Invariant Features from Interest Point Groups

Matthew Brown and David Lowe

Department of Computer Science,
University of British Columbia,
Vancouver, Canada.

{mbrown|lowe}@cs.ubc.ca

Abstract. This paper approaches the problem of finding correspondences between images in which there are large changes in viewpoint, scale and illumination. Recent work has shown that scale-space 'interest points' may be found with good repeatability in spite of such changes. Furthermore, the high entropy of the surrounding image regions means that local descriptors are highly discriminative for matching. For descriptors at interest points to be robustly matched between images, they must be as far as possible invariant to the imaging process.
In this work we introduce a family of features which use groups of interest points to form geometrically invariant descriptors of image regions. Feature descriptors are formed by resampling the image relative to canonical frames defined by the points. In addition to robust matching, a key advantage of this approach is that each match implies a hypothesis of the local 2D (projective) transformation. This allows us to immediately reject most of the false matches using a Hough transform. We reject remaining outliers by selecting a set of correspondences which are consistent with the epipolar geometry. Results show that dense feature matching can be achieved in a few seconds of computation on 1GHz Pentium III machines.

Keywords: Image Features, Object Recognition, Correspondence

1 Introduction

A widely-used approach for finding corresponding points between images is to detect corners and match them using correlation, using the epipolar geometry as a consistency constraint [1, 2]. This sort of scheme works well for small motion, but will fail if there are large scale or viewpoint changes between the images. This is because the corner detectors used are not scale-invariant, and the correlation measures are not invariant to viewpoint, scale and illumination change. The first problem is addressed by scale-space theory, which has proposed feature detectors with automatic scale selection [3, 4]. In particular, scale-space interest point detectors have been shown to have much greater repeatability than their fixed scale equivalents [5, 6]. The second problem (inadequacy of correlation)

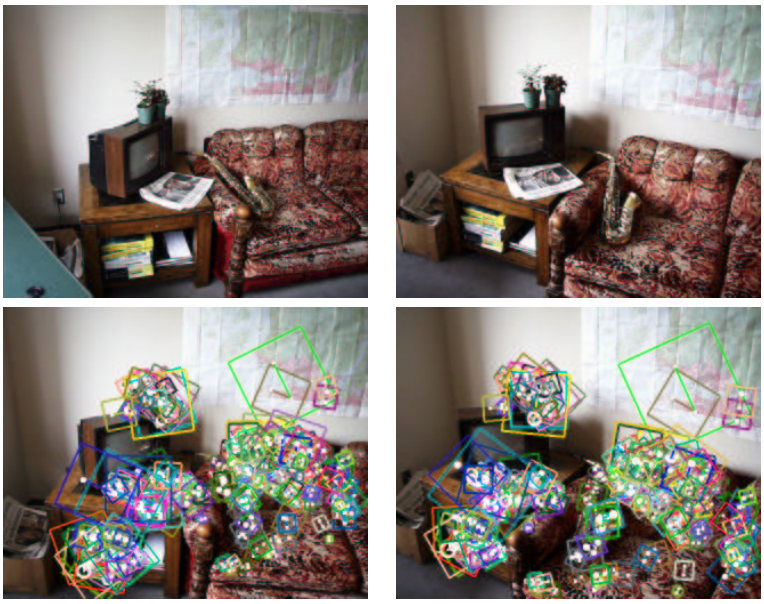


Fig. 1. Matching of invariant features between images with a large change in viewpoint. Each matched feature has been rendered in a different greyscale.

suggests the need for local descriptors of image regions that are invariant to the imaging process.

Describing regions of an image in a manner that is invariant to all possible viewing conditions is impossible, and assumptions must be made. Many authors assume simple geometric transformations between image regions, for example similarities or affinities. Schmid and Mohr [7] use rotationally symmetric Gaussian derivatives to characterise image regions. Lowe's SIFT features [5] use a characteristic scale and orientation at interest points to form similarity invariant descriptors. Baumberg [8] goes further by extracting six affine parameters from the image regions at scale-invariant interest points. The approach presented here is to use *groups of interest points* to define local 2D transformation parameters. Similar schemes have been demonstrated by [9, 10]. Unfortunately, both have suffered from the lack of scale-invariant interest point detectors in their implementation.

