Reliable Feature Matching Across Widely Separated Views

Adam Baumberg
Canon Research Centre Europe Ltd
1 Occam Court, Surrey Research Park
Guildford, Surrey GU2 5YJ
United Kingdom
adamb@cre.canon.co.uk

Abstract

In this paper we present a robust method for automatically matching features in images corresponding to the same physical point on an object seen from two arbitrary viewpoints. Unlike conventional stereo matching approaches we assume no prior knowledge about the relative camera positions and orientations. In fact in our application this is the information we wish to determine from the image feature matches. Features are detected in two or more images and characterised using affine texture invariants. The problem of window effects is explicitly addressed by our method - our feature characterisation is invariant to linear transformations of the image data including rotation, stretch and skew. The feature matching process is optimised for a structure-from-motion application where we wish to ignore unreliable matches at the expense of reducing the number of feature matches.

1. Introduction

The problem being addressed is, given two arbitrary images of a scene or object, can we find a reliable set of feature matches. In our application we wish to use the feature matches to recover camera motion parameters (or camera intrinsics) using standard "Structure-From-Motion" techniques (see for example Beardsley et al [1]).

1.1. Current approaches to stereo correspondence

The stereo matching problem can be broken down into several categories:

Calibrated short baseline stereo: Two cameras can be placed in a known configuration such that the cameras are close together relative to the viewed scene (a "short" baseline). The conventional convergent stereo configuration makes the correspondence problem much easier because for any given point in the left image, the corresponding point in the right image lies on a known epipolar line. The drawback of this approach is obviously that the camera configuration is heavily constrained. There is a large body of literature on this subject (e.g. Falkenhagen [3], Faugeras [4]).

Uncalibrated short baseline stereo: A single camera can be moved in such a way that the displacement of the camera and change in camera orientation between images is small but unknown. For example a sequence taken with a video camera can be used. In such a case points in a given image frame appear to be displaced by a small amount in the next frame. This makes the correspondence problem easier because for any given point in a given image, the corresponding point in the next image in the sequence is known to lie within some small neighbourhood of the original location. The drawback of this approach is that again the camera configuration (or in the case of a video camera, the camera motion) is heavily constrained. There has been much work with short baseline images (e.g. Deriche [2], Xu [20]) as well as tracking features through video sequences (e.g. Tomasi and Shi [19]).

Uncalibrated wide baseline stereo: This is the problem that we are trying to address. The distance between cameras relative to the viewed scene is significant (a "wide" baseline). The epipolar geometry of the scene is unknown and the ultimate aim is to determine this using point correspondences. The camera(s) may also be cyclo-rotated significantly between images. Conventional techniques that use intensity cross-correlation as an affinity measure for potential matches will fail in these situations. Robust statistical methods such as RANSAC [5] can be used to tolerate a significant fraction of mismatched features. However there is

an associated computational cost that grows with the number of mismatched features. This approach alone will fail for large changes in camera parameters or large camera motions.

1.2. Current approaches to wide baseline stereo

Pritchett and Zisserman [15] describe their approach which is to generate sets of local planar homographies and to use these for two purposes. Firstly, to provide a better affinity measure between potentially matching features and secondly, to restrict the search for potential feature matches. The main drawback of their approach is that the generation of possible homographies relies on suitable structures (parallelograms and large planar regions) being present in the scene.

Gouet *et al* describe an approach based on image invariants [7]. Features are detected in colour images and characterised using first order differential rotation invariants. Matching is performed using a relaxation technique which uses semi-local constraints. The relaxation technique approach is computationally quite expensive.

There has also been some relevant work in the general area of image matching and indexing. Schmid and Mohr describe an approach to indexing greyscale intensity images using differential rotation invariants calculated at multiple scales [18]. A voting scheme is used to accumulate the number of features in a query image that match features in each image in a database. The system has not been demonstrated with wide viewpoint variation.

