# Multi-Image Matching using Multi-Scale Oriented Patches

Matthew Brown
*Department of Computer Science*
*University of British Columbia*
mbrown@cs.ubc.ca

Richard Szeliski
*Vision Technology Group*
*Microsoft Research*
szeliski@microsoft.com

Simon Winder
*Vision Technology Group*
*Microsoft Research*
swinder@microsoft.com

## Abstract

*This paper describes a novel multi-view matching framework based on a new type of invariant feature. Our features are located at Harris corners in discrete scale-space and oriented using a blurred local gradient. This defines a rotationally invariant frame in which we sample a feature descriptor, which consists of an $8 \times 8$ patch of bias/gain normalised intensity values. The density of features in the image is controlled using a novel adaptive non-maximal suppression algorithm, which gives a better spatial distribution of features than previous approaches. Matching is achieved using a fast nearest neighbour algorithm that indexes features based on their low frequency Haar wavelet coefficients. We also introduce a novel outlier rejection procedure that verifies a pairwise feature match based on a background distribution of incorrect feature matches. Feature matches are refined using RANSAC and used in an automatic 2D panorama stitcher that has been extensively tested on hundreds of sample inputs.*

## 1 Introduction

Early work in image matching fell into two camps – direct and feature-based. Feature-based methods attempt to extract salient features such as edges and corners and use a small amount of local information e.g. correlation of a small image patch, to establish matches [1, 2]. Direct methods attempt to use all of the pixel values in order to iteratively align images [3, 4].

At the intersection of these approaches are invariant features, which use large amounts of local image data around salient features to form invariant descriptors for indexing and matching. The first work in this area was by Schmid and Mohr [5], who used a jet of Gaussian derivatives to form a rotationally invariant descriptor around a Harris corner. Lowe extended this approach to incorporate scale invariance [6]. Other researchers have developed feature descriptors that are invariant under affine transformations [7, 8, 9]. Interest point detectors range from classic feature detectors such as Harris corners or DOG maxima to more elaborate methods such as maximally stable regions [10] and stable local phase structures [11]. There has also been good progress in evaluating these techniques with respect to interest point repeatability [12] and descriptor performance [13].

Invariant feature-based approaches to matching have been successfully applied to a wide range of problems, including object recognition [6], structure from motion[14], and panoramic image stitching [15]. In this paper, we concentrate on the latter application, where we expect the amount of foreshortening and scale variation to be fairly small.

While a tremendous amount of progress has been made recently in invariant feature matching, the final word has by no means been written. In this paper, we advance the state of the art in several directions. First, we develop a novel adaptive non-maximal suppression algorithm that better distributes features across the image than previous techniques (section 3). Second, we show that with suitable modifications, a direct patch-based sampling of the local image structure can serve as a useful invariant feature descriptor (section 4). Third, we develop a feature space outlier rejection strategy that uses all of the images in an $n$-image matching problem to give a background distribution for incorrect matches (section 5). Fourth, we develop an indexing scheme based on low-frequency Haar wavelet coefficients that greatly speeds up the search for feature correspondences with minimal impact on matching performance (section 6). We close the paper with a discussion of our results and ideas for future work in this area.

## 2 Interest Point Detection

The interest points we use are multi-scale Harris corners [1, 2]. For each input image $I(x, y)$ we form a Gaussian image pyramid $P_l(x, y)$ using a subsampling rate $s = 2$ and pyramid smoothing width $\sigma_p = 1.0$. Interest points are extracted from each level of the pyramid.