A commonality between these approaches is the assumption that the region local to an interest point is approximately planar in the world, so that such regions are by related by homographies between images. The main idea of our approach is that any such 2D transformation can be calibrated from groups of scale-invariant interest points. Groups of interest points which are nearest neighbours in scale-space are used to calibrate a 2D transformation (similarity, affinity or homography) to a canonical frame [11, 12]. The resampling of the image region in this canonical frame is geometrically invariant. Colour invariance

(invariance to changes in the illumination spectrum) is achieved by normalising intensity in each of the R, G, B channels [13].

In addition to enabling robust matching, a key advantage of the invariant feature approach is that each match represents a hypothesis of the local 2D transformation. This fact enables efficient rejection of outliers using geometric constraints. We use broad-bin Hough transform clustering [14, 15] to select matches that agree (within a large tolerance) upon a global similarity transform. This might seem to be a bad assumption for 3D scenes, but in practice we use a sufficiently large tolerance to accommodate underlying 3D structure, whilst still rejecting a large number of outliers. Other authors have used local constraints upon feature matches. For example, there is a consistency constraint between two homographies which are compatible with a fundamental matrix [2, 9]. Given a set of feature matches with relatively few outliers, we compute the fundamental matrix and use the epipolar constraint to reject remaining outliers.

In our implementation, we efficiently construct a scale-space representation by using an image pyramid. This contains the minimum number of samples needed to represent the image at each scale. High-dimensional feature vectors are efficiently matched using a k-d tree, as in [5].

2 Interest Points in Scale-Space

Our interest points are located at extrema of the Laplacian of the image in scale-space. This function is chosen for its response at points with 2-dimensional structure, and for the fact that it can be implemented very efficiently using a Laplacian Pyramid [16, 17]. In a Laplacian Pyramid, a difference of Gaussians is used to approximate the Laplacian.

Pyramid representations have the advantage that the minimum number of samples are used to represent the image at each scale, which greatly speeds up computation in comparison with a fixed resolution scheme. A scale-space sampling of the image and its Laplacian can be efficiently constructed by a series of convolution, subtraction and subsampling steps, as shown in figure 2. The subsampling and Gaussian standard deviation must be chosen such that each layer of the pyramid represents a 'correct sampling' of the image at some scale. In practice we use a Gaussian kernel with standard deviation 1.5 pixels and a sub-sampling of $1.5 : 1$ in the Laplacian Pyramid. The Gaussian kernel is discretised with 7 samples.

To find the maxima and minima of the scale-space Laplacian we first select samples which are extrema of their neighbours ± 1 sample spacing in each dimension. Then, we locate the extrema to sub-pixel / sub-scale accuracy by fitting a 3D quadratic to the scale-space Laplacian

$$L(\mathbf{x}) = L + \frac{\partial L}{\partial \mathbf{x}}^T \mathbf{x} + \frac{1}{2}\mathbf{x}^T \frac{\partial^2 L}{\partial \mathbf{x}^2}\mathbf{x}$$

where $\mathbf{x} = (x, y, s)^T$ is the scale-space coordinate and $L(\mathbf{x})$ is the approximation of the Laplacian. The quadratic coefficients are computed by approximating the

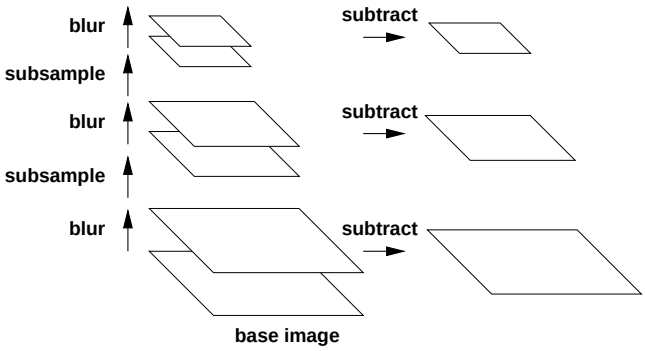


Fig. 2. The Laplacian Pyramid. Each level in the pyramid is a difference of Gaussian approximation to the Laplacian, formed by subtracting a Gaussian blurred version of the image from itself. The blurred image is then subsampled to generate the next pyramid level.

derivatives using pixel differences of the already smoothed neighbouring samples. The sub-pixel / sub-scale interest point location is taken as the extremum of this 3D quadratic

$$\hat{\mathbf{x}} = -\frac{\partial^2 L}{\partial \mathbf{x}^2}^{-1} \frac{\partial L}{\partial \mathbf{x}}$$

Locating interest points to sub-pixel / sub-scale accuracy in this way is especially important at higher levels in the pyramid. This is because the sample spacings at high levels in the pyramid correspond to large distances relative to the base image.

