernel, the size of which equals the support of the point spread function of the sensor.

The image produced as a result of super-resolution processing has a higher and improved resolution in comparison with any of the originally captured images. This is demonstrated in Figure 3, where an original LR (standard definition) frame of the video sequence and the reconstructed HR (super resolved) frame using the 16 preceding frames from the video sequence is shown, in two different posses, in comparison to High Definition Frame from a simulated HD data. In combination with tracking, the method can be used to super-resolve parts of the captured frame that contains the object of interest that is being tracked.



Standard Definition Frames

High Definition Frames
Upsampling factor ~ 2

Super Resolved SD Frames
Upsampling factor ~ 4

*Figure 3: Super resolved frames from Standard Definition Frames compared to High Definition frames*

## Joint Fusion and Blind Image Restoration

Image fusion is the process of combining information from different image realizations that capture the same registered scene in order to enhance the perception of that scene.

Current image fusion approaches detect salient features from the input images and fuse these details to form a new synthetic (fused) image. These image fusion approaches can be classified into two domains:

- Spatial domain
- Transform domain

For spatial domain techniques, the input images are fused in the spatial domain using localised spatial features. The motivation to move to a transform domain comes from the need to work in a framework, where image's salient features are more clearly depicted than in the spatial domain. Transform domain techniques project the 'input images' onto bases, modelling sharp and abrupt transitions (edges) and therefore represent the image into a more meaningful representation that can be used to detect and emphasize salient features, this is important for performing the task of the image fusion.

Image fusion is the process of combining information from different input sensor images in order to form a new composite, synthetic image that contains all the useful information of the input images. In some cases there might be parts of the observed scene where there is only degraded information available. The task in this piece of research is to identify the areas of degraded information in the input sensor images. A very simple identification approach based on local image statistics to trace the degraded areas is adopted.

Image fusion can exhibit poor performance in various situations especially when a specific region is distorted in all of the available image realizations. These distortions can be considered to be of any unknown blurring process; out-of-focus camera, motion and others etc. The current fusion algorithms will fuse all high quality information from the input sensors and for the common degraded areas will form a blurry mixture of the input images, as there is no high quality information available.

Imperial College have developed a Joint Image Fusion and Restoration method of overlapping areas to overcome this problem, while allowing for a simultaneous reduction of additive random noise (smoothing). The question does arise: Why not restore the entire images prior to fusion?

The answer is: Restoration methods enhance edge information, but suffer from various types of distortion such as ringing effects, ghost artefacts etc. This promotes the case for region based restoration, though, how can you estimate areas that are jointly distorted? This is specific to the application, and should be dealt with on case-by-case basis. To aide the estimation of overlapping areas in the multi-focus case, Imperial follow a very simple identification approach, based on local image statistics to trace the degraded areas.

The following algorithm for extracting these areas is used:

- Extract the edge map of the fused image $f$ , in terms of Laplacian kernel, i.e., $\nabla^2 f(r,t)$
- Find the local standard deviations $V_L(r,t)$ for each pixel of the Laplacian edge map $\nabla^2 f(r,t)$ , using $5\times 5$ local neighbourhoods
- Reduce the dynamic range by calculating $\ln(V_L(\mathrm{r},t))$
- Estimate $V_{sL}(\mathrm{r},t)$ by smoothing $\ln(V_L(\mathrm{r},t))$ using a $15\times 15$ median filter
- Create the common degraded area map by thresholding $V_{sL}(r,t)$ by $\mathrm{mean}_r(V_{sL}(\mathrm{r},t))-\xi$

Now an image restoration technique is applied that is based on double weighted regularised image restoration [21] with additional robust functionals to improve the performance in the case of outliers. Blind regularised image restoration uses alternating minimisation technique based on the following function:

$$Q(h(\mathrm{r}),f(\mathrm{r}))=\underbrace{\frac{1}{2}A_1(\mathrm{r})(y(\mathrm{r})-h(\mathrm{r})*f(\mathrm{r}))^2}_{\text{residual}}+\underbrace{\frac{\lambda}{2}A_2(\mathrm{r})(C*f(\mathrm{r}))^2}_{\text{image regularisation}}+\underbrace{\frac{\gamma}{2}A_3(\mathrm{r})(A*h(\mathrm{r}))^2}_{\text{blur regularisation}}$$

Where $*$ denotes 2D convolution, $h(r)$ the degradation kernel and $f(r)$ the estimated image.

This cost function has three distinct terms: Residual term represents the accuracy of the restoration process. The second term – image regularisation imposes a smoothness constraint on the recovered image and the third term acts similarly to the estimated blur. Operators $C$ and $A$ are high-pass Laplacian and PSF respectively. The constants $A_1, A_2$ and $A_3$ represent spatial weights for each optimisation term. Here, $A_1, A_2$ and $A_3$ are essentially masks that aim at letting the corresponding term in the cost function be applied at specific areas only and not the whole image. In our case, we treat $A_1$, $A_3$, as constant to 1 and $A_2$ is a penalising mask (based on an edge map) that lets the image regularising parameter smooth only constant background areas and not edges [21].

The parameters $\lambda$ and $\gamma$ control the trade-off between the residual term and the corresponding regularising terms for the image and the blurring kernel. Since each term of the cost function is quadratic, it can simply be optimised by applying Gradient Decent optimisation [22]. To recover the image using this cost function using the gradients of the cost function in terms of $f(r)$ and $h(r)$ , the iterative scheme as follows:

- At each iteration, update:

$$f^{t+1}=f^t-\eta_1\frac{\partial Q(h^t,f^t)}{\partial f^t}$$

$$h^{t+1}=h^t-\eta_2\frac{\partial Q(h^t,f^{t+1})}{\partial h^t}$$

- Stop, if $\hat{f}$ and $h$ converge.

