

Temporal resolution enhancement from motion – application to airborne imagery

We describe progress in the second year of the EMRS DTC TEP theme project entitled “Temporal Resolution Enhancement from Motion”. The aim is to develop algorithms that combine evidence over time from a sequence of images in order to improve spatial resolution and reduce unwanted artefacts. A C++ implementation of an algorithm was developed in year one [1]. Work in year two has improved the efficiency and extended the applicability of the algorithm. New schemes for information update and scene matching have substantially reduced the processing time and enabled application of the technique to imagery with more complicated viewing geometries. The new technique is demonstrated using airborne infra-red imagery data sets from a Wescam MX series turret on a helicopter.

By

[M. P. Rollason and A. P. Gardner](#)

QinetiQ Limited, 1105/A7, Cody Technology Park, Ively Road, Farnborough, GU14 0LX

Introduction

We set out to develop temporal resolution enhancement (TRE) techniques that exploit a sequence of images in which there is relative motion between sensor and scene/target, to provide step-change improvements in target acquisition, target identification and scene reconstruction performance.

An information theory argument suggests a benefit from processing multiple images. As we increase the number (m) of images processed, the amount of information available about the scene increases proportionately (given sufficient motion). In contrast, the dimensionality of the 'state' to be inferred is almost constant, because it is dominated by the scene description (rather than geometric and photometric transformation parameters that are relatively low-dimensional). Asymptotic analysis implies that spatial resolution will improve by up to m times in each axis compared with a single image frame.

The limitation in exploiting this information is in the ability to formulate and efficiently solve the inference problem. Significant progress has been made in this direction through the development of high dimensional Bayesian inference approaches [2, 3] and research in the image processing related disciplines of super resolution [4], optical flow [5], track before detect [6], structure from motion [7] and scene reconstruction. Case studies using image data from a variety of military application domains will provide quantitative results and immediate push-through into higher technology readiness (TRL4-6) programmes for a range of airborne imaging systems (surveillance EO turret, missile, UAV or fast jet).

In this paper we demonstrate TRE of infra-red airborne imagery to produce resolution enhanced estimates of (i) a tracked target and (ii) a scene swath. Resolution enhanced mosaic sequences are presented for the latter.

Military relevance

TRE will improve the effective resolution of legacy hardware through software-only upgrade. For future systems it will enable the use of smaller, lighter and cheaper sensors for a given required level of performance. We now detail three application areas for the technology.

1. TRE will obtain enhanced resolution target images for presentation to a human or to an automatic target

identification system, leading to increased identification range in both EO and IR imagery. There is a strong military need for the technologies in fast-jet air-to-surface systems. There is also potential for immediate exploitation into airborne maritime surveillance systems, where TRE could improve human identification of fast inshore attack craft (FIAC) and similar threats.

2. TRE will provide enhanced resolution (background) scene images in a moving sensor system (weapon, UAV, conventional aircraft, ship, land vehicle). This will multiplicatively improve target detection performance in modern systems that detect targets as outliers from the background (e.g. using their relative motion), especially in low pixel-count legacy sensors. Technology insertion into low-cost software-only upgrades of a wide variety of systems is feasible, including air-to-air missiles.
3. TRE will provide enhanced resolution scene reconstructions or mosaic images for mapping applications and automatic enhancement of surveillance and reconnaissance imagery. An example is UAV mapping of an urban area in preparation for a military mission.

Technical approach

The technical approach was detailed in [1] presented at the 2007 EMRS conference and is summarised here.

For a Bayesian formulation of the inference problem we require a generative model of the sensor system. This models the process by which the low resolution image I , measured at the detector array, is formed from the actual scene S .

The generative model must account for the optical system and detector performance and so includes the effects of optical blur by the sensor point spread function, detector aliasing, and detector noise.

The aim of TRE is to exploit a sequence of images in which there is relative motion between scene and sensor. Motion is expressed within the framework as transformations which project each sensor image onto the scene estimate. Such transformations comprise both geometric parameters A (e.g. the aspect of the sensor to the scene) and photometric parameters P .

The generative model can be written $p(I | S, P, A)$. The inference problem is to obtain $p(S, P_{1:t}, A_{1:t} | I_{1:t})$ where the time index t has been introduced to represent a sequence of data. Other information that must be made available includes prior distributions on S , P and A . The prior distribution on S will typically encode spatial structure constraints. (These have had extensive research because they are the basis of single frame super resolution and image compression).