3 Invariant Features from Interest Point Groups

Once 'interesting points' in the image have been localised, robust matching requires an invariant description of the image region. Geometrical invariance is typically achieved by assuming that the region is locally planar, and attempting to recover the parameters of a 2D transformation. One approach is to use information from the local image region itself. For example, the 2^{nd} moment matrix can be used to recover affine deformation parameters [8]. Degeneracies can cause problems for this approach. For example, if the local image region were circularly symmetric, it would be impossible to extract a rotation parameter.

An alternative approach is to use groups of interest points to recover the 2D transformation parameters. There are a number of reasons for adopting this approach. Firstly, improvements in the repeatability of interest points mean that the probability of finding a group of repeated interest points is sufficiently large. Secondly, the transformation computation is guaranteed to be non-degenerate. Thirdly, and most importantly, since the interest points are very accurately localised, the 2D transformation estimate is also accurate.

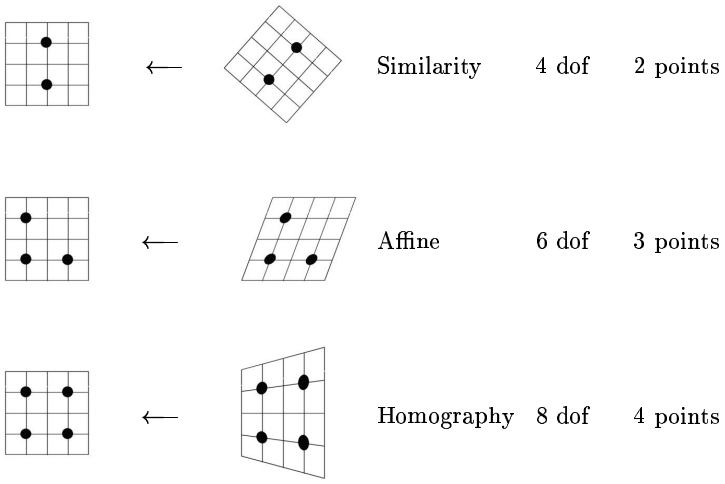


Fig. 3. 2D transformation invariant features based on interest point groups. Groups of interest points which are nearest neighbours are formed, and used to calibrate the 2D transformation to a canoncial frame. The feature descriptor is the resampling of the image in the canonical frame.

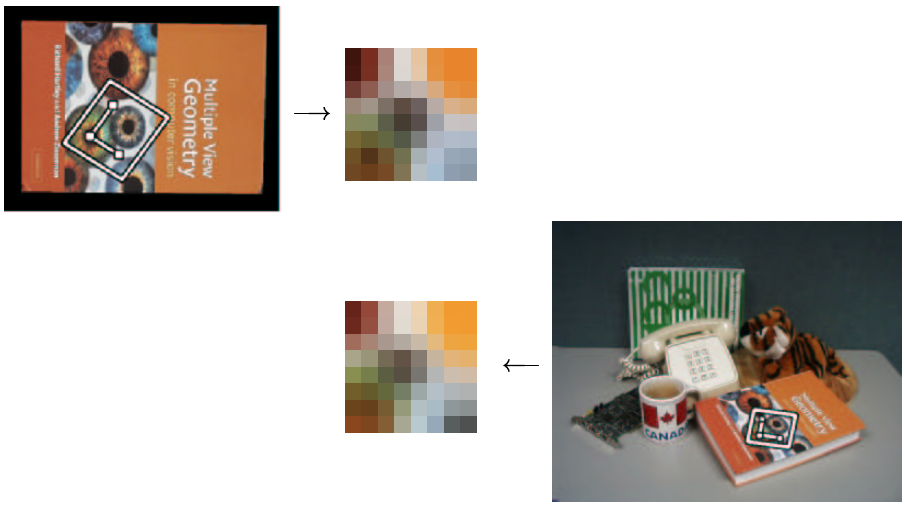


Fig. 4. Extraction of affine invariant features from a pair of images. Groups of 3 interest points are used to calibrate the affine transformation to a canonical frame. The image region is then resampled in the canonical frame to form the feature descriptor.