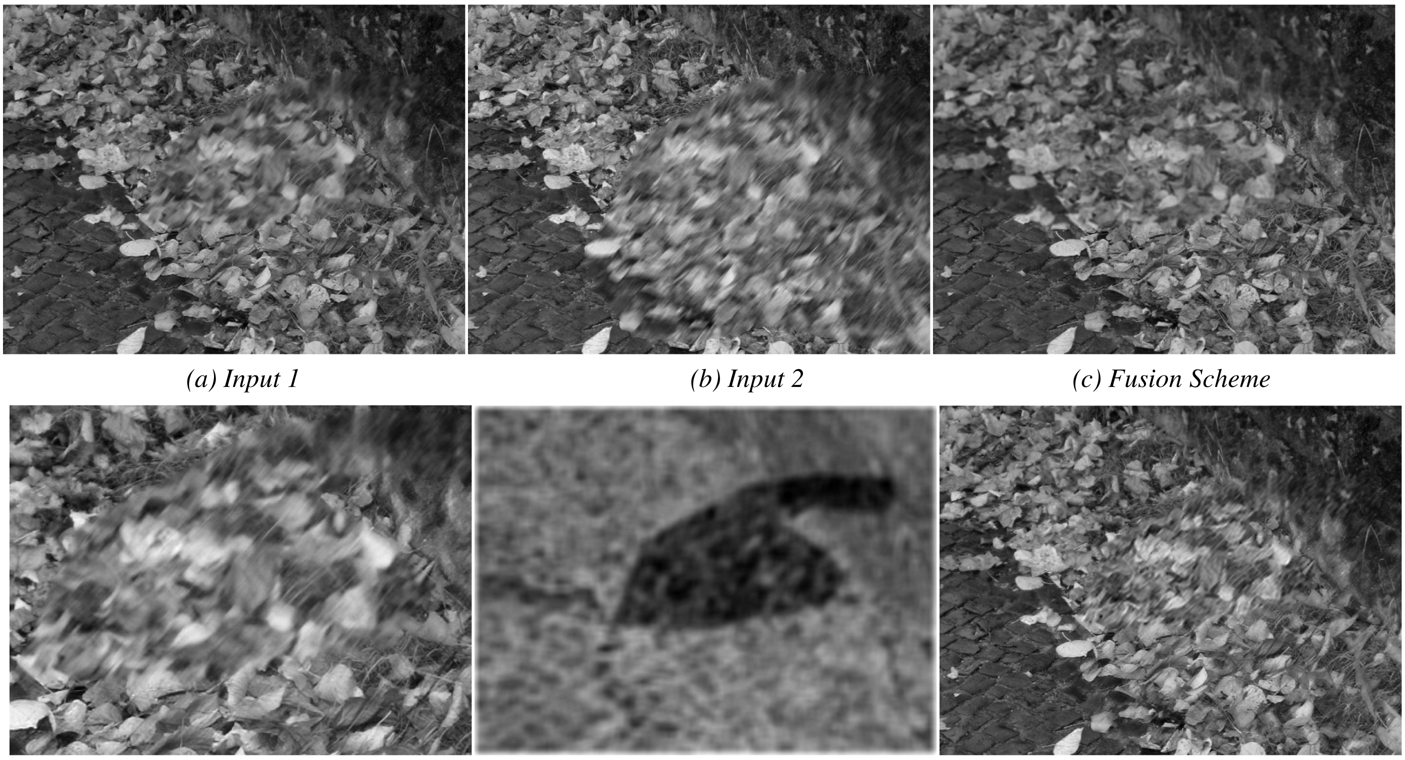
There terms $\eta_1$ and $\eta_2$ are the step size parameters that control the convergence rates for the image and Point Spread Function (blurring image) respectively.

Imperial College has applied robust functionals in the cost functions in order to rectify some of the problems with using double regularisation restoration (e.g., quadratic term penalises sharp grey-level transitions resulting in blurring of image details that are recovered from the images suffering from ringing). This results in modified original cost function:

$$Q\big(h(\mathrm{r}), f(\mathrm{r})\big) = \underbrace{\frac{1}{2} A_1(\mathrm{r}) \rho_n \big(y(\mathrm{r}) - h(\mathrm{r}) * f(\mathrm{r})\big)^2}_{\text{residual}} + \underbrace{\frac{\lambda}{2} A_2(\mathrm{r}) \rho_f \big(C * f(\mathrm{r})\big)^2}_{\text{image regularisation}} + \underbrace{\frac{\gamma}{2} A_3(\mathrm{r}) \rho_d \big(A * h(\mathrm{r})\big)^2}_{\text{blur regularisation}}$$

Three distinct robust kernels, $\rho_n(\cdot)$, $\rho_f(\cdot)$ and $\rho_d(\cdot)$ are introduced in the new cost function and are referred to as the robust, residual and regularising terms respectively. Robust functionals are used to address the problem of outliers that might risk the stability of the update algorithm. Detailed discussion on the derivation and insight to the way these kernels are selected is presented in [24, 25].

Results of fusion, and joint fusion and restoration are presented in Figure 4.



*(a) Input 1* *(b) Input 2* *(c) Fusion Scheme*

*(d) Common Degraded region* *(e) Fusion (Affected Area)* *(f) Fusion + Restoration (Affected Area)*

*Figure 4: Overall fusion improvement using the proposed fusion approach enhanced with restoration.*

## Multi-resolution target detection and identification

An important problem in image analysis is that of finding similar objects in sets of images, where the objects are often at different locations, scales and orientations in the various images. Partial occlusion of objects is also quite common. An effective general approach to this problem is first to find a relatively large number, typically several thousand, of key feature points in each image, and then to develop a more detailed descriptor for each keypoint. This allows points from different images to be compared and matched to create candidate pairings.

Often a reference object is taken from one image and then other instances of the object are searched for in the remaining images, so the number of reference keypoints is quite small (10 – 100), but the number of candidate keypoints can be very large ($10^5$ – $10^7$). Hence it is important to develop keypoint descriptors which allow efficient comparison of pairs of keypoints (reference-to-candidate).

Cambridge University (CU) approached the problem with the Template Matching technique with automatic template update. The technique produced promising results; however, with this scheme the target cannot rotate arbitrarily between consecutive frames. In order to overcome this, CU has adopted a technique of polar matching with dual-tree complex wavelet transform (DTCWT) coefficients.

Polar-matching with DTCWT based technique does not require the dominant orientation(s) to be computed first because it allows efficient matching of descriptor pairs in a rotationally invariant way. Polar matching matrix gives low redundant rotation invariant descriptor in addition to its computational efficiency making it a very effective.

DTCWT is a multi-scale transform with decimated six subbands with complex coefficients. DTCWT is approximately shift invariant, which means that the z-transfer function, through any given subband of a forward and inverse DTCWT in tandem, is invariant to spatial shifts, and that aliasing effect due to decimation within the transform are small enough to be neglected for most image processing purposes.

Another feature of DTCWT is that the complex wavelet coefficients within any given subband are sufficiently bandlimited that it is possible to interpolate between them in order to calculate coefficients that correctly correspond to any desired sampling location or pattern of locations. Hence for a given keypoint location you may calculate the coefficients for an arbitrary sampling pattern centred on that location.