In order to process a sequence of images, any temporal dependency between successive realisations of P and A must be expressed in the form of a dynamics model. For example, the dynamics $p(A_t | A_{t-1})$ on the geometric parameters expresses the rate of change of viewing aspect; i.e. the optical flow.

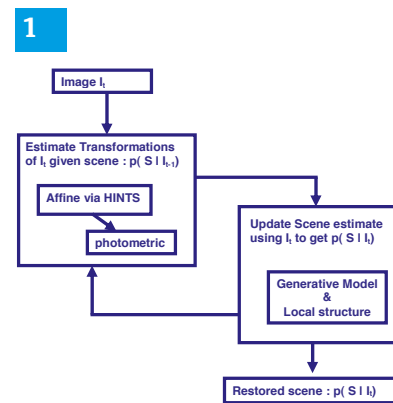
To jointly estimate the scene and a sequence of transformations is a challenging inference problem in Bayesian terms: a target tracker or navigation estimator would typically have 5 or 10 state variables, whereas our super resolved scene has thousands (one per sample point). Therefore decomposition of the inference problem into smaller parts and further approximations are required. The challenge is to ensure that the decomposition does not introduce significant loss of accuracy compared with the full formulation.

We have chosen to decompose the inference problem and apply methods appropriate to the dimensionality of the problem:

- Scene points $\sim 10^5$ (a grid of intensities super sampled with respect to the sensor image)
- Transformations ~ 10 times the number of frames

Algorithm overview

The most important quantities are (i) the “Scene” estimate (which is stored as a grid of points super sampled with respect to the sensor image) and (ii) a list of “View” structures that each contain the geometric and photometric transform parameters for a single sensor image. The “View” structure comprises 6 affine parameters and 2 photometric (gain and offset). Each incoming frame (sensor image) is considered in the context of the current scene estimate. Figure 1 illustrates the processing of each new sensor image, which proceeds in two stages.



Algorithm overview.

The first stage uses an efficient Bayesian inference scheme (HINTS) to infer the geometric transformations bringing the incoming image into alignment with the stored scene estimate. Within this stage, maximum likelihood values for the photometric parameters are obtained directly.

The second processing stage updates the scene estimate to take account of the new image.

The scheme developed in year 1 iterates through all scene points (in pseudo random order) and updates their values based on a conditional distribution relative to their neighbours. This conditional distribution is derived from both the generative model (applied over every low resolution pixel to which the scene point contributes) and from the spatial structure model.

The differentiability of the generative model allows an error gradient to be calculated for individual scene points and they are updated one at a time. However, this process is relatively inefficient compared with the *whole scene* update scheme which has been developed this year. This *gradient based* scheme re-uses some of the generative model calculations and seeks to minimise an error function comprising (i) a local smoothness penalty and (ii) a reconstruction error computed as the difference between measurement image pixel value and the corresponding pixel value predicted via the generative model, given the estimated transformations and scene. This type of gradient update scheme is a Landweber iteration [8] and is very different from temporal averaging because it directly exploits the generative model.

The *whole scene* gradient update scheme has been implemented using whole image operations (ie convolutions) so that the update scheme could be carried out quickly by parallel architectures such as a FPGA.

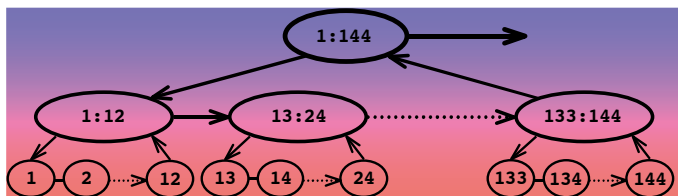
Details of inference of transformations

A multi-scale pairwise state space search was used in year 1 to identify up to 4 geometric states (two translations, plus rotation and horizontal scale). This approach, however, is not scalable as the dimensionality of the transformation state vector is increased to 6 (full affine).

In the pairwise search, evaluating the match score for each state vector requires summing, over every pixel in the incoming image, of the errors between predictions and observations (ie the reconstruction error at the proposed alignment state vector).