We propose a family of 2D transformation invariant features based on groups of interest points as follows. Find groups of $2 \leq n \leq 4$ interest points which are nearest neighbours in scale-space, and compute the $2 \times n$ parameter 2D transformation to a canonical frame. Describe the region local to the interest points by resampling it in the canonical frame. This is shown in figure 3. Aliasing is avoided by sampling from an appropriate level of the already constructed image pyramid. We resample using a linear interpolant.

In addition to geometric invariance, our features are also *colour invariant*. Results from the colour constancy literature, for example [13], have shown that, under practical illuminants[1], the diagonal model is appropriate for describing the change in R, G, B values. The diagonal model implies that there are independent scalings in each colour channel when the illumination spectrum changes. Hence, colour invariance can be obtained simply by normalising in each colour channel in the feature descriptors. This allows for spatial variation in the illumination conditions so long as the feature size is small compared to the scale on which the illumination change occurs. Note that geometrically related illumination effects such as shadows, reflections and specularities are not modelled.

Implementation Issues So far, only similarity invariant and affine invariant feature types have been implemented. When constructing affine features, interest point groups that are nearly degenerate (colinear or coincident) are rejected. This is because the canonical description would be oversampled and hence not discriminating.

3.1 Feature Matching

Features are efficiently matched using a k-d tree. A k-d tree is an axis-aligned binary space partition, which recursively partitions the feature space at the mean in the dimension with the highest variance. We use 8×8 pixels for the canonical description, each with 3 components corresponding to normalised R, G, B values. This results in 192 element feature vectors.

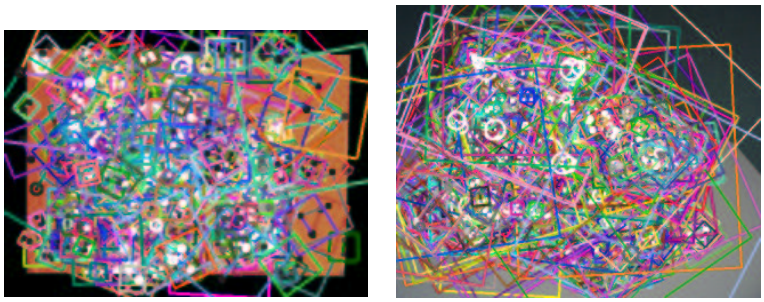
4 Finding Consistent Sets of Feature Matches

Our procedure for finding consistent sets of feature matches consists of two parts. First, we use a Hough transform to find a cluster of features in 2D transformation space. Next, we use RANSAC to refine the transformation before computing the epipolar geometry, rejecting additional outliers.

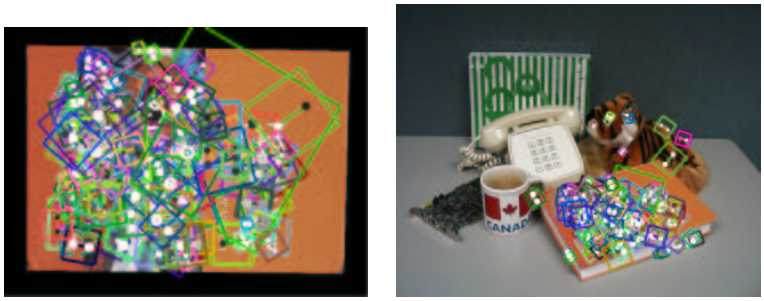
4.1 Hough Transform Clustering

An extremely useful property of the invariant-feature approach is that each match provides us with the parameters of a 2D transformation. This information can be used to efficiently reject outliers whose 2D transform estimates are

[1] For example fluorescent lighting, sunlight.



