

# The Evolution of the TINA Stereo Vision Sub-System

A. J. Lacey, N. A. Thacker, P. Courtney and S. Crossley

Last updated  
21 / 02 / 2002

This document forms part of the **Features and Measurement Series**  
available from [www.tina-vision.net](http://www.tina-vision.net).

- 2002-005 Tutorial: An Empirical Design Methodology for the Construction of Machine Vision Systems.
- 1996-001 Algorithmic Modelling for Performance Evaluation.
- 2006-002 A Statistical Framework for Detection of Connected Features.
- 2006-007 Quantitative Verification of Projected Views Using a Power Law Model of Feature Detection.
- 1997-003 Tutorial: Supervised Neural Networks in Machine Vision.
- 1995-003 Invariance Network Architecture.
- 1992-001 Combining the Opinions of Several Early Vision Modules using a Multi-Layer Perceptron.
- 2005-006 Curve Fitting and Image Potentials: A Unification within the Likelihood Framework.
- 1996-002 Tutorial: The Likelihood Interpretation of the Kalman Filter.
- 1994-003 Using a Switchable Model Kalman Filter.
- 2004-012 Tutorial: Computing 2D and 3D Optical Flow.
- 2005-011 Comparing the Performance of Least-Squares Estimators: when is GTLS Better than LS?
- 1994-001 Tutorial: Overview of Stereo Matching Research.
- 1992-002 Online Stereo Camera Calibration.
- 1995-002 Calibrating a 4 DOF Stereo Head.
- 2000-009 An Evaluation of the Performance of RANSAC Algorithms for Stereo Camera Calibration.
- 2001-011 The Evolution of the TINA Stereo Vision Sub-System.
- 2007-011 A Methodology for Constructing View-Dependent Wireframe Models.



Imaging Science and Biomedical Engineering,  
School of Cancer and Imaging Sciences,  
University of Manchester, Stopford Building,  
Oxford Road, Manchester M13 9PT, U.K.

# The Evolution of the TINA Stereo Vision Sub-System

A. J. Lacey, N. A. Thacker, P. Courtney  
Imaging Science and Bio-medical Engineering  
University of Manchester, Stopford Building  
Oxford Rd., Manchester M13 9PT. UK  
`neil.thacker@manchester.ac.uk`

S. Crossley\*  
Dept. of Electronic & Electrical Engineering  
University of Sheffield  
Mappin St., Sheffield S1 3JD. UK

## Abstract

In the last 10 years the TINA vision system has been under continual development and expansion with the stereo sub-system evolving considerably. In this paper we document this evolution outlining the stereo requirements of the 3D wireframe model matcher in TINA and explaining why PMF was designed as a solution. We describe how the algorithmic constraints within PMF (a classical feature-based solution to the stereo correspondence problem) were then re-formulated as the Stretch Correlator (SC), a feature driven, area-based solution. We explain how the computational benefits of the SC implementation were exploited in the development of the Video Convolution Processor (VCP), a prototype VLSI device designed and fabricated to accelerate the fundamental SC processes. The SC algorithm was extended to integrate data from temporal sequences of stereo images in order to improve robustness and cope with dynamic environments. Finally, we discuss how and why multi-scale techniques were incorporated into the temporal-stereo SC algorithm, leading to the MS-TSC (Multi-Scale Temporal Stretch Correlation) algorithm. As TINA is completely open source, all of the code is available for free download from <http://www.tina-vision.net>.

## 1 Introduction

In the late 1980's researchers from the Artificial Intelligence Vision Research Unit (AIVRU) at the University of Sheffield began developing the TINA vision system. The initial focus of this endeavour was the development of a 3D wireframe model matching system (3DMM) that could locate a known object in a scene from stereo camera data, with sufficient accuracy to guide a robotic arm. Tackling such a task required solutions to several key problems in machine vision including, feature extraction, geometric fitting and robust view-based model matching. It also required the development of a complete depth from stereo system known as PMF [13]. The algorithms developed were published independently as solutions to the problems they targeted, but all were developed within the unifying framework of TINA. Development of the entire TINA vision system continued to extend existing solutions as well as research techniques to tackle new image understanding problems. In this time there have been many developments to the stereo depth estimation sub-system.

### 1.1 The TINA 3D Object Location System

The original version of the 3D model matcher (3DMM) was presented in [15] and [14]. Briefly the system used a sparse edge based depth map extracted from pairs of binocular stereo images together with the corresponding camera calibration information. A geometric interpretation of the scene was constructed by fitting lines and arcs to the depth map data. Statistical matching of 3D scene descriptions to a stored wireframe model enabled the location of the model within the scene to be identified. Recently we have returned to the 3D wireframe model matcher, introducing a verification feedback path that closes the loop on the existing processing architecture. This has the effect of eliminating many of the assumptions required to generate the initial estimate of object location. This work was recently presented at the BMVC [8]. Figure 1 shows how all the components in the 3D model matcher are integrated.