An efficient Bayesian scheme, HINTS, is ideally suited to this problem because it can exploit the additive structure of the error function to obtain a very rapid search of the multi-dimensional parameter space. The scheme is detailed in [3] and uses a hierarchical structure (Figure 2) where changes in state at the top (root) level are the result of sequences of Metropolis-Hastings steps taken at the lower levels. At the lowest level (leaves) error function evaluations are carried out for *individual* (or groups of just a few) pixels.

2



HINTS sampling architecture.

The HINTS scheme provides for a substantial speed up c.f. the multi-scale pairwise search which it replaces (e.g. > 10x for four dimensions in the state vector comprising translation, rotation and horizontal scale).

Algorithm summary

Key elements of the TRE processing can be summarised as follows. The TRE algorithm:

- models spatial structure (local smoothness) in the scene;
- models the sensor point spread function and detector noise;
- combines a generative model of the image formation process with the estimate of the resolution enhanced scene and transformations to predict sensor resolution measurements;
- updates the scene estimate so as to minimise the error between each received image and the image predicted from the scene estimate;
- infers full affine transformations (6 parameters) plus whole image gain and offset between resolution enhanced scene and each received image. This transformation is valid for a plane viewed at distance.

Software description

The TRE process is computationally demanding, and so it was determined at an early stage that an efficient implementation (in a compiled language) would be needed to perform evaluation with significant amounts of test data. Therefore a C++ implementation of the algorithm has been developed during the project. Object-oriented design features of the C++ language have been exploited to obtain a flexible and extensible implementation to support both algorithm research and large-scale evaluation. Parameters defining the input imagery, the generative model, and the processing to be applied, are specified through a settings file. Processing options which may be specified according to data or performance requirements include:

- whether to track a selected object as it moves within the sensor's FOV (e.g. for target identification), or whether to process a defined region that is fixed within the sensor's FOV, (e.g. for scene mapping);
- choice of update scheme – individual scene point or whole scene;
- choice of structure model – defining local spatial smoothness;
- Recursive/Batch update – whether each scene update should take a contribution from a number of input frames, or just the latest – (and for the former, the length of the time window – and whether to revisit previously received input images and improve the estimate of their transformations).

Results

Experiments with synthetic data are useful for validation of the algorithm, however they do not account for 'modelling error': the disparity between the assumed generative model and the way the data was actually formed. This can be catastrophic for Bayesian methods and therefore evaluation with real image data sets are the major focus of this project.

Airborne IIR imagery sequences from an MX series turret have been made available by L-3 WESCAM. This section presents three results for two scenarios:

- A tracked target;
- Scene mapping.

Tracked target

A region of interest (ROI) identified by the operator is tracked as it moves within the airborne sensor's field of view (FOV).

Figure 3 shows an example sensor image and the TRE output after 100 frames. It should be noted however that most of the resolution enhancement has been achieved after ~ 30 frames.

This result shows good resolution enhancement and clarifies details such as the number of spokes in the car's front wheel that are not visible in the sensor imagery *even when viewed as a sequence*.

Scene mapping

A region of interest fixed relative to the sensor axes (e.g. on boresight) has been processed and two outputs are presented:

- Individual resolution enhanced scene estimate;
- Resolution enhanced mosaic.

A sequence comprising 199 images of a scene that was viewed obliquely provides the input to the TRE algorithm. The six parameter affine transformations are a valid approximation for this viewing geometry. Figure 4 shows the six (non-zero) affine coefficients for each frame in this sequence.

3

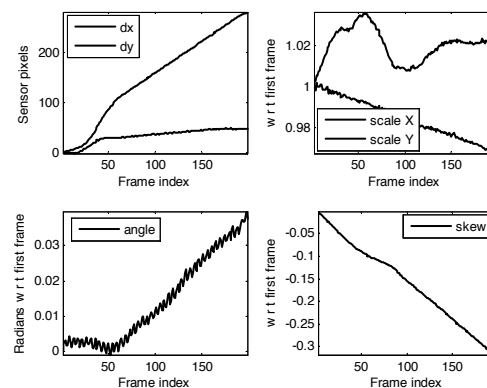


Resolved features:
Wing mirror
Five spokes visible

(top) IR sensor image (41 by 71 pixels); (bottom) TRE result for a target tracked as it moves within the sensor FOV.

The resolution enhanced, and noise reduced, output is presented in Figure 5. Figure 6 provides a comparison of the TRE output with the raw sensor imagery for the part of the scene identified by the orange box in Figure 5.