The main thrust of the polar matching method is to assemble dual-tree complex wavelet coefficients from 13 sampling locations around a circle, at 6 subband orientations, and one or more scales, such that they form a 'polar matching matrix', as shown in Figure 5. Consequently, a rotation of the image about the centre of the sampling pattern corresponds to a cyclic shift of the columns of the polar matching matrix.

Polar matrices are formed from the reference image and the search image. If an object of interest in the search image rotates with respect to the reference image, Fourier transform methods can be employed to match and detect the object by performing correlations between the two matrices.

The main innovation of Cambridge's work is the technique for assembling complex coefficients from the sampling locations, subband orientations, and one or more scales such that they form a 'polar' matching matrix *P*, in which a rotation of the image about the centre of the sampling pattern corresponds to a cyclic shift of the columns of *P*.

The cyclic shift property of the matrix *P*, when rotation occurs, means that Fourier Transform methods are appropriate for performing correlations between two matrices $P_r$ and $P_s$ from the reference and search images respectively. It has been shown that this correlation may be performed efficiently in the Fourier domain followed by a single low complexity inverse FFT to recover the correlation result as a function of rotation $\theta$. The peak of this result is the required rotation-invariant similarity measure between $P_r$ and $P_s$. A key aspect is that phase information from the complex coefficients can be fully preserved in this whole process.

The aim is to sample the directional subbands at a given scale on a grid, centred on the desired keypoint, and then to map the data to a matrix *P*, such that the rotations of the image about the keypoint are converted into linear cyclic shifts down the columns of *P*.

The sampling grid that is centred on the keypoint is shown in Figure 6. It is circularly symmetric and the sampling interval is chosen to be $30^{\circ}$ to match that of subbands. There are 12 samples around the circle $(A, B, ..., L)$ and one at its centre *(M)*. The radius of the circle is equal to the sampling interval of the DTCWT subbands at the given scale, as this is an appropriate interval to avoid aliasing and yet provide a rich description of the keypoint locality.

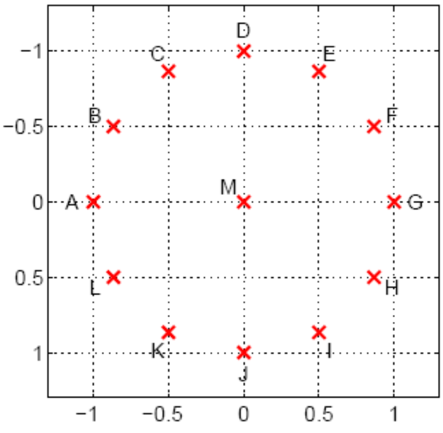


*Figure 5: The 13-point circular sampling pattern for DTCWT at each keypoint location*

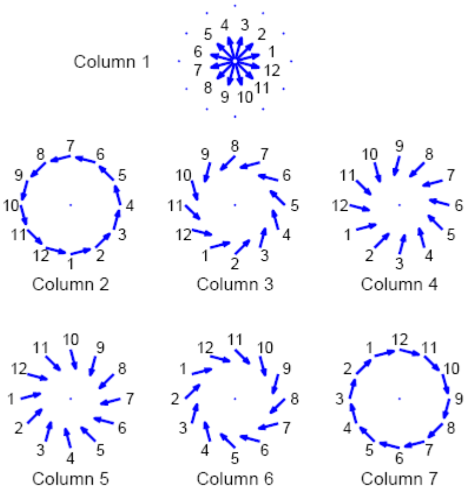


*Figure 6: Shows how each column of the polar matching matrix P comprises of a set of rotationally symmetric samples from the subbands and their conjugates, whose orientations are shown by the arrows. Numbers give the row indices in P.*

Cambridge use a bandpass interpolation technique for obtaining samples on the circular grid around each keypoint.. The information contained in a given directional complex subband is bandlimited to a particular region of 2-D frequency space, which has a centre frequency $\left(w_1, w_2\right)$. Bandpass interpolation may be implemented by:

1. a frequency shift by $\left\{-w_1, -w_2\right\}$ down to zero frequency (i.e. a multiplication of the complex subband coefficients by $e^{[-j(w_1 x_1 + w_2 x_2)]}$ at each sampling point $\{x_1, x_2\}$ ),
2. a conventional lowpass Spline or bi-cubic interpolation to each new grid point,
3. an inverse frequencies shift up by $\left\{w_1, w_2\right\}$ (a multiplication by $e^{[-j(w_1 y_1 + w_2 y_2)]}$ at each grid point $\{y_1, y_2\}$ ).

To simplify notation for the mapping to matrix *P*, for a given keypoint locality $\{A, B, ..., M\}$ in Fig. 5, the 13 subband coefficients are denoted by $\{a_d, b_d, ..., m_d\}$, where $d = 1,2,...,6$ indicates the direction of the subband. The $12 \times 7$ matrix *P* is then formed from the $13 \times 6$ coefficients and their conjugates as shown in the Matrix P.

$$P = \begin{bmatrix} m_1 & j_1 & k_1 & l_1 & a_1 & b_1 & c_1 \\ m_2 & i_2 & j_2 & k_2 & l_2 & a_2 & b_2 \\ m_3 & h_3 & i_3 & j_3 & k_3 & l_3 & a_3 \\ m_4 & g_4 & h_4 & i_4 & j_4 & k_4 & l_4 \\ m_5 & f_5 & g_5 & h_5 & i_5 & j_5 & k_5 \\ m_6 & e_6 & f_6 & g_6 & h_6 & i_6 & j_6 \\ m_1^* & d_1^* & e_1^* & f_1^* & g_1^* & h_1^* & i_1^* \\ m_2^* & c_2^* & d_2^* & e_2^* & f_2^* & g_2^* & h_2^* \\ m_3^* & b_3^* & c_3^* & d_3^* & e_3^* & f_3^* & g_3^* \\ m_4^* & a_4^* & b_4^* & c_4^* & d_4^* & e_4^* & f_4^* \\ m_5^* & l_5^* & a_5^* & b_5^* & c_5^* & d_5^* & e_5^* \\ m_6^* & k_6^* & l_6^* & a_6^* & b_6^* & c_6^* & d_6^* \end{bmatrix}$$

The rational for choosing this mapping can be understood from Fig. 5, which shows each of the columns of *P* in diagrammatic form using arrows on the grid of Fig. 6 to represent the direction of each subband. Hence all the samples in column 1 of *P* are taken at the midpoint *M* and correspond to the 6 subbands and their conjugates taken in sequence.