More recently, Lowe [13] describes a Scale Invariant Feature Transform (SIFT) approach where scale-space features are detected and characterised in a manner invariant to location, scale and orientation. Some robustness to small shifts in local geometry is achieved by representing the local image region with multiple images representing each of a number of orientation planes. This results in an unnecessarily high dimensional SIFT key vector which is not truly invariant to affine distortions.

A general approach to matching uncalibrated images is described by Deriche [2]. This paper is really concerned with the short baseline problem although the relaxation matching scheme described is also applicable to the wider problem. Correlation is used over a significantly sized search window to match features in two images. For wide baseline stereo this will fail because raw image correlation will be sensitive to the affine (and more generally projective) distortions of the image which can occur.

Our approach significantly extends the rotation invariants method described by Gouet to cope with local affine image transformations. Under an unconstrained change of viewpoint a small planar surface patch will undergo a (near) affine transformation in the images. Hence it is important to be able to cope with this class of image transformations.

Affine invariants have been used for recognition purposes (e.g. moment invariants are used by Kadyrov [10], Flusser and Suk [6] and photometric affine invariants by Van Gool [14]). The main problem with these methods is that they either require binary images or segmented bounded regions. These are difficult to obtain robustly for general images. Sato and Cipolla consider the statistics of image texture features over the whole image to recover affine transformations [16]. Again this method ignores window effects which will be discussed in more detail later on.

Finally it is worth mentioning another approach to matching features with affine distortion described by Gruen [8]. The ALSC (Adaptive Least Squares Correlation) technique attempts to match image regions by explicitly recovering the affine transformation between the two. The transformation may be different for every pair of features and thus the alogorithm needs to be run on every candidate pair. Hence this approach becomes prohibitively costly as the number of potential feature matches increases.

2. Outline of our approach

There are 3 basic steps to our system:

1. Detect scale-space features – This step extracts a set of "interest points" in each image, each with an associated scale. We use a multi-scale Harris feature detector (see section 3).

2. Calculate affine invariants – Each interest point is characterised by examining a region in the image centred around that point. We explicitly consider the problem of determining an appropriate window that allows the robust calculation of a set of "characterisation" values that are robust to local linear transformations of the image data (i.e. 2D rotations, stretching and skew). Typically 20-40 invariants are required to adequately characterise an interest point (see section 4).

3. Matching – By comparing two vectors of invariants the similarity between two interest points from different images can be efficiently determined. The matching step uses this information to determine a set of corresponding points from the interest points in two images. The mechanism aims to find a reliable set of matches so that the number of incorrect matches is small compared to the total (see section 5).

Once correspondences are found between a single pair of images the epipolar constraint can be employed to remove a small number of incorrect matches (outliers) in the conventional manner.

By matching across successive pairs of images, the cameras and scene geometry can be determined using a conventional "Structure From Motion" approach.

3. Feature Detection

Any suitable scale-space feature detector could be used in this step. Given two images of a scene we would like to allow for changes in scale due to camera translation or zoom. Ideally scale-space features should be used such that there is a good chance of a feature being detected in both the "zoomed-out" image and the "close-up" image. The scales associated with these features should reflect their differing apparent size in the images.

We have found that in practice true scale-space features such as scale-space maxima of a "cornerness" measure (see Lindeberg [11]) are not reliably detectable. In particular the scale associated with a corner feature is often unreliable because the feature is present over a range of scales (with no one single dominant scale). Hence we have found that a better approach is to detect spatial Harris features [9] at a set of scales and order these features based on a scale-normalised feature strength as follows.

For each image a fixed number of interest points are calculated. We compute the 2nd moment matrix M (at some scale σ), at every point in the image using :-

$$M = \exp -x^T x/2\sigma^2 \otimes ((\nabla I)(\nabla I)^T) \qquad (1)$$

where ∇I is the gradient operator on the intensity image I calculated at some scale t and $\otimes$ is the image convolution operator over $x \in \Re^2$. A Harris corner strength measure is calculated from the determinant and the trace of this matrix as follows:

$$\text{strength} = \det M - 0.04 * (\text{trace}(M))^2$$

Corners are placed at local maxima of the corner strength measure. The corner strength measure can then be used to order the corners in order of significance.