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

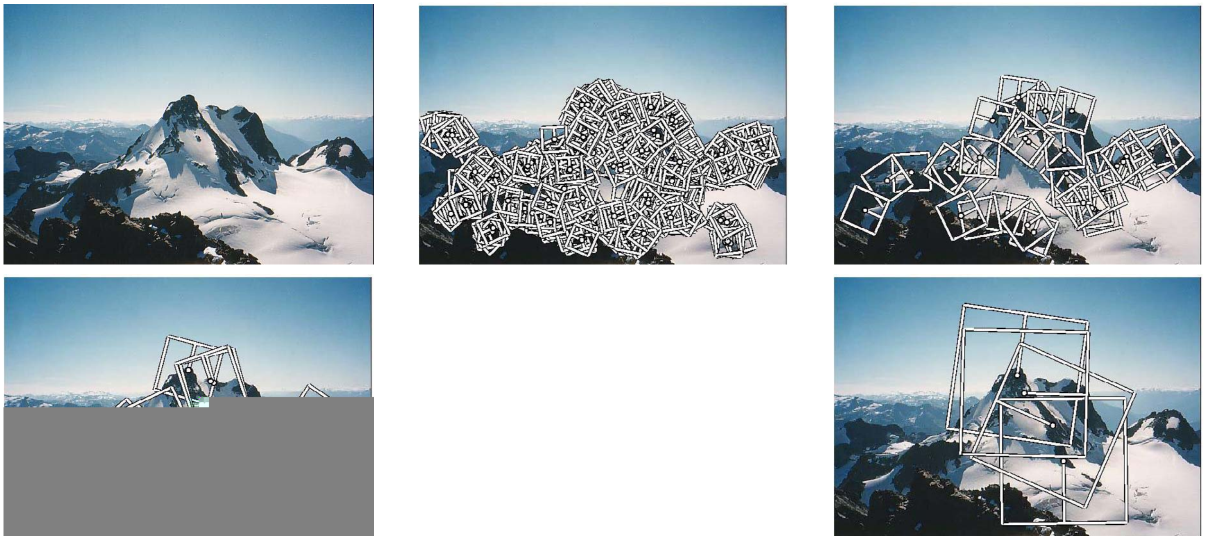


*Figure 1. Multi-scale Oriented Patches (MOPS) extracted at five pyramid levels from one of the Matier images. The boxes show the feature orientation and the region from which the descriptor vector is sampled.*

The Harris matrix at level $l$ and position $(x, y)$ is the smoothed outer product of the gradients

$$\mathbf{H}_l(x, y) = \nabla_{\sigma_d} P_l(x, y) \nabla_{\sigma_d} P_l(x, y)^T * g_{\sigma_i}(x, y)$$

We set the integration scale $\sigma_i = 1.5$ and the derivative scale $\sigma_d = 1.0$. To find interest points, we first compute the "corner strength" function

$$f_{HM}(x, y) = \frac{\det \mathbf{H}_l(x, y)}{\operatorname{tr} \mathbf{H}_l(x, y)} = \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2}$$

which is the harmonic mean of the eigenvalues $(\lambda_1, \lambda_2)$ of $\mathbf{H}$. Interest points are located where the corner strength $f_{HM}(x, y)$ is a local maximum in a $3 \times 3$ neighbourhood, and above a threshold $t = 10.0$. Once local-maxima have been detected, their position is refined to sub-pixel accuracy by fitting a 2D quadratic to the corner strength function in the local $3 \times 3$ neighbourhood and finding its maximum.

For each interest point, we also compute an orientation $\theta$, where the orientation vector $[\cos\theta, \sin\theta] = \mathbf{u}/|\mathbf{u}|$ comes from the smoothed local gradient

$$\mathbf{u}_l(x, y) = \nabla_{\sigma_o} P_l(x, y)$$

The integration scale for orientation is $\sigma_o = 4.5$. A large derivative scale is desirable so that the gradient field $\mathbf{u}_l(x, y)$ varies smoothly across the image, making orientation estimation robust to errors in interest point location.

## 3 Adaptive Non-Maximal Suppression

Since the computational cost of matching is superlinear in the number of interest points, it is desirable to restrict the maximum number of interest points extracted from each image. At the same time, it is important that interest points are spatially well distributed over the image, since for image stitching applications, the area of overlap between a pair of images may be small. To satisfy these requirements, we have developed a novel adaptive non-maximal suppression (ANMS) strategy to select a fixed number of interest points from each image.

Interest points are suppressed based on the corner strength $f_{HM}$, and only those that are a maximum in a neighbourhood of radius $r$ pixels are retained. Conceptually, we initialise the suppression radius $r = 0$ and then increase it until the desired number of interest points $n_{ip}$ is obtained. In practice, we can perform this operation without search as the set of interest points which are generated in this way form an ordered list.