The arrow labelled '1' is from the $15^\circ$ subband, arrow '2' is from the $45^\circ$ subband, arrow '7' is from the conjugate of the $15^\circ$ subband (i.e. the $195^\circ$ subband), and so on. The circle of arrows for column 2 shows the location and subband from which each element in column 2 of *P* is taken, and this is also shown for the remaining columns. Thus you see that each column of *P* represents a particular pattern of rotationally symmetric combinations of sampling location and subband orientation, such that if an object is rotated clockwise about the centre of the sampling pattern by $k \times 30$ (*k* integer), then each column of *P* will be cyclically shifted *k* places downwards.

In order to perform rotation-invariant object detection, a matching technique is required which measures the correlation between a candidate locality in the search image and all possible rotations of a reference object in an efficient way. The Fourier transform is well-known to be a useful aid to performing cyclic correlations and in conjunction with the mapping to the *P* matrix, as above, it turns out to be effective at performing rotational correlations too. The basic idea is to form matrices $P_{r,i}$ at every keypoint *i* in the reference image, and to form matrices $P_{s,j}$ at all candidate keypoints *j* in the search image.

The pairwise correlation process for each transformed matrix pair $P_{r,i}$ and $P_{s,j}$ then becomes:

1. Multiply each Fourier component of $P_{s,j}$ with the conjugate of the equivalent Fourier component of $P_{r,i}$ to get a matrix $S_{i,j}$ ($12 \times 7 = 84$ complex multiplies).
2. Accumulate the $12 \times 7 = 84$ elements of $S_{i,j}$ into a 48 – element spectrum vector $s_{i,j}$ (84 complex adds).

3. Take the real part of the inverse FFT of $s_{i,j}$ to obtain the 48-point correlation result $s_{i,j}$ ($48\times\log_2(48) = P_{r,i}$ 270 complexes multiply-an-adds).

Here we have concentrated on the theory of CU's technique, that is admittedly quite complicated, and as there is a limited space, a brief account of results will be presented. Cambridge has shown how rotational correlations may be performed using interpolated complex samples from the DTCWT, utilising both phase and amplitude information. There is considerable scope for extending these ideas to increase the robustness to typical image distortions (e.g. due to change of viewpoint or lighting) and small mis-registration of keypoints.

Results of the Polar matching DT-CWT base tracking scheme are shown in Figure 7. The blue graph represents the best polar match score for each frame of the test sequence. It is equivalent to a measure of confidence. The top images show the moment that the target enters the occlusion zone. The corresponding confidence drops dramatically. When this occurs the search window moves in the direction of the last known target trajectory and widens to a larger search area. This is indicated by a red circle in the top left image. The bottom images show the moment that the target exits the occlusion zone. Here, the confidence measure rises significantly and passes a threshold which indicates that the target has been successfully reacquired. The green circle in the bottom left image indicates that the confidence measure is above the defined threshold. Currently, the threshold is chosen by experimentation. Future work will attempt to derive a threshold automatically.

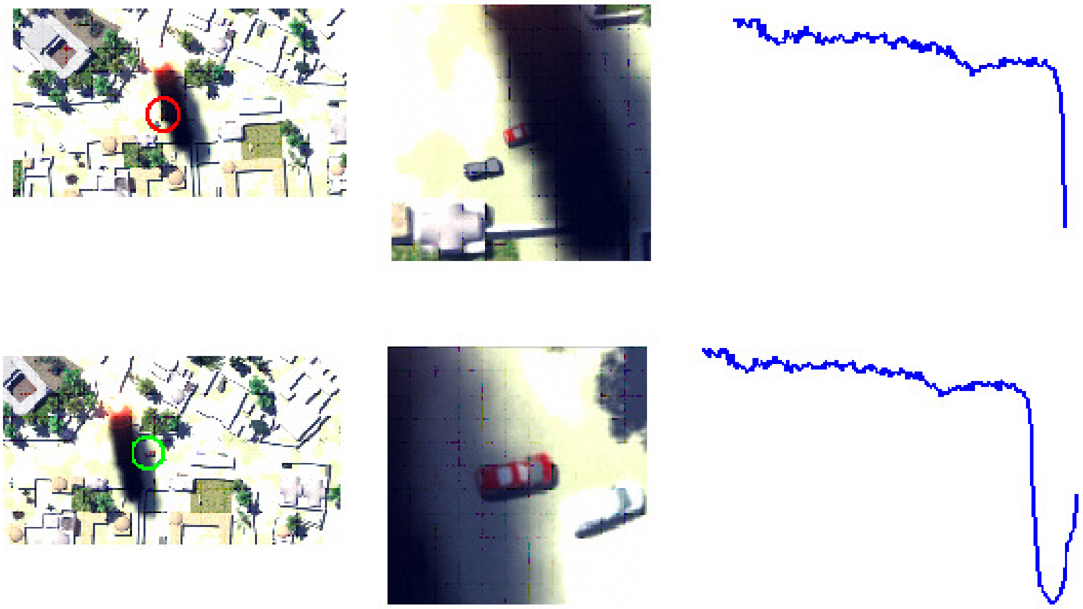


*Figure 7: Tracking results with confidence level shown in blue, entering and exiting occluded zone (smoke)*

## Task based image and video fusion assessment

The widespread use of image fusion methods has led to a rising demand of pertinent quality assessment tools in order to compare the results obtained with different algorithms and systems or to derive an optimal setting of parameters for a specific fusion algorithm.

For man-in-the-loop applications, the performance of the fusion algorithm can be measured in terms of improvement in operator performance in different tasks like detection, recognition, classification, and tracking. This approach requires a well defined task, for which quantitative measurement can be made, and it usually involves costly and time consuming field trials.

Computational image fusion quality assessment metrics that relate to human observer performance are therefore of great value. The assessment can either be done by comparing the fused result with a reference image that provides the ground-truth, or (since such ground-truth is not available in most applications) by relating the fused result (or some of its features) to each of the input images (the so-called non-reference approach). Video fusion assessment is even more challenging as the spatio-temporal characteristics of the inputs need to be taken into account.

Previous experiments conducted at Bristol University have shown that, unfortunately, subjective ranking of fused images/video and computational metric results on the one hand, and human performance for particular tasks, such as tracking, on the other hand, do not correlate well. In other words, fused images and videos that are highly ranked by computational metrics or even by human observers because of their high image quality, do not necessarily lead to improved task performance when shown to human observers.