### 1.2 Feature-based vs. Area-based Stereo

As is clear from Fig. 1, stereo forms a key component in the 3DMM system and has therefore been the focus of much research within TINA. Algorithms designed to solve the stereo correspondence problem can basically be divided into two groups; area-based (ABS) and feature-based (FBS). The usual reason for employing area-based solutions

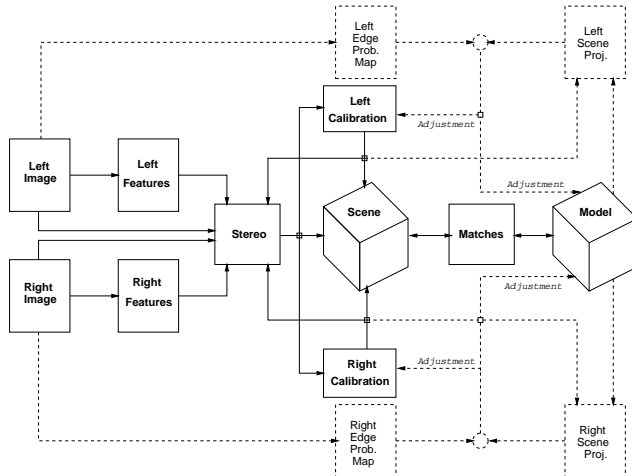


Figure 1: Block diagram of the 3D object location system in TINA. The dashed regions show the closed loop validation stages used to optimise the final object location.

is to produce dense estimates of disparity (typically at a density equivalent to the number of pixels in the image) in order to approximate a continuous surface for the scene. Such techniques are commonly used in photogrammetry applications such as terrain mapping from satellite images [12]. ABS algorithms work best on highly textured scenes and often fail in man-made environments due to the low signal-to-noise ratio in continuous areas of the image, e.g. plain coloured walls. In contrast feature-based approaches only attempt to solve correspondence at image locations where sufficient signal-to-noise has already been identified. In typical man-made scenes this leads to sparse estimates of disparity (usually less than 5% of image pixels). However, the reliability of these matches is nearly always better than that of average ABS results. Solving the correspondence problem using features is aided by incorporating individual feature properties as well as their spatial relationships with neighbouring features. For example, an edge element (edgel) can have properties of strength (the grey-level gradient across the edge) and orientation, both of which can be used as parameters in a similarity metric.

The location accuracy of the 3D results from FBS is higher than ABS as feature detection algorithms such as the Canny edge detector [4] can locate to sub-pixel accuracies. Accepting the inevitable problem of outliers [18] the expected error on the 3D reconstruction of data from an edge-feature-based matching algorithm is given in Appendix A. The dominant factors affecting accurate 3D reconstruction are the proportional errors on distance for both X and Y and the quadratic dependence on distance of the Z measurement. Another important factor is the effect of structural feature orientation, which conspires with any inaccuracies in camera calibration to destabilise measurements of Z for horizontal features. As constraints for positional location of world features can only be determined at such image features these error characteristics are fundamental to the data available in both FBS and ABS. However, for ABS there is also the additional problem of selecting the correct local fitting function with which to interpolate the data. In theory this can only be solved if we know enough about the scene to select the correct interpolation function for each part of the image, i.e. we need to know what objects we are looking at. In fact we believe that a system capable of extracting dense stereo from arbitrary scenes can only be done reliably in the context of a larger vision system that also has quite sophisticated scene interpretation capabilities.

Computationally, whilst the raw load may be lower for FBS algorithms, they tend to utilise complex control mechanisms together with complex data structures. Therefore they are not usually as amenable to hardware acceleration using DSP devices and parallel processing techniques as ABS. However, we will show how a FBS algorithm can be reformulated into an area-based framework in order to exploit the inherent computational benefits.

## 2 Camera Calibration

Much has been published on the reconstruction of scene structure using projective geometry, (the so-called uncalibrated techniques). These methods are generally designed to work with static scenes [1]. In our system we wished to be able to extract structure from scenes containing unconstrained motion. In addition, for robotic tasks, although it is true that the visual feed-back loop can be closed using projective geometry [2], much more can be achieved with good Euclidean geometry. Indeed, calibration of Euclidean geometry is not difficult if we are prepared to make use of all of the information available in a scene and apply relevant statistical methods. Camera calibration

in TINA works in this way, operating in the context of an iterative robust optimisation with optimal combination of covariances via the use of a regularisation term. It is able to track an existing calibration, integrating statistical information from known epipolar errors, robot motion, known object as well as standard calibration targets [19]. It was also extended to support calibration of a 4DOF stereo head system [17].

### 3 PMF

The issue of outliers (due to incorrect matches, specularities, illumination artefacts or smooth occluding boundaries) together with the difficulty of selecting an appropriate interpolant function (which must also deal with discontinuities), means that it is not possible to model for accuracy the performance of ABS with a fixed regional constraint function in the same way as FBS. Also it follows that empirical evaluations will always vary depending upon the test data set selected. This is an important point if error estimates are needed for reliable use of the data in subsequent stages of processing within a vision system [18]. This is why we have adopted a FBS approach to the representational form of the data from our algorithms.

PMF [13] was the original depth from stereo algorithm developed within TINA and has been cited many times as a successful feature based algorithm. PMF uses edges detected using the Canny [4] operator. Potential edgel correspondences are identified as edgels with the same contrast polarity and similar orientations. The algorithm is built upon several constraints for solving the stereo correspondence problem that were derived from the properties of the surfaces in the world being viewed.

#### 3.1 Epipolar Constraint

Solving the correspondence problem can be greatly simplified because, for all practical stereo camera configurations, epipolar geometry can be exploited to reduce a 2D search of order  $N^2$  (for an image of  $N \times N$  pixels) to a 1D search of order  $N$  for each corresponding pair of points. For parallel cameras the epipolar lines lie parallel to the image rasters, however parallel cameras are rarely used in practice. This is because the errors in the depth estimates are inversely proportional to the camera baseline (separation), but the larger the baseline, the less scene overlap there is between the cameras and the less depth computed. To overcome this problem, cameras are most often deployed in a convergent configuration. Unfortunately, this also has the effect of mis-aligning the epipolars and image rasters as well as causing coincidental changes in the local image characteristics. In order for convergent systems to use the epipolar constraint the image data (or the extracted features) must be rectified from their original camera co-ordinates into a parallel camera co-ordinate system. This is possible as the epipolar geometry is entirely defined by the relative positions of the two stereo cameras and can be recovered from the calibration process even without knowledge of the 3D geometry (so called uncalibrated stereo).