Lindeberg points out that there are two scales that can be varied in the calculation of M – the integration scale, σ and the "local scale" at which derivatives are calculated, t. The approach taken here is to fix the local scale t proportional to the integration scale σ. In our system the Harris detector is run at multiple integration scales using a geometric progression of fixed scale settings.

We use the determinant and trace of the scale normalised 2nd moment matrix (as defined by Lindeberg [11]) to calculate a scale normalised corner strength measure. This means that corner strengths can be compared across different scales and the top 'n' corners over all detected scales can be determined.

4. Calculating affine invariants

The motivation for choosing this class of transformations is that a small planar surface patch when viewed from a varying viewpoint undergoes an *affine* distortion. Smooth surfaces can locally be approximated by planar surface patches. Hence the approach is valid for the majority of everyday scenes which contain some smooth surface regions.

We intend to match scale-space features – features with a spatial 2D image location and an associated scale. If we hypothesise a match between two such features in two images then assuming an affine model there are 3 unknown transformation parameters for the local image distortion:

- stretch, skew – can be parameterised by two parameters (a direction angle and a stretch factor).
- rotation – can be parameterised by an angle

This information is illustrated graphically in figure 1. The translation (2 parameters) is determined by the spatial position of the features and the scale change (1 parameter) is determined by the relative scale of the features.

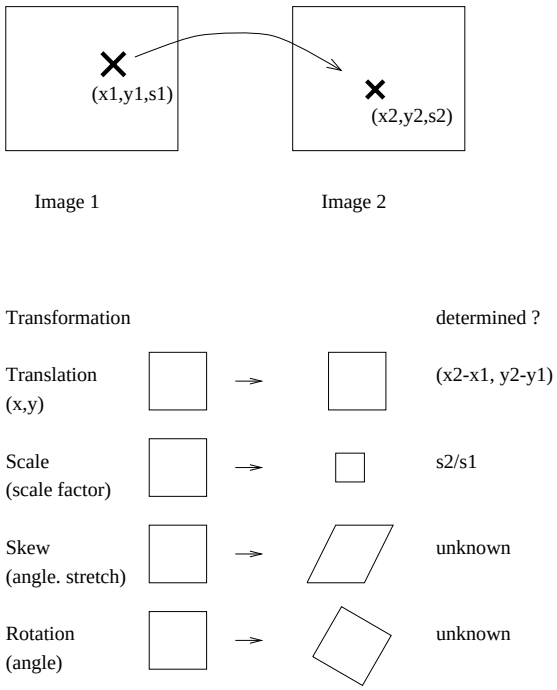

Figure 1. Hypothesised correspondence

4.1. Overview

The calculation of invariants can be broken down into three stages:

1. remove stretch and skew effects
2. normalise for photometric intensity changes
3. remove rotation effects

Previous approaches to calculating affine invariants ignore the issue of choosing a suitable window function. If we ignore this problem and used a circular Gaussian window function to calculate differential affine invariants then the resulting characterisation will not in fact be invariant to affine transformations.

For example, suppose a circular window centered around a given interest point is always used when calculating invariants. After an affine transformation the image structure in the circle is mapped to an elliptical region. If we place a circle around the transformed image feature that contains this elliptical region there will be additional image structures in the region that will distort any invariant measures calculated. Hence the window function needs to be adapted for each image feature.

4.2. Choosing a window function and removing stretch and sk ew

We propose a novel method for determining a stretch and skew normalised image patch for further processing. The key point is to adapt the shape of a window function based on local image data. The algorithm extends the idea of shape-adapted texture descriptors as described by Lindeberg [12]. Lindeberg extends the notion of scale space to "affine Gaussian scale-space". Consider the 2nd moment matrix defined by equation (1). In this equation a rotationally symmetric Gaussian window function is used to calculate the moment descriptor.