It was thus decided to focus in the future on developing video fusion assessment metrics that correlate well with and can predict human performance for a particular task. Work on such task-dependant metrics has begun and the plan is to incorporate them into a general framework of metrics that will work for a broad range of tasks.

One of the main focuses of the AMDF Cluster project is to study the effects of resolution (SD vs. HD) and multi-sensor (visible and IR) video fusion on target tracking. Hence, it was decided to study in more detail the influence of pixel-level video fusion on object tracking using a variety of multi-sensor datasets, i.e. visible, FLIR and hyperspectral synthetic sequences from QinetiQ, visible and IR datasets from the Eden Project [9] (see Fig.8) and another visible and IR dataset available in the public domain [10]. The object tracking was done in house in collaboration with the DIF DTC Tracking Cluster project using two different trackers (mean shift [26] and particle filters [27]) available in Bristol.

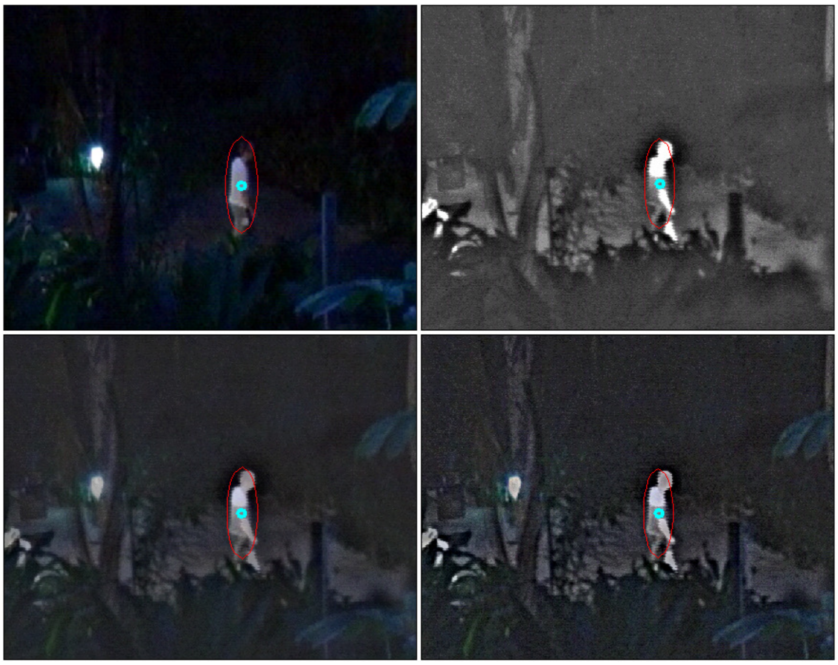


*Figure 8: Tracking in an Eden sequence. Clockwise from top left the results correspond to: visible, infrared; DT-CWT fused and average fusion*

The experimental results suggest strongly that on average, the IR mode is the most useful when it comes to tracking objects that are well seen in the IR spectrum. However, under some circumstances fusion is beneficial. In addition, in a situation when the task is not to simply track a single target, but to determine/estimate its position with respect to another object that is not visible in the IR video, video fusion is essential in order to perform the task successfully and accurately. This is due to the inclusion of complementary and contextual information from all input sources, making it more suitable for further analysis by either a human observer or a computer program. However, metrics for fusion assessment clearly point towards the supremacy of the multi-resolution methods, especially DT-CWT. Thus, a new, tracking-oriented, metric is needed that would be able to reliably assess the tracking performance on a fused video sequence.

## Independent evaluation of the image fusion results

It is generally agreed that image fusion techniques can produce fused images that appear to be at least as good, and hopefully better than, the sum of the input parts. But *proving* how much better a fused image is over the original source images is notoriously difficult. This is essential if the additional costs of multi-sensor systems and associated processing are to be justified.

The method of assessment is greatly dependent upon the application; empirical studies using human observers [15] have illustrated the benefits of fusion for tasks such as object detection and identification, and general situational awareness. These tasks can be performed with higher accuracy and greater confidence when compared to using the source imagery alone. Such experiments also compared grey-scale and colour fusion and concluded that the utility of colour fusion is highly dependent upon the colour mapping.

The benefits of systems whose outputs are interpreted by automatic processing algorithms (for example, target tracking) are generally easier to quantify because clearly defined metrics exist for the tasks that they perform. No equivalent standard set of metrics currently exists for fusion systems that provide imagery for human interpretation.

Part of Waterfall Solution's work is concerned with assessing the performance of image fusion schemes which provide outputs for a number of automatic tasks, principally target detection and tracking. This can be achieved by quantifying the improvement to image quality engendered by the fusion process through the use of appropriate metrics.

Measures of the quality of an image are diverse, from simple image moments (i.e. mean, standard deviation etc.) to edge densities and other image content metrics. These measures are very flexible and can therefore be chosen to match the salient image features that might be exploited - for example, a target detection algorithm.

However, single-frame image metrics are only sensitive to the contents of the current frame, and can therefore give misleading results in the presence of noise or other time-dependent image features. Metrics may also give very different results when presented with two scenes of the same quality, and so must be chosen carefully.

The sensitivity of single-frame metrics to frame content can be mitigated by normalising the metric to the results from one of the sources images to show the relative change in the metric caused by the fusion algorithm. Although this removes sensitivity to changing image content, the result is not bounded.

A novel method to visualise a potentially large number of single-frame image metrics, first proposed by Smith [16], is the polar plot (also sometimes termed Kiviat diagram when used in a control system validation context). In this representation, each normalised metric is plotted on a spoke of the polar plot along with the results for the input images so that an instantaneous comparative 'snap-shot' can be given which encompasses all metrics of interest. An example for five metrics is show below in Figure 9.

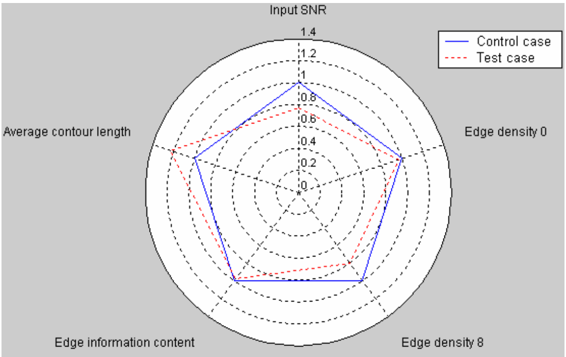


*Figure 9: Polar plot representation of image metrics*

Investigations have shown that single-frame image metrics can sometimes disagree with interpretation of the fused imagery by eye and some (even the more advanced) metrics may exhibit behaviour counter to task-driven measures of performance.