(a) All feature matches



(b) Hough transform clustering using similarity parameters

(c) RANSAC solution for the homography

Fig. 5. Finding consistent sets of feature matches. To find inliers to a homography we use a Hough transform followed by RANSAC. In this example, the test image was 640×512 pixels. The number of initial matches to consider was 5469, which was reduced to 279 by broad-bin Hough transform clustering, and further to 104 matches in the final solution using RANSAC.

inconsistent. We do this by finding a cluster of features in transformation space. This is known as a generalised Hough transform.

The transformation parameters for each feature match are entered into a broad-bin histogram. This can be implemented efficiently in practice by using a hash table for the histogram bins. To avoid boundary effects, votes are entered into two adjacent histogram bins in each dimension. For similarity transforms, the parameters we use are translations (t_1, t_2), log scale $(\log s)$ and rotation (θ). Typical bin sizes are $1/8$ of the image size for translation, one octave for scale, and $\pi/8$ radians for rotation.

Note that assuming a single cluster in 2D transformation space is equivalent to assuming a global homography. Since this is not necessarily the case, we must allow a large tolerance for inliers to account for the underlying 3D structure. This corresponds to using broad bins in the Hough transform histogram.

4.2 RANSAC Transformation Estimation

We refine the 2D transformation estimate using Random Sample Consensus (RANSAC). RANSAC has the advantage that it provides a homography estimate that is largely insensitive to outliers, but it will fail if the fraction of outliers is too great. This is why we use Hough transform clustering as a first step. See figure 5.

If the scene is 3-dimensional, we first select inliers which are loosely consistent with a 2D transformation using the above methods, using a large error tolerance. Then, given a set of points with relatively few outliers, we compute the fundamental matrix. This is used to find a final set of feature matches which is consistent with the epipolar geometry. See figure 6.

5 Results

We have found good results when using our invariant features to solve correspondence problems involving large viewpoint changes in 3-dimensional scenes, as shown in figures 1 and 6. In both of these examples, a set of feature matches which are consistent with the epipolar geometry has been found. Figure 6 shows a pencil of epipolar lines. It can be seen that this epipolar geometry is consistent with the images, which are related by camera translation along the optical axis.

We have also found good results for object recognition problems, as is demonstrated in figure 7. In this example, we have solved for a homography between the object in the two views. Note the large scale changes between the objects in the images in this case.

5.1 Numerical Results

For quantitative assessment of our features we have used a sequence of images from the Annapurna region in Nepal (see figure 9). This sequence was chosen for the fact that the images are related by rotation about the camera centre, and

(a) UBC Rose Garden

(b) Correct feature matches

(c) Epipolar geometry

Fig. 6. A pair of images from the UBC Rose Garden. Similarity invariant features formed from groups of 2 interest points are extracted and matched. Outliers are rejected by requiring (loose) consistency with a global similarity transformation. Then, the fundamental matrix is computed and the epipolar constraint used to select a final set of consistent matches. Note that the epipolar geometry is consistent with a camera translation along the optical axis.

Fig. 7. Object recognition using invariant features from interest point groups. The white outlines show the recognised pose of each object according to a homography model. For 3D objects, this model is appropriate if the depth variation is small compared to the camera depth.

hence by (a 3 parameter family of) homographies. For images related by homographies we can compute *repeatability* and *success rate* metrics to characterise the performance of our matching scheme.

Repeatibility measures the fraction of interest points which are repeated in a pair of images. An interest point is said to be repeated if its position in a pair of images is consistent with the homography between them, up to some tolerance. See [18] for a precise definition. The aim of a point based matching scheme is to correctly match all of the repeated points. Hence it seems natural to define the *success rate* of matching as follows

$$\text{Descriptor Success Rate, } s = \frac{\text{Correctly Matched Interest Points}}{\text{Repeated Interest Points}}$$

That is, the success rate is the fraction of the repeated interest points that are correctly matched using the descriptors. Note that by definition $0 \leq s \leq 1$.

Figure 10 demonstrates the effect of sub-pixel / sub-scale accuracy of interest point location on repeatability. This correction gives a clear improvement in the accuracy of interest point location. It is particularly important for high levels of the pyramid, where sample spacings correspond to large distances in the base image. In addition to increasing the accuracy of transformation compu-

tations, accurate interest point localisation also enables more accurate feature descriptors, which improves matching.