4



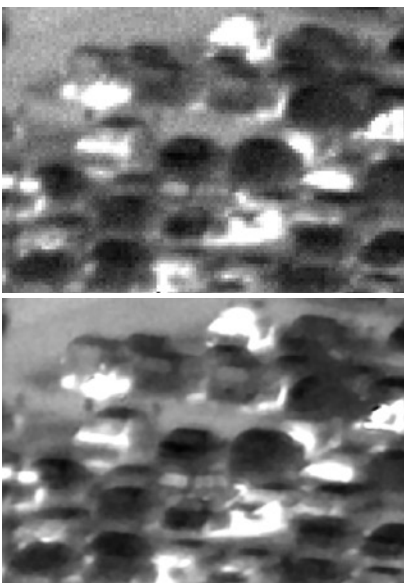
Affine parameters in scene mapping imagery sequence.

5



TRE output for obliquely viewed scene.

6



Enlarged view of selected region in Figure 5. (top) sensor image; (bottom) TRE output.

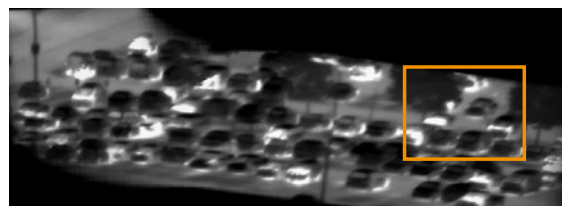
Mosaicing

Resolution enhanced mosaics can be constructed from the sequence of affine transformations and resolution enhanced scene estimates.

Degrees of freedom in the design of mosaic construction include (i) the weighting of pixels in each contributing frame to pixels in mosaic; and (ii) the choice of portion of the contributing frame over which matching is computed.

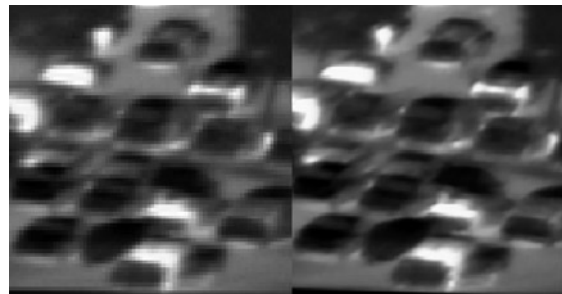
A mosaic scheme which seeks to construct a mosaic whose intensity values are the median of contributing pixels was used to produce the result in Figure 7. This mosaic was constructed using the sequence of resolution enhanced scene outputs from the TRE algorithm, and its transformations output. A sensor resolution mosaic was constructed (not shown) with the affine transformations between frames computed using *optic-flow* [10]. Figure 8 provides a comparison of the TRE mosaic and the sensor resolution counterpart for the patch identified by the box in Figure 7.

7



Mosaic constructed from TRE output.

8



Comparison of mosaics constructed using (left) sensor resolution imagery; (right) TRE output; (enlarged box from Figure 7).

The mosaic constructed using the TRE output is less sharp than the corresponding individual resolution enhanced scene estimates. A likely cause would be the cumulative effect of distortion due to the affine approximation. Projects designed to tackle distortion in mosaicing exist within this DTC [9].

Number of frames required for TRE

In the scene mapping mode of operation, ie when processing a region of interest (ROI) that is fixed relative to the sensor's FOV, the number of frames that are required before the output achieves a stable level of resolution enhancement is dependant upon two factors (i) the rate of progress of the ROI across the scene and (ii) the size of the ROI. Fast progress requires a faster learning rate such that newly observed bits of the scene are learned quickly before they progress across the scene estimate and degrade the inference of transformations. A faster learning rate limits the potential resolution enhancement as it causes the scene estimate to be more influenced by the later measurement images. Increasing the size of the ROI enables a lower learning rate to be used since the newly observed parts of the scene represent less of a fraction of the whole scene and their influence on the inference of transformations is reduced (for a given rate of progress).

For the tracked target (Figure 3) a stable level of resolution was achieved after approximately 30 frames. The TRE output for the scene mapping sequence in Figure 5/6 stabilised after fewer frames (<20).

Processing requirement

For the results in the proceeding section, the TRE algorithm was configured to process each incoming image only once (a recursive processing scheme) and used the gradient based scheme for *whole scene* update.