The first entry in the list is the global maximum, which is not suppressed at any radius. As the suppression radius decreases from infinity, interest points are added to the list. However, once an interest point appears, it will always remain in the list. This is true because if an interest point is a maximum in radius $r$ then it is also a maximum in radius $r' < r$. In practice we robustify the non-maximal suppression by requiring that a neighbour has a sufficiently larger strength. Thus the minimum suppression radius $r_i$ is given by

$$r_i = \min_j |\mathbf{x}_i - \mathbf{x}_j|, \text{ s.t. } f(\mathbf{x}_i) < c_{\text{robust}} f(\mathbf{x}_j), \ \mathbf{x}_j \,\epsilon\, \mathcal{I}$$

where $\mathbf{x}_i$ is a 2D interest point image location, and $\mathcal{I}$ is the set of all interest point locations. We use a value $c_{\text{robust}} = 0.9$, which ensures that a neighbour must have significantly

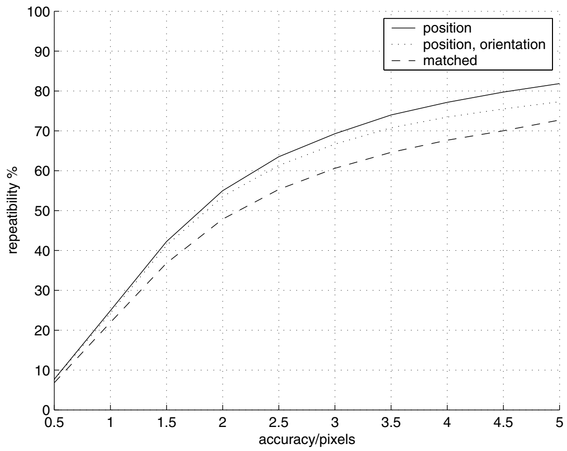


*Figure 3. Repeatability of interest points, orientation and matching for multi-scale oriented patches at the finest pyramid level. These results were obtained using the 7 images of the Matier dataset, each matched to 2 other images. Orientation repeatability means accurate to 3 standard devations (= 3 × $18.5\,^\circ$).*

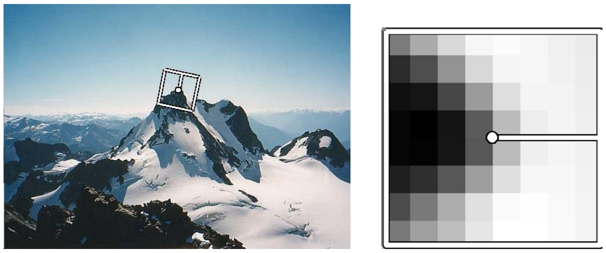


*Figure 4. Descriptors are formed using an $8 \times 8$ sampling of bias/gain normalised intensity values, with a sample spacing of 5 pixels relative to the detection scale. This low frequency sampling gives the features some robustness to interest point location error, and is achieved by sampling at a higher pyramid level than the detection scale.*

higher strength for suppression to take place. We select the $n_{ip} = 500$ interest points with the largest values of $r_i$.

Figure 2 compares the results of using regular feature detection (strongest responses) to adpative non-maximal suppression. Note how the latter features are much better distributed across the image. Experiments on a large database of panoramic images (section 7) show that distributing interest points spatially in this way, as opposed to selecting based on max corner strength, results in fewer dropped image matches.

### 3.1 Repeatability

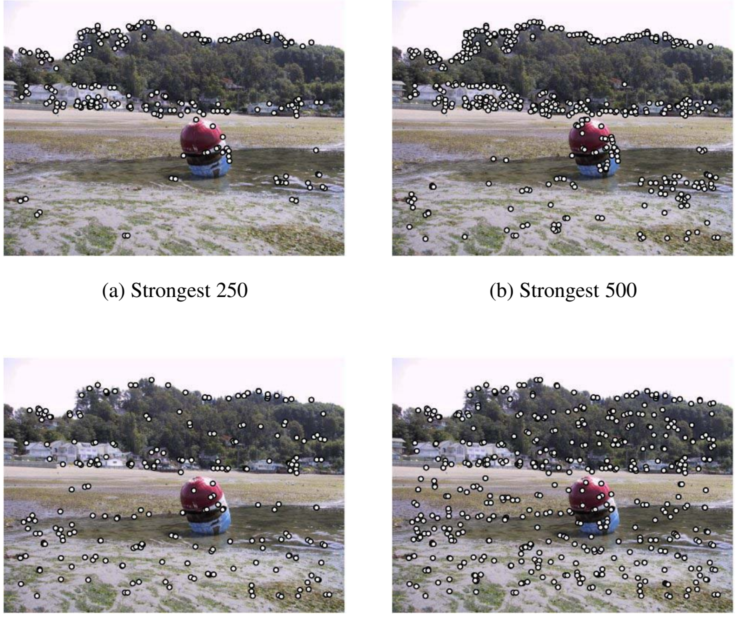
To evaluate the accuracy of interest point detection and orientation estimation, we have computed repeatability curves [12]. Figure 3 shows the fraction of interest points whose transformed position is correct (consistent with the global homography) up to some tolerance $\epsilon$. Note that contrary to the popular perception that Harris corners are subpixel accurate, the majority of interest points have location errors in the 1-3 pixel range.