More generally using "affine Gaussian scale-space" elliptical window functions can be used with associated covariance matrices (or "shape matrices"). Using Lindeberg's notation we define the following second moment descriptor, μ_L by

$$\mu_L(.;\Sigma_t,\Sigma_s) = g(.;\Sigma_s) \otimes ((\nabla L)(.;\Sigma_t)(\nabla L)(,;\Sigma_t)^T)$$

where $L(.;\Sigma)$ is the affine Gaussian scale-space representation for an intensity image $I(.)$, Σ_t is a covariance matrix corresponding to the local scale and Σ_s is a covariance matrix corresponding to the integration scale.

Given a covariance matrix Σ, the associated non-uniform Gaussian kernel used to generate $L(.;\Sigma)$ is given by

$$g(x;\Sigma) = \frac{1}{2\pi\sqrt{\det\Sigma}} \exp(-x^T\Sigma^{-1}x/2)$$

Hence whereas conventional scale space is generated by convolution with a rotationally symmetric Gaussian, affine scale-space is generated by a linearly transformed (stretched, skewed etc) elliptical gaussian kernel.

Lindeberg describes an iterative procedure for adapting the shape matrices such that the following "fixed point" property holds for some matrix M_L.

$$\begin{aligned} \mu_L(q_L;\Sigma_{t,L},\Sigma_{s,L}) &= M_L \\ \Sigma_{t,L} &= tM_L^{-1} \\ \Sigma_{s,L} &= sM_L^{-1} \end{aligned}$$

Lindeberg shows that if moment image descriptors are calculated under these conditions then the image descriptors will be relative invariant under arbitrary affine transformations (see [12] for details).

We observe that the shape adaptation scheme of Lindeberg can be used to determine a stretch-skew normalised image patch as follows. Consider a 2D image $I_R(x)$ and a linearly transformed image $I_L(x) = I_R(Bx)$. Suppose the shape adapted second moment matrices are calculated for both images at q_L and $q_R = Bq_L$ respectively. For both images we can transform the image data to a normalised frame using the square root of the second moment matrix. We define the transformed image by

$$I_{L'}(M_L^{-\frac{1}{2}}x) = I_L(x)$$

where $M_L^{-\frac{1}{2}}$ is the square root matrix of M_L and similarly for $I_{R'}$. Note this square root matrix is well defined (up to a rotation) since the second moment matrix is symmetric positive definite. We use Cholesky decomposition to calculate this matrix.

Lindeberg derives the following transformation property for affine scale-space second moment matrices. Under a linear transformation of image coordinates B

$$\mu_L\left(q;\Sigma_t,\Sigma_s\right) = B^T\mu_R\left(Bq;B\Sigma_tB^T,B\Sigma_sB^T\right)B$$

Using this transformation property we can show that in our normalised frame:

$$\mu_{L'}(q';tI,sI) = I$$

where I is the 2×2 identity matrix.

Hence in this transformed domain the second moment matrix calculated using circular symmetric smoothing kernels is the identity for both images. Given we have linearly transformed copies of one image, there exists a linear transformation B' from I'_L to I'_R and applying the transformation properties of the second moment matrix again we can show that

$$I = \mu_{L'} = B'^T\mu_{R'}B' = B'^TB'$$

and hence B' is a rotation. The situation is illustrated in figure 2.