PMF exploits the epipolar constraint to reduce the search process, not by rectifying the image data but by rectification of the edgel locations.

#### 3.2 Disparity Gradient Limit

The disparity gradient limit [3] (DG) originated from the psychophysical observation on human stereo vision that objects are less likely to be successfully stereo fused when there are nearby objects at different depths from the observer. The DG between a pair of stereo points is estimated in the cyclopean camera space, and is given by the ratio of the difference in the two disparities to the separation in the cyclopean camera image. Burt [3] demonstrated how the DG for the human vision system has a critical value of 1, above which stereo fusion fails. The PMF algorithm rejects disparity estimates that fail the disparity gradient limit of 1 within a 7 pixel radius. It is interesting to note that this limit implicitly embodies the ordering constraint which ensures that the order of matched points in the left images is preserved in the right.

#### 3.3 Uniqueness

The uniqueness constraint arises from the assumption that any point on a physical surface has a unique position in space at any one time. A relaxation algorithm [16] is employed to enforce the uniqueness constraint. First any matches with the highest match strength for both of the image primitives that formed them are chosen as correct (i.e. mutual preference). Then other matches associated with these points are eliminated from further consideration. This allows further matches, not previously either accepted or rejected to be selected as correct because they now have the highest match strengths for both constituent features.

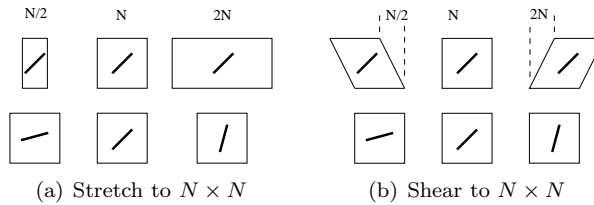


Figure 2: The warping process applied to single edge

## 4 Stretch Correlation

The stretch correlation algorithm [10] is an area based reformulation of the PMF algorithm, matching discrete blocks in the left image to blocks in the right. As in PMF the matching process is focused on information rich areas of the image (those containing non-horizontal edges). However, unlike PMF which uses complex edge string based matching, the stretch correlator uses a dot product correlation to generate match hypothesis for small blocks containing edge enhanced pixel data. This virtually eliminates the photometric dependence that is present in many correlation matching algorithms.

Most area based correlation techniques assume front-o-parallel surfaces. However, such an assumption compromises the correlation process by failing to account for surface rotations caused by the changing views. The stretch correlation technique models these rotations by warping the blocks, stretching or shearing the right image blocks, which improves the accuracy of the correlation process. This in turn improves the accuracy of the resulting disparity estimates.

The stretching process takes an  $N \times M$  pixel block and performs first order interpolation in the horizontal direction using the two pixels either side of the desired pixel to produce an  $N \times N$  block. This process is demonstrated in Fig. 2(a), where three blocks representing three different stretch factors are setup and interpolated. Typically,  $M$  might vary between  $3N$  and  $\frac{N}{3}$ , which is consistent with enforcing a cyclopean disparity gradient limit of 1 along the epipolar. Within this framework there is also the potential for shear-correlation. This takes a sheared rectangle from the source image and, using interpolation as for stretch correlation, produces an  $N \times N$  block, as in Fig. 2(b). In practice the shear-correlation technique has not proved necessary.

### 4.1 Algorithm

In order to drive the correlation matching process with edge information rather than grey level pixel data, the images are first pre-processed using an edge enhancement filter. This filter is a Gaussian smoothed horizontal differencing filter that suppresses high frequency image noise and enhances non-horizontal (non epipolar aligned) edges.

To exploit the epipolar constraint image rectification is performed on the left and right images, aligning the image rasters with the epipolars. Rectification requires sub-pixel interpolation with sufficient accuracy to preserve the first and second order image derivatives required by subsequent processing stages of the algorithm, which can be achieved using quadratic or sinc interpolation functions [20].

The algorithm then uses a left-to-right block matching scheme where the left image is divided into  $N \times N$  even sized blocks (typically  $6 \times 6$  or  $8 \times 8$ ). Left image blocks containing one or more edgels are correlated with raster equivalent regions in the right image over a range of stretches/compressions. For a given block in the left image this gives rise to a 2D space of correlation scores with disparity as one axis and stretch/compression factor as the other. A dot product correlation metric is used to compute the similarity at each point in the disparity-warp space and potential block matches are identified as peaks in this disparity-warp correlation search space. A disparity gradient constraint is first applied at the block level, ensuring consistency amongst neighbouring blocks and resolving any ambiguous matches. Using the disparity-warp parameters the algorithm then identifies the corresponding edgel location in the right image, within tolerance. The locations of the left and right edgels are then used to reference the corresponding edge elements detected using Canny [4], which has a location accuracy of approximately 0.1 pixels ( $\Delta_{xf}$  in Appendix A). Another local disparity gradient filter is applied this time at the pixel level to reject any points breaking the DG limit (potential mismatches are still possible for instance in blocks containing more than one edge). The sub-pixel locations of the remaining matched edgels are then used to calculate depth. Thus the stretch correlation algorithm provides accurate disparity estimates gained from matched Canny edges, while using a computationally efficient, hardware realisable correlation based framework.