An alternative family of metrics is the set of image validation metrics which calculate the difference (or similarity) between two images for a given characteristic [17]. One or more of the original input modalities is chosen as a reference when assessing output from an image fusion technique using image validation metrics. The majority of image validation metrics also have the advantage that they are automatically bounded between 0 and 1. In most cases values approaching unity indicate that the images are nearly identical.

Examples of image validation metrics are: Cross-correlation, image quality, image structural similarity and peak signal-to-noise ratio. Image quality and image structural similarity metrics have both been proposed by Wang [18] et al. and the peak signal-to-noise ratio was developed by Fisher [19].

A cautionary note on the interpretation of image validation metrics is illustrated by looking at the image structural metric. A value close to 1 would indicate that the fused image has retained much of the structure of the reference input channel. However, a fused image may be of high quality but have a low score. This would occur if the fused image had retained complementary detail present in another input channel. These types of issues can be detected by running the validation metrics using each source image as the control case.

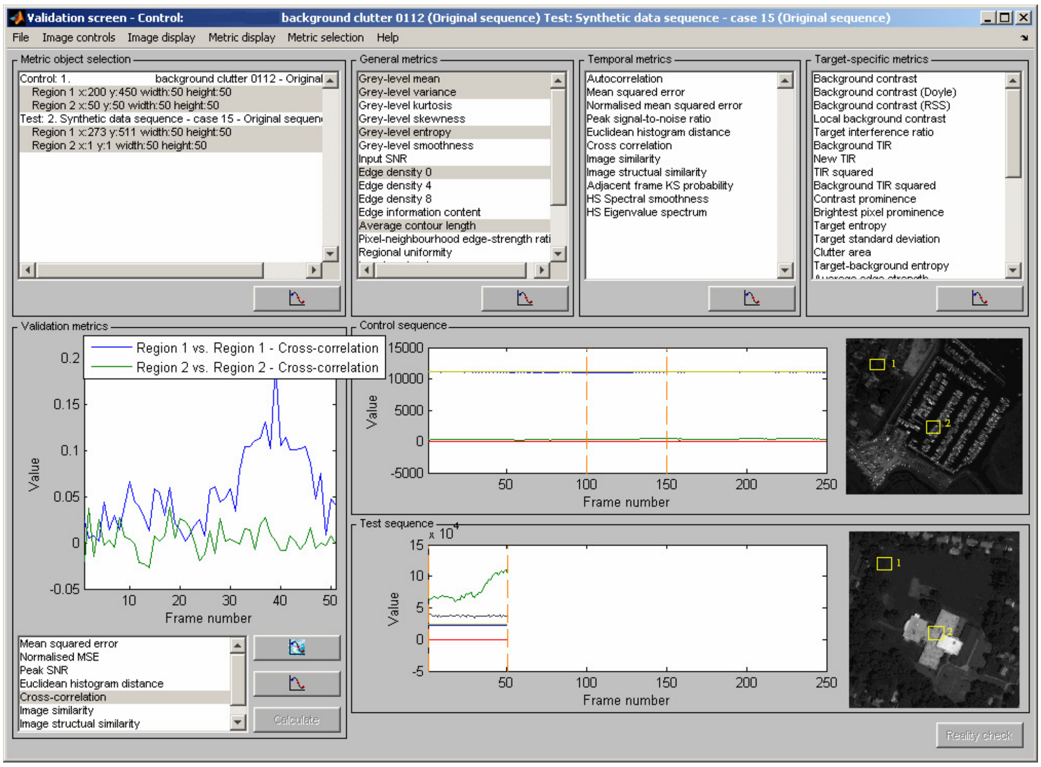

Validation screen - Control: background clutter 0112 (Original sequence) Test: Synthetic data sequence - case 15 (Original sequence)
File Image controls Image display Metric display Metric selection Help
Metric object selection
Control: 1. background clutter 0112 - Original
Region 1 x:200 y:450 width:50 height:50
Region 2 x:50 y:50 width:50 height:50
Test: 2. Synthetic data sequence - case 15 - Original sequen
Region 1 x:273 y:511 width:50 height:50
Region 2 x:1 y:1 width:50 height:50
General metrics
Grey-level mean
Grey-level variance
Grey-level kurtosis
Grey-level skewness
Grey-level entropy
Grey-level smoothness
Input SNR
Edge density 0
Edge density 4
Edge density 8
Edge information content
Average contour length
Pixel-neighbourhood edge-strength rati
Regional uniformity
Temporal metrics
Autocorrelation
Mean squared error
Normalised mean squared error
Peak signal-to-noise ratio
Euclidean histogram distance
Cross correlation
Image similarity
Image structual similarity
Adjacent frame KS probability
HS Spectral smoothness
HS Eigenvalue spectrum
Target-specific metrics
Background contrast
Background contrast (Doyle)
Background contrast (RSS)
Local background contrast
Target interference ratio
Background TIR
New TIR
TIR squared
Background TIR squared
Contrast prominence
Brightest pixel prominence
Target entropy
Target standard deviation
Clutter area
Target-background entropy
Validation metrics
Region 1 vs. Region 1 - Cross-correlation
Region 2 vs. Region 2 - Cross-correlation
0.2
0.15
0.1
0.05
0
-0.05
Value
10
20
30
40
50
Frame number
Mean squared error
Normalised MSE
Peak SNR
Euclidean histogram distance
Cross-correlation
Image similarity
Image structual similarity
Calculate
Control sequence
15000
10000
5000
0
-5000
Value
50
100
150
200
250
Frame number
1
2
Test sequence
x 10^4
15
10
5
0
-5
Value
50
100
150
200
250
Frame number
1
2
Reality check

*Figure 10: Screen shot of Image Analysis Tool, I2DB*

The image validation metrics have equal applicability to temporal sequences in which the reference image is replaced by the previous (or next) image in the sequence. This analysis provides information on flicker and the severity of image transitions which can be a contributory factor in operator stress over long periods.

A particular family of image metrics deserving of mention are spectral metrics. The project will generate synthetic co-incident panchromatic, multi-spectral and hyperspectral imagery each at a progressively coarser resolution.