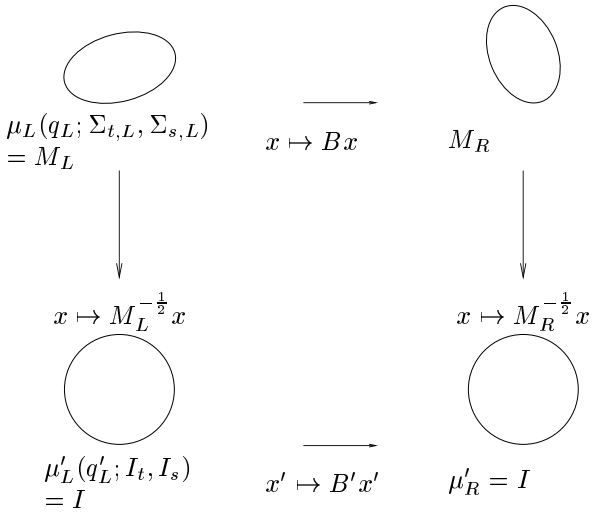

Figure 2. Diagram illustrating transformation of second moment matrices

We have shown that by calculating a shape adapted moment descriptor at a 2D interest point and transforming using the square root matrix of this descriptor we can obtain a normalised image patch. We have shown that any two such normalised patches originating from an image and a linearly distorted copy are related by a rotation.

In practice for this part of the algorithm, we use a simplified version of Lindeberg's shape adaptation scheme. The integration scale s is fixed proportional to the local scale t and this is set proportional to the detection scale of the interest point. For convenience, our iterative adaptation scheme works in the transformed image domain. We calculate the second moment matrix using the conventional rotationally symmetric smoothing kernels. We then transform the local image structure using the square root of this second moment matrix but scaled to have unit determinant. This process is repeated until convergence (i.e. until the second moment matrix is sufficiently close to the identity). Our adaptation scheme assumes the scale associated with our scale-space features is reasonably consistent across images and further scale adaptation is unnecessary. An obvious future step would be to implement the full scale adaptation approach of Lindeberg.

It is also worth noting that a simple linear rescaling of intensity values will rescale each element in the second moment matrix. Hence in order to be reasonably robust to lighting changes we have fixed the determinant of the second moment matrix to be 1 when determining the "skew-stretch normalisation" transformation.

4.3. Normalising for lighting changes

Once the image patch has been normalised for stretch and skew we use a conventional intensity (or colour) normalisation algorithm such as that described by Gouet et al [7].

4.4. Removing rotation

In order to obtain an affine invariant characterisation of the local image structure, all that remains is to remove the effects of rotation. We could use any conventional set of rotation invariants to generate our characterisation data. In practice we use a variant of the Fourier-Mellin transformation [17]. Explicitly, we calculate a set of complex-valued coefficients $u^X_{n,m}$ for each colour component X (i.e. red, green, blue or intensity) defined by:

$$u^X_{n,m} = \int \frac{d^n}{dr^n} G_\sigma(r) \exp(im\phi) J^X(r,\phi) r \, dr \, d\phi$$

where $J^X(r,\phi) = I^X(r\cos\phi + x_0, r\sin\phi + y_0)$ is the relevant colour component of the image patch, $\{r, \phi\}$ are polar coordinates defined about (x_0, y_0) the centre of the image patch (i.e. the location of the interest point) and $G_\sigma(r)$ is a 1D Gaussian window function with standard deviation σ set proportional to the size of the image patch. The integral can be efficiently approximated as a weighted sum of pixel values. The weights can be precomputed from the defining integral above.

Under a rotation of the image $J'(r,\phi) = J(r, \phi + \theta)$ these complex coefficients transform as follows:

$$u'_{n,m} = \exp(im\theta) u_{n,m} \tag{2}$$

Hence to calculate rotation invariants, we normalise all the coefficients $u^X_{n,k}$ by dividing by a unit-length complex number proportional to $u^X_{0,k}$. Note that for colour images the same normalisation coefficient may be used across all the colour components.

We calculate around 13 invariants for greyscale images and 43 invariants for RGB colour images.

5. Matching

The above features are sufficiently well characterised to use a very simple one-shot matching scheme. Each feature is represented by a set of affine invariants which are combined into a feature vector. A Mahalanobis distance metric is used to measure similarity between any two feature vectors.