Algorithm	Additions	Multiplications
Fixed correlation	$57 \times 10^6$	$57 \times 10^6$
Stretch correlation	$331 \times 10^6$	$334 \times 10^6$
Fast stretch correlation	$81.5 \times 10^6$	$117 \times 10^6$

Table 1: Comparison of computational load between the original stretch correlation and the fast version. The fixed correlation algorithm provides the lowest possible bound on the correlation approach.

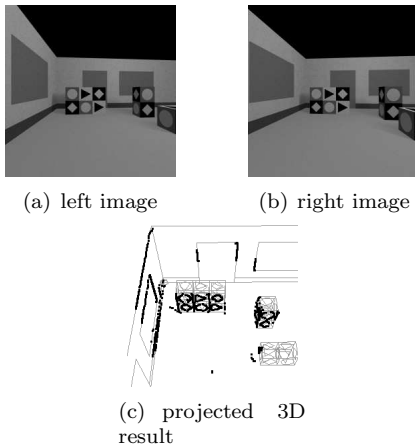


Figure 3: Artificially rendered cube-room scene images with 3D result

## 4.2 Fast Stretch Correlation

If the process of window shaping (stretch/shear) were to be incorporated directly into a correlation algorithm it would greatly increase the computational requirements of the algorithm by a factor proportional to the quantisation of the warp parameter. However, in [9, 11] we presented a formulation of the SC algorithm called the Fast Stretch Correlation (FSC) algorithm, which exploits the inherent redundancies in the stretch process to massively reduce the computational overhead. This was achieved by reformulating the dot product correlation between the blocks using a series of 1D multiply-accumulate operations between columns of pixels from the left block and across the disparity search range in the right image. Whilst the algorithm must calculate a new set of cross correlation 1D column summations for each block used from the left image, the right normalisation 1D column summations only have to be calculated once per row of blocks creating additional load savings. A comparison of the computational loads in terms of the numbers of multiplies and additions required for a typical  $512 \times 512$  stereo scene is given in Table 1. It clearly shows the benefit of FSC achieving the same results as the standard SC algorithm but requiring between a third and a quarter of the multiply-additions. As another computational saving, the pre-processing stages of Gaussian smoothing, horizontal differencing and image rectification are all linear filtering algorithms and can therefore be combined into a single convolution process.

## 4.3 Performance

Robustness and accuracy are the important criteria when considering the performance of depth estimation techniques. Accuracy is independent of the correspondence solving algorithm. As was shown using error propagation in [6] for a fixed set of camera parameters, it is the accuracy with which the features can be detected that dictates the depth estimate accuracy. The equations for the accuracy of world co-ordinates are given in Appendix A. These equations have been validated using simulated data with ground truth. The number of matches/mis-matches dictates the robustness of any correspondence algorithm. In the examples that follow the FSC algorithm was used to estimate disparities for a number of image pairs. Using structural knowledge (available for artificial data such as in Fig. 3) it was found that, typically the FSC could be expected to mis-match between 1%-2% of the data and in the worst case up to 8%. Figure 4 shows the performance of the technique on real image sets.

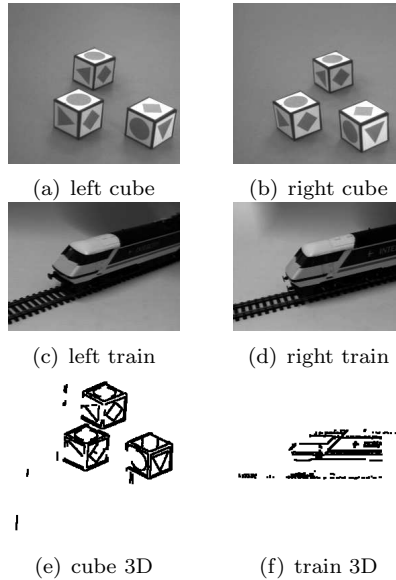


Figure 4: Real cube and train scenes images with 3D results

## 5 Temporal-Stereo SC

The quality of any stereo algorithm depends directly on its ability to solve the correspondence problem with the minimum number of mismatches. Therefore, any technique that can minimise the mismatch probability must improve the robustness of the algorithm. We have shown in [18] that for any stereo matching algorithm, the mismatch probability can be minimised by reducing the correspondence search area. However, in the absence of information regarding the range of expected disparities, a suitable minimum search area cannot be established. One way to acquire this information is to encapsulate the matching algorithm within a temporal loop where disparities from the previous frame are used to seed matching in the current frame. In this way, the search ranges can be kept as small as possible, given the expected degree of movement in the scene, reducing the probability of a mismatch. This process also improves the computational load of the algorithm, directly reducing the number of block correlations required.

### 5.1 Temporal Bootstrapping

Our first attempt to introduce temporal consistency into the stretch correlation involved a process of temporal bootstrapping. In the original SC/FSC algorithms block matching was performed along the complete epipolar (raster in the rectified image). In temporal stretch correlation (TSC) the size and location of the epipolar search bands is biased, constraining feature matching in the current stereo pair, using the disparity information from previous frames. This is done by centring the matching around the location of the previous best match with a disparity range proportional to the previous local disparities. Bootstrapping the matching algorithm in this way seeds the matcher with the most probable hypotheses for each match location given prior evidence. However, because full stereo constraints are re-applied at each stage, there is still the possibility of identifying and resolving mismatches from previous results in the light of new stereo information.