## 4 Feature Descriptor

Once we have determined where to place our interest points, we need to extract a description of the local image structure that will support reliable and efficient matching of features across images. A wide range of such local feature vectors have been developed, including local intensity patches [1, 2], Gaussian derivatives [5], shift invariant feature transforms [6], and affine-invariant descriptors [7, 8, 9]. In their comparative survey, Mikolajczyk and Schmid [13] evaluated a variety of these descriptors and found that SIFT features generally perform the best. Local patches oriented to the dominant local orientation were also evaluated, but found not to perform as well.

In this section, we show how such patches can be made less sensitive to the exact feature location by sampling the pixels at a *lower frequency* than the one at which the interest points are located. Given an oriented interest point $(x, y, l, \theta)$, we sample a $8 \times 8$ patch of pixels around the sub-pixel location of the interest point, using a spacing of $s = 5$ pixels between samples (figure 4). To avoid aliasing, the sampling is performed at a higher pyramid level, such that the sampling rate is approximately once per pixel (the Nyquist frequency). Figure 5 shows how varying the sample spacing $s$ affects the reliability of feature matching. We have found that performance increases up to a value $s = 5$, with negligible gains thereafter.

After sampling, the descriptor vector is normalised so that the mean is 0 and the standard deviation is 1. This makes the features invariant to affine changes in intensity (bias and gain). Finally, we perform the Haar wavelet transform on the $8 \times 8$ descriptor patch $d_i$ to form a 64 dimensional descriptor vector containing the wavelet coefficients $c_i$. Due to the orthogonality property of Haar wavelets, Euclidean distances between features are preserved under this transformation. The first three non-zero wavelet coefficients $c_1, c_2, c_3$ are used in an indexing strategy described in section 6.



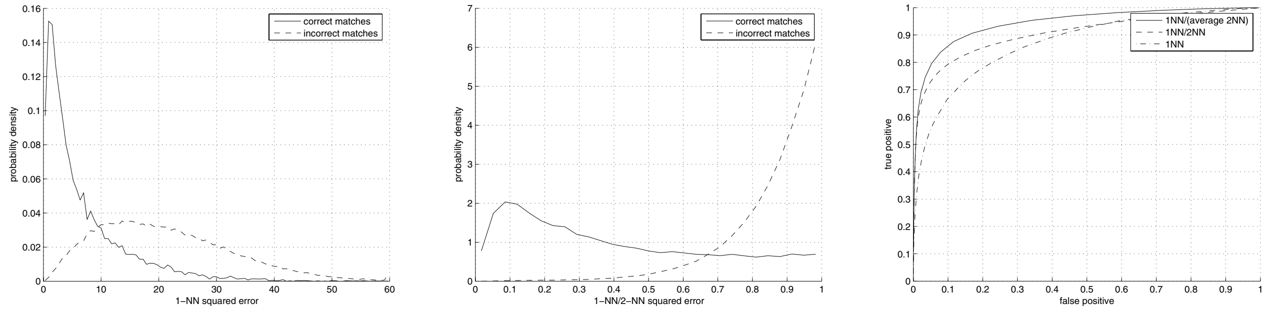
(a) Strongest 250

(b) Strongest 500

(c) ANMS 250, $r = 24$

(d) ANMS 500, $r = 16$

*Figure 2. Adaptive non-maximal suppression (ANMS). The two upper images show interest points with the highest corner strength, while the lower two images show interest points selected with adaptive non-maximal suppression (along with the corresponding suppression radius $r$). Note how the latter features have a much more uniform spatial distribution across the image.*



(a)

(b)

(c)

*Figure 6. Distributions of matching error for correct and incorrect matches. Note that the distance of the closest match (the 1-NN) is a poor metric for distinguishing whether a match is correct or not (figure (a)), but the ratio of the closest to the second closest (1-NN/2-NN) is a good metric (figure (b)). We have found that using an average of 2-NN distances from multiple images (1NN/(average 2-NN)) is an even better metric (figure (c)). These results were computed from 18567 features in 20 images of the Abbey dataset, and have been verified for several other datasets.*