Hence for two images we have a set of feature vectors $v^{(i)}$ and $w^{(j)}$ and a distance measure $d(v,w)$. The matching scheme proceeds as follows:

- Calculate distance matrix $m_{i,j}$ between pairs of features across the 2 images.

$$m_{i,j} = d(v^{(i)}, w^{(j)})$$

- Identify potential matches (i, j) such that feature i is the closest feature in the first image to feature j in the second image and vica versa.
- Score matches using an ambiguity measure.
- Select unambiguous matches (or best "n" matches).

We use an ambiguity measure similar to that used by Deriche [2]. The ambiguity score measures the relative distance between the two matching features and the next closest distance between one of the matched pair and any other feature.

We have found that for Structure-From-Motion algorithms the strength of a feature match is less important than the ambiguity of the match. Furthermore, we have realised that the ambiguity measure is the only information required for deciding whether to accept a possible match.

Note that at this stage we do not require any high level prior information. The reliability of the outlined scheme allows the matching and camera solving (fundamental matrix calculation) steps to be decoupled avoiding computationally expensive relaxation schemes.

6. Experimental results

6.1. Adapted windows for synthetic data

Figure 3 shows the adapted windows obtained around 4 corner features for a synthetic image of a square and the features at the same scale for an affine distorted copy of the image.

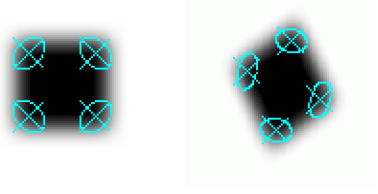

Figure 3. Shape of window functions around corner features

Figure 4 shows one of the features from each image mapped into the skew-normalised frame. It can be seen that these image patches are related by a rotation.

Figure 4. Resampled image patches in skew-normalised frame

6.2. Real intensity images

In this example two small images (approx 200 by 300 pixels) of a box viewed from different viewpoints were used. The images were obtained with a static camera and by rotating the object about 15 degrees. Hence the the light incident on each surface across the images varies (variable illumination).

The system takes about 1-2 minutes to process the pair of images on a standard 400Mhz Pentium PC. (The majority of processing time is spent in detection of interest points). Around 10 affine intensity invariants were used to characterise each feature and the strongest 200 features were detected in both images.

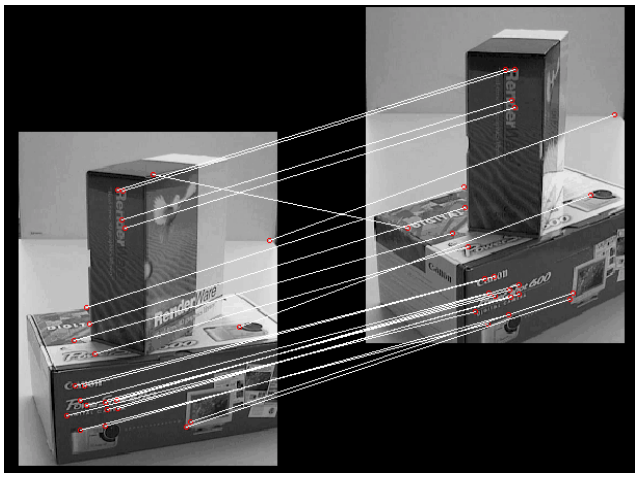

Figure 5. Example output with greyscale images

The correspondences found are indicated by the white lines joining the corresponding points in each image (figure 5). There were 24 correspondences found and there is only one mismatch. The epipolar constraint could be used to remove this outlier.

6.3. Real colour images

A sequence of full sized (768 by 576) colour images of a dinosoar toy were taken. The angular separation between images (i.e. the angle between camera1 - object - camera2) was typically 10 to 15 degrees. The algorithm takes about 2 minutes to process each frame on a Pentium PC (400 MHz). Around 40 colour affine invariants were used to characterise each feature and the strongest 200 features were detected in each image.