### 5.2 Comparison of FSC and TSC Algorithms

The cube images used in the previous demonstration actually form part of a sequence of stereo images where the cubes move across the field of view, parallel to the camera system. This sequence is used here to demonstrate the performance of the TSC algorithm in comparison to the original FSC technique. Figure 5 shows the 3D projection of the extracted depth for the first, middle and last frames in the sequence for both the FSC and TSC algorithms.

In general the FSC algorithm matches the majority of the edgel features correctly. However, there are many instances throughout the sequence where outliers are generated and also where sections of previously correctly matched edges disappear. For the TSC algorithm, the initial 3D result is poorly matched for the front and rear cubes in the scene as both cubes are located outside the default disparity search range. However, the motion of the

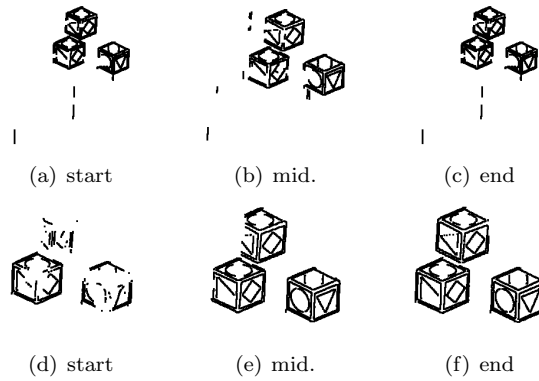


Figure 5: Real cube sequence 3D results for FSC (top) and TSC (bottom) algorithms

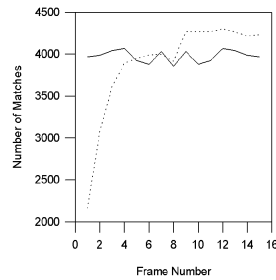


Figure 6: Match rates for the FSC (solid line) and TSC (dashed line)

cubes combined with the propagation of disparities from correctly matched regions allows the whole scene to be eventually captured without any obvious gross mismatches. Figure 6 shows the number of matches generated by each of the algorithms. We have found [5] that the TSC algorithm manages to produce 3D results that are more robust and contain more data than the FSC algorithm, whilst consistently using 50% or less of the computation load of the FSC algorithm. Using an artificial temporal-stereo image sequence like that shown in Fig. 3 it was possible to show that the TSC algorithm computed an average of 14% more matches with less than 1% typical outliers. The TSC algorithm avoids the need to recover motion correspondences and therefore does not impose any rigid body constraints on the scene. These results conclusively prove the value of including temporal constraints in scene reconstruction.

### 5.3 Multi-scale Temporal Stretch Correlation

There is a potential problem with the TSC algorithm. The extra robustness can only be delivered when the TSC algorithm has a good prior representation of the scene with which to bootstrap correspondence processes in subsequent frames. This often results in poor matching performance from the TSC algorithm at start-up, when no prior disparity information is available. This situation is identified as being particularly critical for applications such as autonomous navigation, where it is considered important that a robust 3D result should be available from the very first frame of the sequence.

One approach to solving this problem is to incorporate a coarse-to-fine (CTF) bootstrap at the start of the temporal analysis. Coarsening stereo images using scaling reduces the false maxima due to fine detail, leaving the larger, less ambiguous scene structure intact for matching. CTF bootstrapping involves tackling the correspondence problem hierarchically using a series of matching stages, starting with the coarsest scene structure and refining disparity measurements up to the 100% image scale. Unfortunately, problems can arise when new feature data appears from outside the FOV, or from occlusion, after the start of the sequence, when the CTF bootstrapping phase has finished. Under the right conditions (the new data appears close to existing data) the TSC-CTF algorithm would be able to ‘capture’ the new data. However, we regard this as unsatisfactory and instead have developed a multi-scale solution.

The multi-scale temporal stretch correlation [5] (MS-TSC) algorithm can match individual regions of stereo images using different correlation block scales and disparity search ranges, based on the past matching performance of



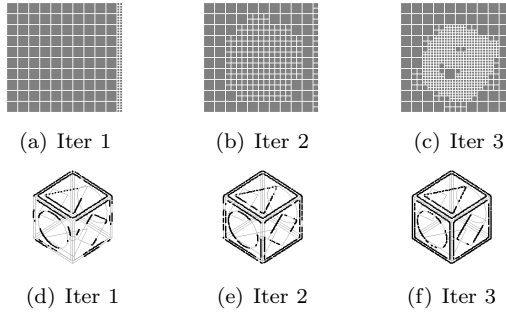


Figure 7: Multi-scale matching of static scene (3 iterations). Block scales across image are shown top. 3D results are shown below. The final number of edgels matched was 1043 with 1 remaining outlier.

individual blocks. Matching of new or previously unmatched feature blocks is attempted using large search ranges to increase the chances of encountering the correct correspondences, with robustness being maximised using coarse correlation blocks in the regions where large search ranges are employed. However, the rest of the correlation blocks in the scene can continue to be matched at the finest image scale as normal, utilising the superior accuracy and robustness of the basic TSC technique. Figure 7 illustrates the evolution of the correlation blocks as the MS-TSC algorithm iterates over a single frame-pair from synthetic sequence. We have found that 3 scales of coarse-to-fine are sufficient, 25%, 50% and 100% image scale.