Figure 11 shows the success rate of matching for each image in the test set using similarity invariant features. The figure shows success rates when using one feature per point, and when forming multiple features per interest point using multiple nearest neighbours. In these examples, we have extracted 1000 interest points. For the simplest case of constructing one similarity invariant feature per point, the success rate is around 50%. Typically, around 500 interest points are repeated between the images, so a 50% success rate corresponds to 250 correct matches. Constructing two features per point using the nearest two neighbours significantly increases the number of matched points, but there are diminishing returns when we use even more points.

6 Conclusions

In this paper we have introduced a family of features based on groups of scale-invariant interest points. The geometrical and illumination invariance of these features makes them particularly applicable for solving difficult correspondence problems. We have shown the importance of sub-pixel / sub-scale localisation of interest points, which critically improves the accuracy of descriptors. To reject outliers we use Hough transform clustering followed by RANSAC to select a set of feature matches that are loosely consistent with a global 2D transformation. We then compute the fundamental matrix, and use the epipolar constraint to reject remaining outliers. These techniques enable practical recognition and registration tasks to be performed in a few seconds of computation using 1GHz Pentium III machines.

Future work will look at more efficient parameterisation of feature descriptors, and alternative methods for computing local canonical frames.

References

1. Zhengyou Zhang, Rachid Deriche, Olivier Faugeras, and Quang-Tuan Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry,. *Artificial Intelligence, December 1995*, 78:87–119, 1995.
2. R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, ISBN: 0521623049, 2000.
3. T. Lindeberg. Scale-Space Theory: A Basic Tool for Analysing Image Structures at Different Scales. *Jounal of Applied Statistics*, 21(2):224–270, 1994.
4. T. Lindeberg. Feature Detection with Automatic Scale Selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.
5. D. Lowe. Object Recognition from Local Scale-Invariant Features. In *Proceedings of the International Conference on Computer Vision*, pages 1150–1157, Corfu, Greece, September 1999.
6. K. Mikolajczyk and C. Schmid. Indexing Based on Scale Invariant Interest Points. In *Proceedings of the International Conference on Computer Vision*, pages 525–531, 2001.

Fig. 8. Cylindrical panorama of the Annapurna sequence of images. This was computed by estimating the (3 dof) rotations between images from feature matches.

Fig. 9. Images from the Annapurna sequence. There are 15 images in total. These images were used to compute the numerical results presented in this paper. The images are related by rotation about the camera centre.

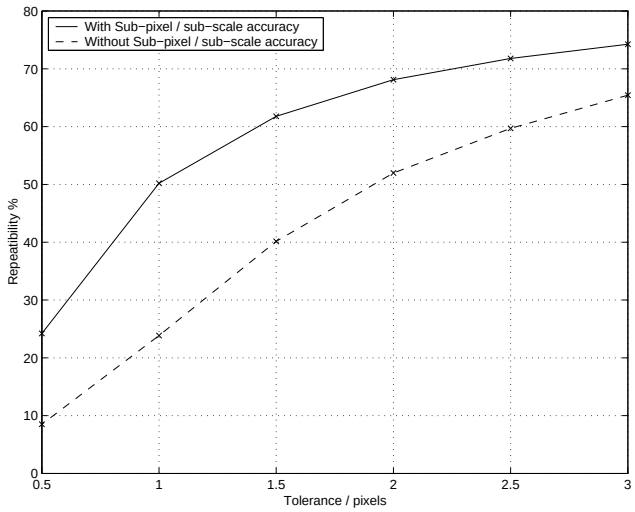


Fig. 10. Repeatibility of interest points with and without sub-pixel / sub-scale accuracy. Repeatibility is defined for images related by homographies as the fraction of interest points that are consistent with the homography, to within some tolerance.