Teams will utilise the spectral imagery for tasks such as target detection, identification and tracking. Man made targets often posses spectral qualities that are distinct from their natural backgrounds and this contrast can be used to enhance task performance. However, spectral imagery is frequently collected at the expense of spatial resolution. Hence the addition of spectral imagery provides a rich feature space in which to investigate spectral and spatial fusion. The spectral imagery may be fused with other spectral or panchromatic sources to enhance spatial details and the tasks may be executed utilising a mixture of spectral and spatial features. In some cases, the fusion process may be required to reduce the number of spectral bands in an image to limit processing overhead or as necessary for a further stage of processing. Thus, spectral metrics can be used in the same way as spatial metrics to attempt to quantify the changes in the image quality, including spectral information content, engendered by the fusion process. Furthermore, the same metrics can be used a-priori to help to optimise the down-selection of spectral bands in a fusion process.

Image metrics - single-frame, validation, temporal and spectral – underpin any assessment of image fusion. Waterfall Solutions' integrated software tool will be used for trusted, repeatable and multi-faceted image analysis to support the assessment activities.

Figure 10 shows a screen shot of Waterfall Solutions' integrated software tool that will be used for trusted, repeatable and multi-faceted image analysis to support the assessment activities. The figure shows two image sequences, selected from the database, displayed in the bottom right corner. Two complementary regions have been selected in the images. The selected general image metrics are displayed as a function of frame number in the two graphs at the bottom of the figure. In this case, the lower graph is truncated as it is limited by the number of frames in the second sequence. A selected image validation metric is displayed as a function of frame number for the two regions in the graph in the bottom left corner of the image. The data generated by this tool can be automatically imported into a data validation environment for further analysis.

## Future Work

Hyperspectral imaging is widely used in Earth observation systems and remote sensing applications. Modern hyperspectral imaging sensors produce vast amounts of data. Thus, autonomous systems that can fuse "important" spectral bands and then classify regions of interest are required. Hyperspectral image analysis has proved useful in a variety of applications including target detection, pattern classification, material mapping and identification, etc.

At Bristol University, research in this direction has focused on the development of novel algorithms for band reduction in hyperspectral images as well as for subsequent image classification. Different state-of-the-art techniques for dimensionality reduction have been investigated, which are based on entropy, mutual information and independent component analysis (ICA). New techniques based on the universal image quality index instead of entropy or mutual information have been developed and they showed considerable improvement over existing techniques.

Another important topic that is being studied is how to improve the classification of hyperspectral images using image fusion techniques. The main idea is to capture the most important features and salient points of the input bands using image fusion. The information obtained using image fusion techniques can lead to improved target detection and (supervised or unsupervised) classification in hyperspectral imagery. As fusion methods relying on the dual-tree complex wavelet transform and ICA have been shown to be the best performing in the multimodal image fusion, these approaches are now being generalized to work with hyperspectral image data.

Initial developments were made using the wavelet transform, which constitutes a powerful framework for implementing image fusion algorithms [5, 6]. The theoretical limits of many image fusion algorithms are determined by the underlying statistical model. Consequently, the focus was on studying prior probability models that have the potential to better characterize the different bands of hyperspectral images, as well as their associated transform coefficients.

In order to cope with more appropriate statistical model assumptions, the original weighted-average method [7] that combines images based on their local saliency was reformulated and modified. The candidate prior probability models included: generalised Gaussian distributions and alpha-stable distributions. Both were previously applied successfully to modelling natural images and were found to model the heavy-tailed image distributions more precisely than the conventional Gaussian distribution [8, 9].

Additionally, the models of image wavelet coefficients have been amended to account for both interscale dependencies and noise presence in the data. This has been achieved by incorporating bivariate shrinkage functions, derived from the underlying statistical models, into the fusion scheme. Simple and efficient implementations have been achieved with analytic estimators for special cases of the above distributions, namely the Laplace and the Cauchy distributions. In order to estimate all statistical parameters involved in the fusion algorithms a relatively novel framework that of Mellin transform theory was used.

The new method has been shown to perform very well with noisy datasets, outperforming conventional algorithms. The method has also been shown to significantly reduce the noise variance in the fused output images. Figure 11 shows an example of hyperspectral data, taken from [10], fused with the proposed method, compared to the conventional choose-max algorithm. More details on this fusion algorithm can be found in [11, 12].

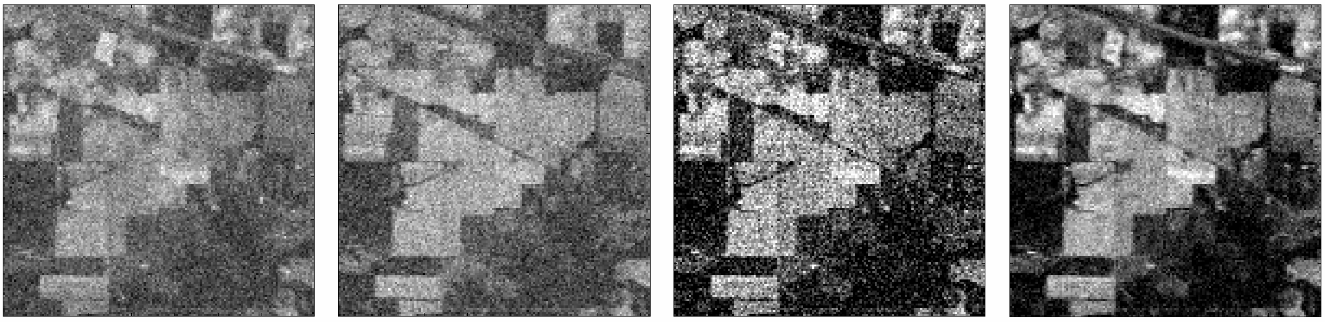


*Figure 11: Statistical fusion of hyperspectral imagery, from left to right: input images from two different spectral bands, a fused image with the choose-maximum method in the wavelet domain and statistical fusion result using Laplacian modelling and bivariate shrinkage functions.*

## Conclusions

AMDF set out to develop academic research into applicable products. Work done to date in the area of super-resolution, joint image fusion and restoration, multi-resolution tracking and task based image fusion metric has yielded exciting results. At the same time practical issues of synthetic scenario generation have been raised. This research work has shown the care needed when using synthetically generated video sequences; the SR research exposed the absence of sub-pixel detail within the synthetic data, whilst the stable synthetic view and 'quiet' simulated environment lacked the real-world occlusions and dynamically changing views of the objects that would prove the efficacy of the new techniques. The military customer has also desired presence of random fires, and high traffic densities that characterise much of the challenges in real-world surveillance data. To this effect commercial partners are working to develop new version of data set to address these issues.