#### 5.4 Mutual Consistency (Uniqueness)

Identifying non-unique matches is a simple task. In the case of a one-way matching algorithm, e.g. left-to-right point matching only, non-unique matches occur when two points map to a common feature in the other image. In this situation, it may be possible to resolve the correct match based on constraints such as surface smoothness, but if this fails, the algorithm will have to discard both matches despite the fact that one is probably correct. The need for a slightly more sophisticated approach to the uniqueness constraint leads to two-way matching algorithms that employ matching in both the left-to-right and right-to-left directions. The extra matching information makes possible a left-to-right-to-left or mutual consistency check that encompasses the uniqueness constraint. Although two-way matching leads to an inevitable increase in computational load, the extra information can help resolve not only non-unique matching ambiguities but can also identify problems such as left point A matching right point B, but B matching left point C. In this case, the lack of mutual consistency is a good indication that those points are not sufficiently unambiguous for robust stereo matching and should therefore be dropped from any further disparity calculation.

The mutual consistency constraint is implemented in the MS-TSC algorithm by executing parallel left-to-right and right-to-left correspondence solving processes that generate two sets of feature matches. These sets are then merged by checking each edgel match pair to see if there is a left-to-right-to-left match consistency. Only then is the local disparity gradient constraint applied to the merged matches.

#### 5.5 Demonstration

In these demonstrations depth was extracted on the train sequence 4 using the TSC and MS-TSC algorithms. The results for the first, middle and final frames are given in Fig. 8. As is shown the MS-TSC algorithm performs better than the TSC technique. The multi-scale search enables the technique to find a larger initial quantity of matches and to retain and refine them as the sequence moves on. Using the artificial blocks sequence we have been able to assess that the MS-TSC has a typical outlier score of less than 0.25% and a worst case of 5%.

## 6 Demonstration of CLV 3DMM

The 3DMM was outlined at the start of this paper. The original 3D model matcher uses only the 3D extracted features leading to errors due to; a) inaccuracies or failings in segmentation that generate incorrect feature geometry, b) incorrect feature matches and c) simplistic treatment of feature primitive statistics. The new closed loop validation stage (CLV) closes the loop on the process, testing the generated hypothesis against the original image data and refining this estimate without the constraints imposed by the previous algorithmic stages, Fig. 1.

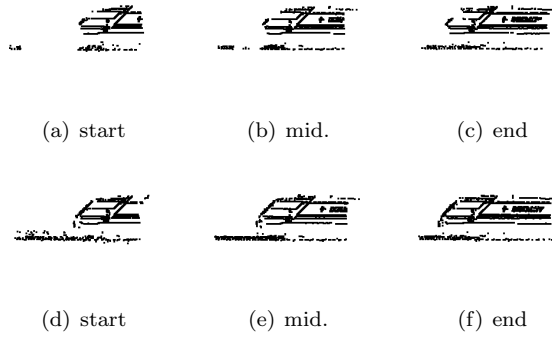


Figure 8: Train sequence 3D data for TSC (top) and MS-TSC (bottom) reprojected as a side view

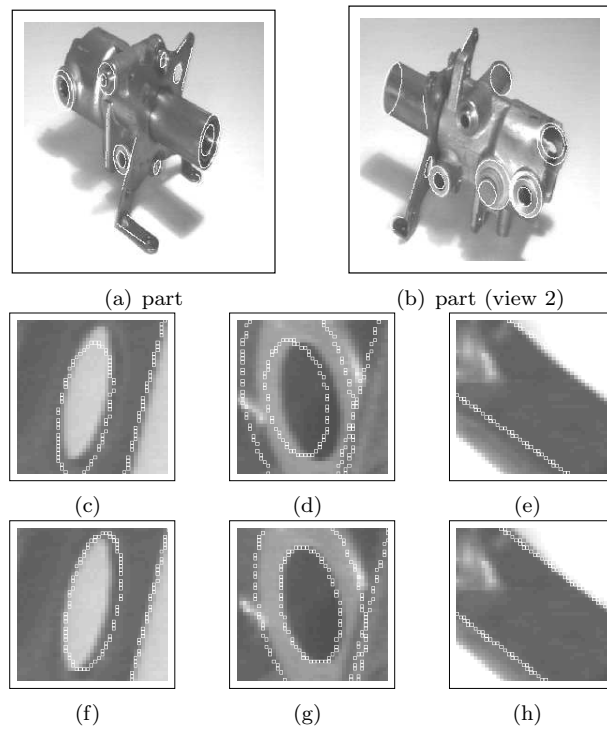


Figure 9: Full view and enlargements of an industrial part model fitted to data both with and without the use of the CLV. The enlargements show the effects of feature duality (c & f), specularity (d & g) and shadow (e & h)

The images in Fig. 9 show where the 3DMM system has been employed to locate a model of the scene using stereo data. In many cases it is obvious that the projected features could never locate onto edge data. This occurs when the feature is self occluded or view direction and illumination conspire to prevent the detection of a definite edge, i.e. the data is not present. Non-the-less the combined requirement of each feature having to be maximally consistent with the projected model seems to be enough for the consistent model features to be selected in preference to those that are an artefact of view point or illumination. The most significant change in the camera parameters adjusted by the CLV was a change of 5% in the aspect ratio of the cameras (a parameter that is often particularly difficult to satisfy due to the nature of the hardware). Agreement between the model and the scene cannot be achieved without allowing this change.