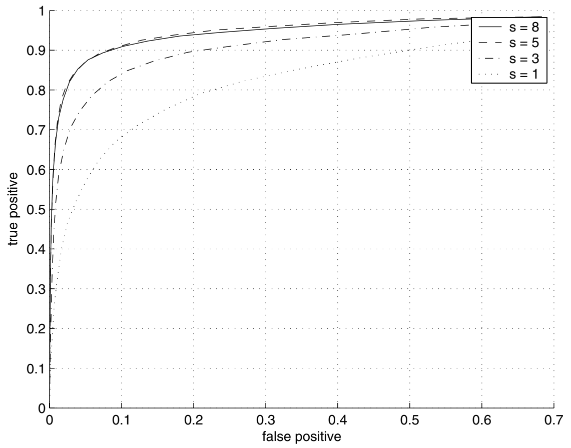


*Figure 5. Effect of changing the descriptor sample spacing on performance. These ROC curves show the results of thresholding feature matches based on normalised match distance as in section 5.1. Performance improves as the sample spacing increases (larger patches), but gains are minimal above a sample spacing of 5 pixels.*

# 5 Feature Matching

Given Multi-scale Oriented Patches extracted from all $n$ images, the goal of the matching stage is to find geometrically consistent feature matches between all images. This proceeds as follows. First, we find a set of candidate feature matches using an approximate nearest neighbour algorithm (section 6). Then we refine matches using an outlier rejection procedure based on the noise statistics of correct/incorrect matches. Finally we use RANSAC to apply geometric constraints and reject remaining outliers.

## 5.1 Feature-Space Outlier Rejection

Our basic noise model assumes that a patch in one image, when correctly oriented and located, corresponds to a patch in the other image modulo additive Gaussian noise:

$$\begin{aligned} I'(\mathbf{x}') &= \alpha I(\mathbf{x}) + \beta + n(\mathbf{x}) \\ \mathbf{x}' &= \mathbf{A}\mathbf{x} + \mathbf{t} \\ \mathbf{A} &= s \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \end{aligned}$$

$$n(\mathbf{x}) \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I})$$

where $I(\mathbf{x})$ and $I'(\mathbf{x})$ are the corresponding patches, and $n(\mathbf{x})$ is independent Gaussian noise at each pixel. However, we have found this model to be inadequate for classification, since the noise distributions for correctly and incorrectly matching patches overlap significantly (see figure 6(a)). Hence, it is not possible to set a global threshold on the matching error $e = \sum_{\mathbf{x}} n(\mathbf{x})^2$ in order to distinguish between correct and incorrect matches.

This behaviour has also been observed by Lowe [6], who suggested thresholding instead on the ratio $e_{1-NN}/e_{2-NN}$. Here $e_{1-NN}$ denotes the error for the best match (first nearest neighbour) and $e_{2-NN}$ denotes the error for the second best match (second nearest neighbour). As in Lowe's work, we have also found that the distributions of $e_{1-NN}/e_{2-NN}$ for correct and incorrect matches are better separated than the distributions of $e_{1-NN}$ alone (figure 6(b)).

The intuition for why this works is as follows. For a given feature, correct matches always have substantially lower error than incorrect matches. However, the overall scale of errors varies greatly, depending upon the appearance of that feature (location in feature space). For this reason it is better to use a discriminative classifier that compares correct and incorrect matches for a particular feature, than it is to use a uniform Gaussian noise model in feature space.

Lowe's technique works by assuming that the 1-NN in some image is a potential correct match, whilst the 2-NN in that same image is an incorrect match. In fact, we have observed that the distance in feature space of the 2-NN and subsequent matches is almost constant[1]. We call this the *outlier distance*, as it gives an estimate of the matching distance (error) for an incorrect match (figure 7).

We have found that in the $n$ image matching context we can improve outlier rejection by using information from all of the images (rather than just the two being matched). Using the same argument as Lowe, the 2-NN from each image will almost certainly be an incorrect match. Hence we *average* the 2-NN distances from all $n$ images, to give an improved estimate for the outlier distance. This separates the distributions for correct and incorrect matches still further, resulting in improved outlier rejection (figure 6(c)). Note that the extra overhead associated with computing 2-nearest neighbours in every image is small, since if we want to consider every possible image match, we must compute all of the 1-nearest neighbours anyway.