When inferring the affine view transformation with the HINTS scheme, the processing time taken per frame (on a standard desktop PC) is ~20 seconds, with time approximately equally split between inference of transformations and scene update. This time relates to a sensor image patch of 151 by 151 pixels (with the scene grid having a 3:1 ratio of linear points).

Smaller image patches, e.g. the 41 by 71 car sequence of Figure 1 proceed at a rate of approximately 1 frame per second.

We have designed the algorithm so that it is implementable on an FPGA architecture (e.g. the whole scene gradient update scheme is based on convolutions). Real time implementation should be feasible for small ROIs (e.g. that in Figure 1). Potential algorithmic speed up via further approximations should allow larger patches to be processed in real time. A real time feasibility study will be carried out in year three of this project.

Conclusions and future work

We have developed a flexible algorithm for temporal resolution enhancement that has broad applicability for improving the performance of imaging sensor systems for a variety of tasks including acquisition and target identification.

We have focused on the need for an efficient decomposition of the inference problem that reduces computational costs but retains accuracy, and on addressing the modelling error that is inevitable when working with genuine sensor data.

Work in year two has improved the efficiency and extended the applicability of the algorithm. New schemes for information update and scene matching have substantially reduced the processing time and enabled application of the technique to imagery with more complicated viewing geometries.

The algorithm has been shown to be effective for resolution enhancement of airborne infra-red imagery. Very good resolution enhancement was demonstrated for a target that was selected and subsequently tracked within the sensor's FOV. A second processing mode, scene mapping, performs TRE of a selected region that is fixed within the sensor's FOV. Resolution enhanced mosaics have been constructed using a sequence of resolution enhanced scene estimates, for imagery of a scene which was viewed obliquely.

The TRE algorithm is able to account for gradual depth variation across the scene geometry and out of plane rotations of target or scene.

In year three we will need to account for distortion due to depth variation in the scene geometry. This requirement suggests the need to augment the scene with depth information, which will be exploited by a new set of 3D transformations that will replace/augment the affine transformations describing the 2D projection. A rigid body transformation would be one possible 3D transformation. Separately moving objects within the scene would require multiple models. The newly included efficient Bayesian inference scheme, HINTS, will be used in the new framework to seek the expanded set of transformation states.

The ability of the 3D algorithm to produce resolution enhanced imagery of a scene which includes depth discontinuities will be assessed using airborne IR imagery.

A formulation designed to provide a 3D reconstruction of passively imaged targets will be compared, in terms of recognition performance, with LIDAR techniques using the PATRICA data set.

Acknowledgements

The work reported in this paper was funded by the Electro-Magnetic Remote Sensing (EMRS) Defence Technology Centre, established by the UK Ministry of Defence and run by a consortium of SELEX Sensors and Airborne Systems, Thales UK, Roke Manor Research and Filtronic.

The authors are grateful for the airborne IR imagery data provided by L3-Wescam.

References

1. Strens, M J A and. Rollason, M P, 'Temporal Resolution Enhancement from Motion' Electro-Optic Systems, Embedded Processing and Devices, EMRS DTC Annual Technical Conference, 2007.
2. Strens M J A et al. Markov Chain Monte Carlo Sampling using Direct Search Optimization, Proceedings of the Nineteenth International Conference on Machine Learning, San Francisco: Morgan Kaufmann, 2002.
3. Strens M J A et al. Efficient Hierarchical MCMC for Policy Search, accepted for the Twenty-first International Conference on Machine Learning, 2004.
4. Baker S and Kanade T. Limits on super resolution and how to break them. IEEE Transactions on Pattern Analysis and Machine Intelligence 24:1167–1183, 2002.
5. Irani M and Peleg S. Super Resolution From Image Sequences, ICPR, 2:115–120, June 1990.
6. Strens M J A and Gregory I N. Tracking in Cluttered Images, Journal of Image & Vision Computing 21(10), pp 891–911, Elsevier, 2003.
7. Faugeras O. Three-Dimensional Computer Vision: A Geometric Viewpoint. MIT Press, 1993.
8. Landweber L. An iterative formula for Fredholm integral equations of the first kind. American Journal of Mathematics 73:615–624, 1951.
9. Turkbeyler E and Harris C. Bundle adjustment of mosaics using features.
10. Lucas B D and Kanade T. An iterative image registration technique with an application to stereo vision. Proceedings of Imaging understanding workshop, pp 121–130, 1981.