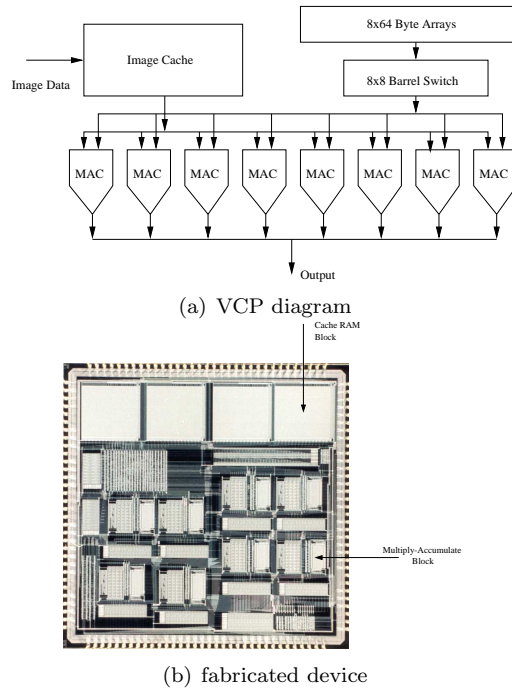


Figure 10: The video convolution processor architecture (top) and finished device (bottom).

## 7 Hardware

### 7.1 A Note on Camera Hardware

Sophisticated stereo algorithms such as these require the support of capable hardware. Our current system employs two Hitachi KFP1E cameras coupled to a Leutron PicPort Stereo PCI based framegrabber. This system, whilst adequate, represents the minimum acceptable for any accurate stereo image analysis. The cameras are able to run in non-interlaced mode and have a full frame shutter<sup>1</sup>, both of which are essential for any moving object analysis. The CCD array has square pixels and a reasonable sensitivity providing 8 bits per pixel with an estimated 1-2 bits of noise (again this really is a minimum specification and we would prefer 12 bits or more. The grabbing hardware is able to synchronise capture from both cameras simultaneously, however this is limited to exporting the horizontal and vertical drive signals. Ideally the framegrabber should be able to drive the internal clocks used by the cameras and although the PicPort Stereo can do this, the Hitachi cameras are unable to accept external pixel clocks. Pixel clocking would stabilise jitter noise to less than  $\pm 1$  pixel. The PicPort has two A/D converters and so can digitise the two channels of data simultaneously at frame rates (25fps). Such performance is unachievable with cheap colour cameras (which suffer from poor pixel alignment) and lower quality capture hardware.

### 7.2 The Video Convolution Processor

The FSC stereo algorithm was developed specifically with real-time performance in mind. Working towards that end, a parallel processing ASIC called the video convolution processor (VCP) chip was developed and fabricated and was presented in [9, 11]. The main VCP chip architecture is shown in Fig. 10, comprising; 8 parallel multiply-accumulators and 1 output accumulator, on-chip storage for 64, 8x8 pixel convolution kernels, and an input image data cache. Unlike conventional DSP chips, which only support fast raster order pixel processing, the VCP pixel cache is designed to support arbitrary movements of the correlation window within an input image, without having to stall the processor for long periods of time as new image data is retrieved from external memory. This data bottleneck is likely to be the largest problem in making efficient use of non-specific systems.

The VCP version of the FSC algorithm was implemented with a fixed sub-pixel accuracy of 1/8 of a pixel. By fixing the accuracy of the interpolation process, image rectification can be reduced to convolutions with fixed sub-pixel location kernels that can be stored on the VCP. All of the FSC pre-processing stages can be combined into a single

<sup>1</sup>the term ‘progressive scan’ is sometimes used to imply full frame shuttering but does not necessarily mean odd/even fields are captured simultaneously.

set of sub-pixel located convolution kernels. Therefore, all of the pre-processing stages can be accomplished in a single pass through the VCP chip. The VCP chip can also operate in a 1D convolution mode to allow the stretch correlation stage of the FSC algorithm to run optimally on the same device. The device was fabricated to run at 20MHz. However, even at this relatively low speed it was still possible to compute disparity for a  $256 \times 256$  image at a rate of between 2-5Hz. The architecture was designed to be scalable and thus, given the more aggressive fabrication technologies available to industry, it should be possible to extend the device to work at frame rates.

## 8 Discussion & Conclusions

This paper summarises a large body of work that has sought to incrementally refine a stereo vision system for a robotic vision application. The specific requirement of the use of 3D measurements in a practical working system has led us to arrive at conclusions regarding the useful characteristics of such algorithms that do not necessarily intersect with naive expectation. In particular we do not attempt to recover dense stereo data and we believe that knowledge of errors on derived results is essential for future use of the data.

We maintain that accurate determination of dense 3D data can only be done in the context of knowledge of the scene contents. A broader vision system also raises the possibility of obtaining truly robust dense stereo constrained by use of definitive top down knowledge. Although our own TINA vision system has many of the required capabilities to solve the dense stereo task, with limited abilities to recognise and locate familiar objects [8], we still have to accept that a robust dense stereo algorithm is a long way off. Paradoxically, by the time we get to the later stages of scene analysis, such as object recognition and location, the sparse data has already provided all of the quantitative information we needed from stereo, so a robotic vision system has little requirement for dense stereo data anyway.

We have found that once an FBS approach has been adopted, the spatial accuracy of the algorithm is largely fixed and can be treated independently from the stereo correspondence algorithm. The largest remaining practical problem is thus the reduction of outliers resulting from poor correspondence matching. We have therefore focused our efforts on understanding the statistical limits of correspondence algorithms and how best to make use of the inclusion of temporal constraints. We have shown that reformulation of standard feature based approaches for hardware implementation can bring potential rewards in this respect, allowing stereo algorithms that benefit from temporal consistency to reduce computational load and increase robustness. At all stages the data generated from our algorithms has been used for the practical purpose of 3D object location for robot manipulation. Such use brings the wider benefits of supporting self calibration by making use of the knowledge of the objects located to maintain a consistent Euclidean geometry.