In general the feature-space outlier rejection test is very powerful. For example, we can eliminate 80% of the false matches for a loss of less than 10% correct matches. This allows for a significant reduction in the number of RANSAC iterations required in subsequent steps.

# 6 Fast Approximate Nearest Neighbours using Wavelet Indexing

To efficiently find candidate feature matches, we use a fast nearest-neighbour algorithm based on wavelet index-

[1] This is known as the shell property. The distances of a set of uniformly distributed points from a query point in high dimensions are almost equal.

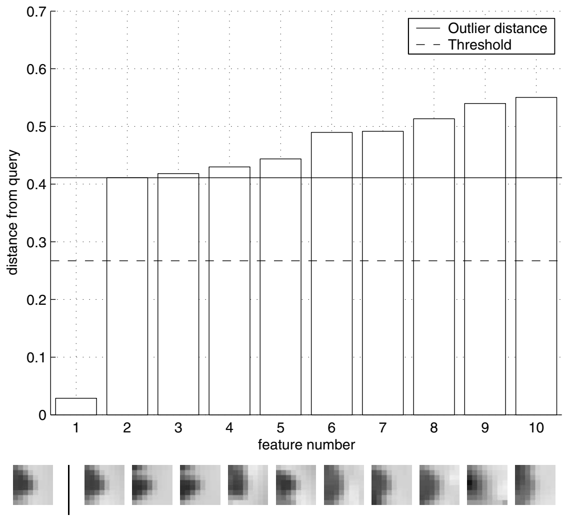


*Figure 7. Thresholding based on outlier distance. This figure shows the best 10 matches for a sample feature. The first is a correct match, and the rest are incorrect matches. Thresholding based purely on matching error gives poor results, since matching errors vary greatly depending upon the position in feature space. However, thresholding at a fraction of the outlier distance gives better results.*

ing. Features are indexed in a three-dimensional lookup table with dimensions corresponding to the first 3 non-zero wavelet coefficients $c_1, c_2, c_3$ (estimates of $\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial^2 I}{\partial x \partial y}$ over the patch). The lookup table has $b = 10$ bins per dimension, which cover $\pm n_\sigma = 3$ standard deviations from the mean of that dimension. Note that the means are typically around zero except for the first derivative that is aligned with the feature orientation, which is significantly positive.

The bins are overlapped so that data within half a bin width, i.e. $\frac{2n_\sigma}{b-1}\frac{1}{2} = \frac{\sigma}{3}$, are guaranteed to be matched against the query. These are approximate nearest neighbours as it is possible (but unlikely) that the true nearest neighbour lies outside $\frac{\sigma}{3}$ in one of the 3 dimensions. The query is exhaustively matched to all features in the query bin, and $k$ approximate nearest neighbours are selected. We then apply the outlier distance constraint as described in section 5.1 to verify correct matches and eliminate outliers. Indexing with $b$ bins on 3 dimensions gives a speedup of $b^3/2^3$ (assuming features are evenly distributed in the histogram) at the expense of some potential for lost feature matches.

Table 1 compares low frequency wavelet indexing to indexing using random grey values from the descriptor. At our chosen operating point of 10 bins per dimension, wavelet indexing retains 10% more matches than indexing on random grey values. At this operating point our indexing scheme gives a speedup of $10^3/2^3 = 125$ for a loss of less than 10% of the correct matches.

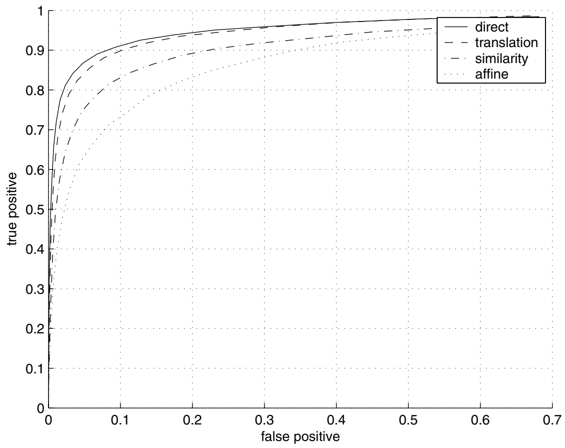


*Figure 8. ROC curves for patch refinement with different alignment models (Matier dataset). Each additional free parameter degrades the matching performance.*

| *Indexing Method* | *Dataset* | *Number of bins / dimension* | | | |
|---|---|---|---|---|---|
| | | *1* | *5* | *10* | *15* |
| Wavelet | Matier | 100 | 99.6 | 91.4 | 72.4 |
| | Abbey | 100 | 99.9 | 95.1 | 80.2 |
| Random pixel | Matier | 100 | 97.9 | 79.8 | 57.8 |
| | Abbey | 100 | 96.7 | 77.8 | 56.3 |