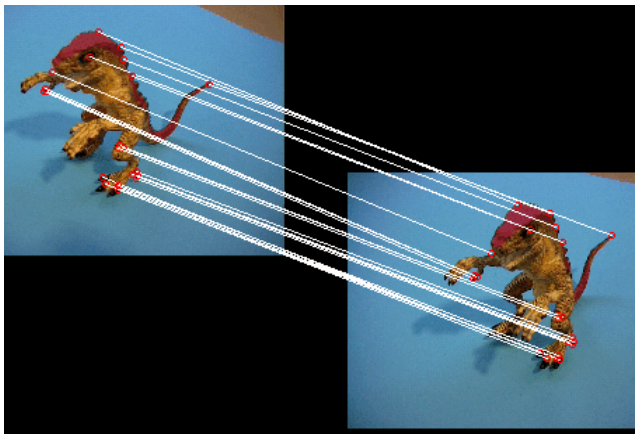

Figure 6. Matches obtained with images 4 and 5 for colour toy

Typically over 50 matches are obtained between successive image frames with few incorrect matches. A typical image pair is shown (figure 6) - 22 correspondences were found with no incorrect matches. For the purposes of visualisation the matches have been thinned (i.e. matches removed which are too close to neighbouring matches).

6.4. Matching across variable turn angle

A sequence of colour images of a toy house were taken. Five of the images are shown in figure 7 (labelled image 0, 1, 2, 3 and 4).

The matching algorithm was ran independently on pairs of images containing image 0 (i.e. (0,1), (0,2), (0,3), (0,4)). The turn angle between the cameras was also determined (using user-verified feature matches). For each image pair the following data was obtained:

- Number of potential matches - This is the total number of matches as defined in section 5 including matches with poor ambiguity scores.
- First incorrect match - The matches are ordered by ambiguity score into a list. This is the rank of the first incorrect match in the list.

Figure 7. Image sequence of a toy house

- Number of correct matches in top 20 - This is the number of verified matches in the list of 20 least ambiguous matches.

The results are summarised in figure 8.

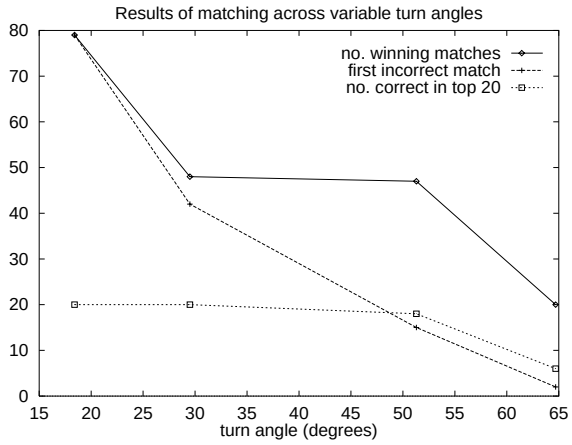

Figure 8. Graph showing performance of matcher for variable turn angle

The method appears to only break down for the final image pair where the turn angle is around 65 degrees. Figure 9 shows the best 20 matches between image 0 and image 3 (a turn angle of 51 degrees). Note the two mismatches where the corners of the window in image 0 are matched to the wrong window in image 3. However the first of these mismatches only occurs at position 15 in the ranked list of matches.