One of the fundamental problems in the development of stereo vision algorithms is performance evaluation, and in particular the determination of ground truth. For the work presented here we have avoided this problem by the use of simulated data or particularly simple scenes. However, the closed loop validation stages of our vision system provide output that is by definition the expected location of the salient 3D information from the known (or required) features. Therefore, although we have not yet made use of this property in our own work, it would be totally appropriate to evaluate the result of alternative stereo correspondence algorithms using the results from the CLV, thus allowing the automatic determination of salient ground truth on realistic industrial scenes. This is an area of work that we would be interested in exploring together with others present at this workshop. All the algorithms described in this paper as well as many others are available for free download in the TINA vision system from <http://www.tina-vision.net>.

## A Stereo Error Propagation Results

We state the disparity errors derived for feature based stereo for each of the world direction vectors  $X$ ,  $Y$  and  $Z$ . The propagation of disparity errors was previously described by ourselves in [6] and similarly only in  $Z$  by Krotkov [7]. The disparity errors for edge-based stereo have been shown [6] to be

$$\Delta d = \sqrt{2\Delta x_f^2 + \frac{2\Delta e^2}{\tan^2 \theta}} \quad (1)$$

where  $\Delta x_f$  is the error due to edge location,  $\Delta e$  is the error in epipolar placement and  $\theta$  is the acute angle between the physical edge orientation and the epipolar.  $\Delta d$  is small when attributed to feature location errors, so we can assume that  $2hf \gg Z\Delta d$ , where  $h$  is the interocular separation of the optical centres of the cameras and  $f$  is the shared focal length. Hence

$$\Delta Z = \frac{Z^2 \Delta d}{2hf} \quad (2)$$

Assuming that  $\Delta x_l = \Delta x_r = \Delta x$  the sensitivity of  $X$  with errors in  $x_l$  (left image  $x$  location) and  $x_r$  (right image  $x$  location) is then given by;

$$\Delta X = \frac{Z \Delta x}{\sqrt{2}f} \left( \frac{X^2}{h^2} + 1 \right)^{\frac{1}{2}} \quad (3)$$

Similar analysis for  $\Delta Y$  gives the result

$$\Delta Y = \frac{Z}{\sqrt{2}f} \left( \frac{Y^2}{h^2} \Delta x^2 + \Delta y^2 \right)^{\frac{1}{2}} \quad (4)$$

## References

- [1] A Zisserman P Beardsley and I Reid. Metric calibration of a stereo rig. In *IEEE Workshop on Representation of Visual Scenes*, pages 93–100, Boston, 1995.
- [2] M Brown, T Drummond, and R Cipolla. 3D model acquisition by tracking 2D wireframes. In *Proc. of BMVC*, volume 2, pages 656–665, 2000.
- [3] P Burt and B Julesz. Modifications of the classical notion of panum’s fusional area. *Perception*, 9:671–682, 1980.
- [4] J F Canny. A computational approach to edge detection. *IEEE PAMI*, 8(6):679–741, 1986.
- [5] S Crossley. *Robust Temporal Stereo Computer Vision*. PhD thesis, Electrical and Electronic Engineering, University of Sheffield, 2000.
- [6] A J Harris, N A Thacker, and A J Lacey. Modelling feature based stereo vision for range sensor simulation. In *Proc. of the European Simulation Multiconference*, pages 417–421, 1998.
- [7] E P Krotkov. *Active computer vision by cooperative focus and stereo*. Springer series in perception engineering. Springer-Verlag, New York, 1989.
- [8] A J Lacey, N A Thacker, P Courtney, and S B Pollard. TINA 2001: The closed loop 3D model matcher. In *Proc. of BMVC*, volume 1, pages 203–212, 2001.
- [9] R A Lane. *Edge Based Stereo Vision with a VLSI Implementation*. PhD thesis, University of Sheffield, 1995.
- [10] R A Lane, N A Thacker, and N L Seed. Stretch correlation as a real-time alternative to feature based stereo matching algorithms. *Image and Vision Computing*, 12:203–212, 1994.
- [11] R A Lane, N A Thacker, N L Seed, and P A Ivey. A generalised computer vision chip. *Real Time Imaging*, 2:203–213, 1994.
- [12] G P Otto and T K W Chau. A region growing algorithm for matching of terrian images. In *Proc. 4th Alvey Vision Conference*, pages 123–128, 1988.
- [13] S B Pollard. *Identifying Correspondences in Binocular Stereo*. PhD thesis, Psychology: AIVRU, University of Sheffield, 1985.
- [14] J Porill, S B Pollard, T Pridmore, J Bowen, J E W Mayhew, and J P Frisby. *TINA: A 3D Vision System for Pick and Place*. MIT Press, 1991.
- [15] J Porill, S B Pollard, T Pridmore, J Bowen J E W Mayhew, and J P Frisby. TINA: A 3D vision system for pick and place. In *Proc. 3rd Alvey Vision Conference*, pages 65–72, 1987.
- [16] A Rosenfield, R Hummel, and S W Zucker. Scene labelling by relaxation operations. *IEEE Trans. on Systems, Man and Cybernetics*, 6:420–433, 1976.
- [17] N A Thacker and P Courtney. Calibration of a 4 DOF stereo head. In *Proc. BMVC*, volume 2, pages 528–537, 1992.
- [18] N A Thacker and P Courtney. Statistical analysis of a stereo matching algorithm. In *Proc. BMVC*, pages 316–326, 1992.
- [19] N A Thacker and J E W Mayhew. Optimal combination of stereo camera calibration from arbitrary stereo images. *Image and Vision Computing*, 9(1):27–32, 1991.
- [20] G Wolberg. *Digital Image Warping*. IEEE Computer Society Press, 1990.