Natural convergence of super-resolution with self organising map based tracking technique will allow view based technique to be applied to difficult problems such as tracking of object exhibiting affine transformations. Work has also shown the ability of the Polar matching Dual Tree Complex Wavelet Transform based tracking technique even under the difficult conditions exhibited by full occlusion. It is planned to bring together all the tracking techniques together into a combined multi-object tracking solution to allow the user to apply these techniques without discrimination. In addition enhancement due to fusion techniques, being developed in the programme, with the ability to register images automatically, super-resolve sections in the frame, enhance common degraded areas in fused images, and track objects in any condition under any circumstances even rotation and scale invariantly are taking the programme to an integrated solution that is applicable to complex urban surveillance and target tracking.

Academic research work carried out to date has shown great potential showing convergence towards more application oriented approach. The next key stage will be to apply research to an enhanced synthetic scenario as a tool where efficacy of the research work to urban surveillance and target tracking can be proved.

***Acknowledgements***: This work has been funded by the UK Data and Information Fusion Defence Technology Centre (DIF DTC) AMDF and Tracking cluster projects.

## References

1. http://www.multigen-paradigm.com/products/standards/openflight/index.shtml
2. N. G. Kingsbury: "Rotation-invariant local feature matching with complex wavelets", Proc. European Conference on Signal Processing (EUSIPCO), Florence, 4-8 Sept 2006
3. N. G. Kingsbury: "Complex wavelets for shift invariant analysis and filtering of signals", Journal of Applied and Computational Harmonic Analysis, vol. 10, no 3, May 2001
4. A. Katartzis and M. Petrou: "Robust Bayesian estimation and normalised convolution for super-resolution image reconstruction", British Machine Vision Conference 2007.
5. S. G. Nikolov, P. Hill, D. Bull, and N. Canagarajah: "Wavelets for image fusion," in Wavelets in Signal and Image Analysis, A. Petrosian and F. Meyer, Eds., pp. 213–244. Kluwer Academic Publishers, 2001.
6. A. M. Achim, C. N. Canagarajah, and D. R. Bull: "Complex wavelet domain image fusion based on fractional lower order moments," in Proc. of the 8th International Conference on Information Fusion, Philadelphia PA, USA, 25–29 July, 2005.
7. P. Burt and R. Kolczynski: "Enhanced image capture through fusion," in Proc. 4th International Conference on Computer Vision, Berlin 1993, pp. 173–182.
8. E. P. Simoncelli: "Modelling the joint statistics of images in the wavelet domain," in Proceedings of SPIE 44th Annual Meeting, vol. 3813, Denver, CO, USA, Jul 1999, pp. 188–195.
9. A. Achim and E. Kuruoglu: "Image denoising using bivariate-stable distributions in the complex wavelet domain," IEEE Signal Processing Letters, vol. 12, no. 1, pp. 17–20, Jan 2005.
10. Online: "Airborne visible/infrared imaging spectrometer," available at http://aviris.jpl.nasa.gov/
11. A. Loza, A. Achim, D. R. Bull, and C. N. Canagarajah: Statistical image fusion with generalized Gaussian and alpha-stable distributions. In 15th IEEE International Conference on Digital Signal Processing (DSP 2007), Cardiff, UK, July 2007 (accepted).
12. A. Achim, A. Loza, D. R. Bull, and C. N. Canagarajah: Statistical modelling for wavelet domain image fusion. In Image Fusion: Theory and Applications, T. Sthathaki Ed. Academic Press, 2007 (to appear).
13. N. Cvejic, S. G. Nikolov, H. Knowles, A. Loza, A. Achim, D. R. Bull and C. N. Canagarajah: "The Effect of Pixel-Level Fusion on Object Tracking in Multi-Sensor Surveillance Video," Workshop on Image Registration and Fusion at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR 2007), Minneapolis, Minnesota, 23 June 2007

14. J. J. Lewis, S. G. Nikolov, A. Loza, E. Fernandez Canga, N. Cvejic, J. Li, A. Cardinali, C. N. Canagarajah, D. R. Bull, T. Riley, D. Hickman, M. I. Smith: The Eden Project multi-sensor data set, Technical report TR-UoB-WS-Eden-Project-Data-Set, University of Bristol and Waterfall Solutions Ltd, UK, 6 April 2006
15. M. Smith, J. Heather: “Review of Image Fusion Technology in 2005”, Waterfall Solutions, UK; Defence and Security Symposium 2005, Orlando, 28 March - 1 April, Conference 5782: Thermosense XXVII: Thermal Image Fusion Applications
16. M. Smith: “The design, verification and validation of a generic electro-optic sensor model for system performance evaluation”, PhD Thesis, University of Glasgow, UK, June 1999
17. C Angell: “Fusion Performance Using a Validation Approach”, Waterfall Solutions, Information Fusion 2005, Philadelphia, 25 - 28 July
18. Wang: “Zhou Wang’s Research Work”, www.cns.nyu.edu/~zwang/files/research .html
19. Y Fisher et al.: “Fractal Image Compression”, (Springer Verlag, 1995), section 2.4, “Pixelized Data”
20. Knutsson H &Westin C (1993). Normalized and differential convolution, *in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 93)*, New York, USA, pp. 515–523.
21. Y. L. You and M. Kaveh: “A regularisation approach to joint blur identification and image restoration.” IEEE Transaction on Image Processing, 5(3):416-428, 1996
22. H. C. Andrews and B.R. Hunt: “Digital Image Restoration” Prentice-Hall 1997
23. Blake, A. and Zisserman, A. *Visual Reconstruction*. MIT Press, Cambridge, MA, 1987.
24. Zervakis, M. and Kwon, T. (1993) On the application of robust functionals in regularized image restoration. Proceedings of the International Conference on Acou tics, Speech and Signal Processing, April, pp. V289- V292.
25. Mitianoudis N., Stathaki T., "Joint Fusion and Blind Restoration for Multiple Image Scenarios with Missing Data”, submitted to Computer Journal.
26. Comaniciu, D.; Meer, P., Mean shift: a robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.24, Iss.5, pp:603-619, May 2002
27. Perez, P.; Hue, C.; Vermaak, J.; Gangnet, M., Color-Based Probabilistic Tracking Computer Vision - ECCV 2002 : Proc. 7th European Conference on Computer Vision, Copenhagen, Denmark, pp:661-675, May 28-31, 2002.