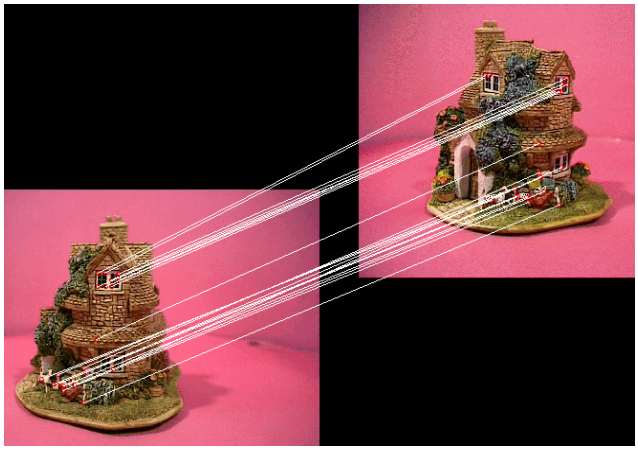

Figure 9. Best 20 matches between image 0 and 3

7. Summary and Conclusions

In this paper we describe a novel method for matching image features that is robust to local affine distortions. We have shown how an adaptive window function can be obtained that allows true affine-invariant descriptors of local image structure to be efficiently calculated. This characterisation is sufficiently descriptive to allow unconstrained image feature matching to be performed at low computational cost. The resulting feature matches can be used as reliable input to camera parameter computations (for example for 3D reconstruction problems).

We have also shown how reliable matches can be selected based on an ambiguity measure. This reduces the number of false matches obtained at the expense of total number of matches.

Future work will look at improving the computational efficiency of this approach and extending the shape adaptation to incorporate true scale adaptation.

References

[1] P. Beardsley, P. Torr, and A. Zisserman. 3d model acquisition from extended image sequences. *ECCV*, 2:683–695, 1996.

[2] R. Deriche, Z. Zhang, Q. Luong, and O. Faugeras. Robust recovery of the epipolar geometry for an uncalibrated stereo rig. In *ECCV94*, pages A:567–576, 1994.

[3] L. Falkenhagen. Depth estimation from stereoscopic image pairs assuming piecewise continuos surface. In Y. Paker and S. Wilbur, editors, *Image Processing for Broadcast and Video Production*, pages 115–127. Springer Great Britain, 1994.

[4] O. Faugeras. Three-dimensional computer vision. *MIT Press*, 1993.

[5] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Comm. Assoc. Comp. Mach*, 24(3):81–95, 1981.

[6] J. Flusser and T. Suk. Pattern recognition by affine moment invariants. *PR*, 26:167–174, 1993.

[7] V. Gouet, P. Montesinos, and D. Pel. A fast matching method for color uncalibrated images using differential invariants. In P. Lewis and M. Nixon, editors, *British Machine Vision Conference*, volume 1, pages 367–376. BMVA Press, 1998. ISBN 1-901725-04-9.

[8] A. W. Gruen. Adaptive least squares correlation: A powerful image matching technique. *Jornal of Photogrammetry, Remote Sensing and Cartography*, 14(3):175–87, 1985.

[9] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey88*, pages 147–152, 1988.

[10] A. Kadyrov. Triple features for linear distorted images. In *Computer Analysis of Images and Patterns (CAIP)*. International Association of Pattern Recognition, Springer Verlag, 1995.

[11] T. Lindeberg. Scale-space theory in computer vision. *Kluwer*, 1994.

[12] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *IVC*, 15(6):415–434, June 1997.

[13] D. Lowe. Object recognition from local scale-invariant features. In *ICCV99*, pages 1150–1157, 1999.

[14] E. Pauwels, T. Moons, L. VanGool, P. Kempenaers, and A. Oosterlinck. Recognition of planar shapes under affine distortion. *IJCV*, 14(1):49–65, January 1995.

[15] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV98*, pages 754–760, 1998.

[16] J. Sato and R. Cipolla. Extracting the affine transformation from texture moments. In *ECCV94*, pages B:165–172, 1994.

[17] R. Schalkoff. Digital image processing and computer vision. In *Wiley*, 1989.

[18] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19(5):530–535, May 1997.

[19] C. Tomasi and J. Shi. Good features to track. In *CVPR94*, pages 593–600, 1994.

[20] G. Xu. A unified approach to image matching and segmentation in stereo, motion, and object recognition via recovery of epipolar geometry. *Videre*, 1(1):22–55, 1997.