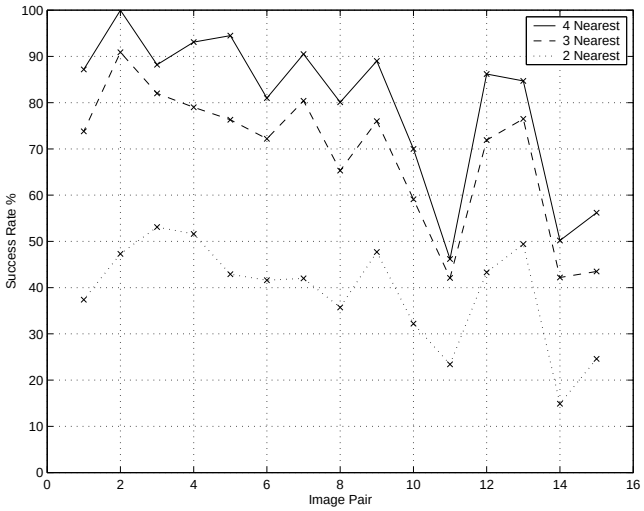


Fig. 11. Success rate of interest point matching for similarity invariant features. Success rate is the fraction of repeated interest points which are correctly matched. The curves show the results of constructing features from all combinations of the n nearest interest points, with $n = 2, 3$ and 4. Image pair refers to the pair of images from the Annapurna image sequence.

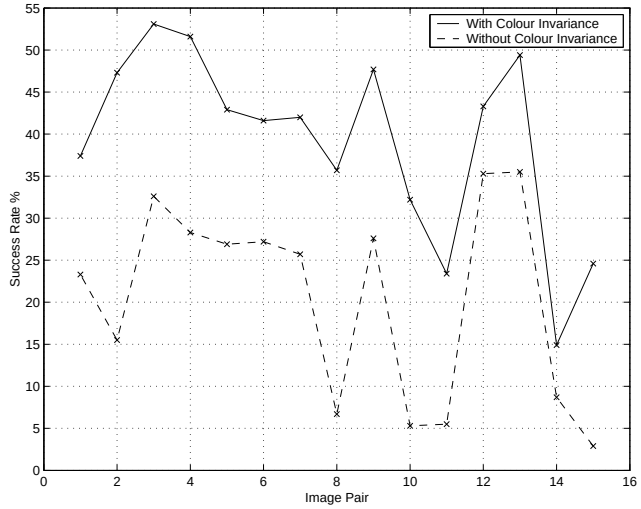


Fig. 12. Success rate of interest point matching with and without colour invariance. Colour invariance means that features are invariant to changes in the illumination spectrum. This is approximated in practice by normalising in each of the R, G, B channels.

7. C. Schmid and R. Mohr. Local Grayvalue Invariants for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, May 1997.
8. A. Baumberg. Reliable Feature Matching Across Widely Separated Views. In *Proceedings of the Interational Conference on Computer Vision and Pattern Recognition*, pages 774–781, 2000.
9. T. Tuytelaars and L. Van Gool. Wide Baseline Stereo Matching based on Local, Affinely Invariant Regions. In *Proceedings of the 11th British Machine Vision Conference*, pages 412–422, Bristol, UK, 2000.
10. D. Tell and S. Carlsson. Wide Baseline Point Matching Using Affine Invariants Compute From Intensity Profiles. In *Proceedings of the European Conference on Computer Vision*, pages 814–828, 2000.
11. C.A. Rothwell, A. Zisserman, D.A. Forsyth, and J.L. Mundy. Canonical frames for planar object recognition. In *Proceedings of the European Conference on Computer Vision*, pages 757–772, 1992.
12. C. Rothwell, A. Zisserman, D. Forsyth, and J. Mundy. Planar Object Recognition Using Projective Shape Representation. In *International Journal of Computer Vision*, number 16, pages 57–99, 1995.
13. B. Funt, K. Barnard, and L. Martin. Is Machine Colour Constancy Good Enough? In *Proceedings of 5th European Conference on Computer Vision (ECCV'98)*, pages 445–459, 1998.
14. P. Hough. Methods and Means for Recognizing Complex Patterns. U.S. Patent 3069654, December 1962.
15. D. Ballard. Generalizing the Hough Transform to Detect Arbitrary Shapes. *Pattern Recognition*, 13(2):111–122, 1981.
16. P. Burt and E. Adelson. The Laplacian Pyramid as a Compact Image Code. *IEEE Transactions on Communications*, 9(4):532–540, 1983.
17. J. Crowley and A. Parker. A Representation for Shape Based on Peaks and Ridges in the Difference of Low-Pass Transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):156–170, 1984.
18. C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of Interest Point Detectors. In *Proceedings of the International Conference on Computer Vision*, pages 230–235, Bombay, 1998.