*Table 1. Indexing on wavelet coefficients vs. pixel values - percent recall in database matching. Using 10 bins per dimension, indexing on the 3 non-zero low frequency Haar wavelet coefficients (x and y derivatives and the cross derivative) gives about 10% better recall than indexing on random dimensions (pixels) of the descriptor.*

## 7 Experimental Results

### 7.1 Patch Refinement

In [13], Mikolajczyk and Schmid note that “It would be interesting to include correlation with patch alignment which corrects for these errors and to measure the gain obtained by such an alignment.” Since sensitivity to localization errors has been touted as one of the weaknesses of pixel-based descriptors, we decided to implement this suggestion to see how much it would help. Rather than computing sum-squared error on pixel patches (or wavelet coefficients) directly, we included a stage of Lucas-Kanade

[3] refinement to bring the patches more closely into spatial alignment before computing the pairwise descriptor distance. Since this has elements in common with the use of tangent distances [16] we expected that there might be an improvement in the separation of good and bad matches. Instead we found the opposite to be true.

We used four motion models (direct, translation, similarity and affine) with 0, 2, 4 and 6 parameters respectively. The results are shown in figure 8. Note that matching performance is degraded for each new parameter that is added to the model.

Since correct matches are already fairly well aligned, but bad matches typically have large errors, refinement tends to overfit the incorrect matches, whilst making only small improvements to the correct matches. This means that Lucas-Kanade refinement actually makes it more difficult to distinguish between correct and incorrect matches than before.

### 7.2 Panoramic Image Stitching

We have successfully tested our multi-image matching scheme on a dataset containing hundreds of panoramic images. We present results for the Matier and Abbey datasets in figure 9. See www.research.microsoft.com/∼szeliski/StitchingEvaluation for more examples.

## 8 Conclusions

We have presented a new type of invariant feature, which we call Multi-Scale Oriented Patches (MOPs). These features utilise a novel adaptive non-maximal suppression algorithm for interest point location, and a simple sampling of the (oriented) local image intensity for the feature descriptor. We have also introduced two innovations in multi-image matching. First, we have demonstrated an improved test for verification of pairwise image matches that uses matching results from all $n$ images. Second, we have shown that an indexing scheme based on low frequency wavelet coefficients yields a fast approximate nearest neighbour algorithm that is superior to indexing using the raw data values.

Future work will extend the features to incorporate true scale invariance, colour, and more robust rotation estimates. We would also like to perform a more detailed, quantitative analysis of feature matching performance on large databases of panoramic images.

## References

[1] W. Förstner. A feature-based correspondence algorithm for image matching. *Int.l Arch. Photogrammetry & Remote Sensing*, 26(3):150–166, 1986.

[2] C. Harris. Geometry from visual motion. In A. Blake and A. Yuille, editors, *Active Vision*, pages 263–284. MIT Press, 1992.

[3] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[4] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani. Hierachical model-based motion estimation. In *Proceedings of the 2nd European Conference on Computer Vision (ECCV92)*, pages 237–252. Springer-Verlag, May 1992.

[5] 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.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[7] A. Baumberg. Reliable feature matching across widely separated views. In *Proceedings of the Interational Conference on Computer Vision and Pattern Recognition (CVPR00)*, pages 774–781, 2000.

[8] 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 (BMVC00)*, pages 412–422, Bristol, UK, 2000.

[9] M. Brown and D. Lowe. Invariant features from interest point groups. In *Proceedings of the 13th British Machine Vision Conference (BMVC02)*, pages 253–262, Cardiff, 2002.

[10] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the 13th British Machine Vision Conference (BMVC02)*, 2002.

[11] G. Carneiro and A. Jepson. Multi-scale local phase-based features. In *Proceedings of the Interational Conference on Computer Vision and Pattern Recognition (CVPR03)*, 2003.

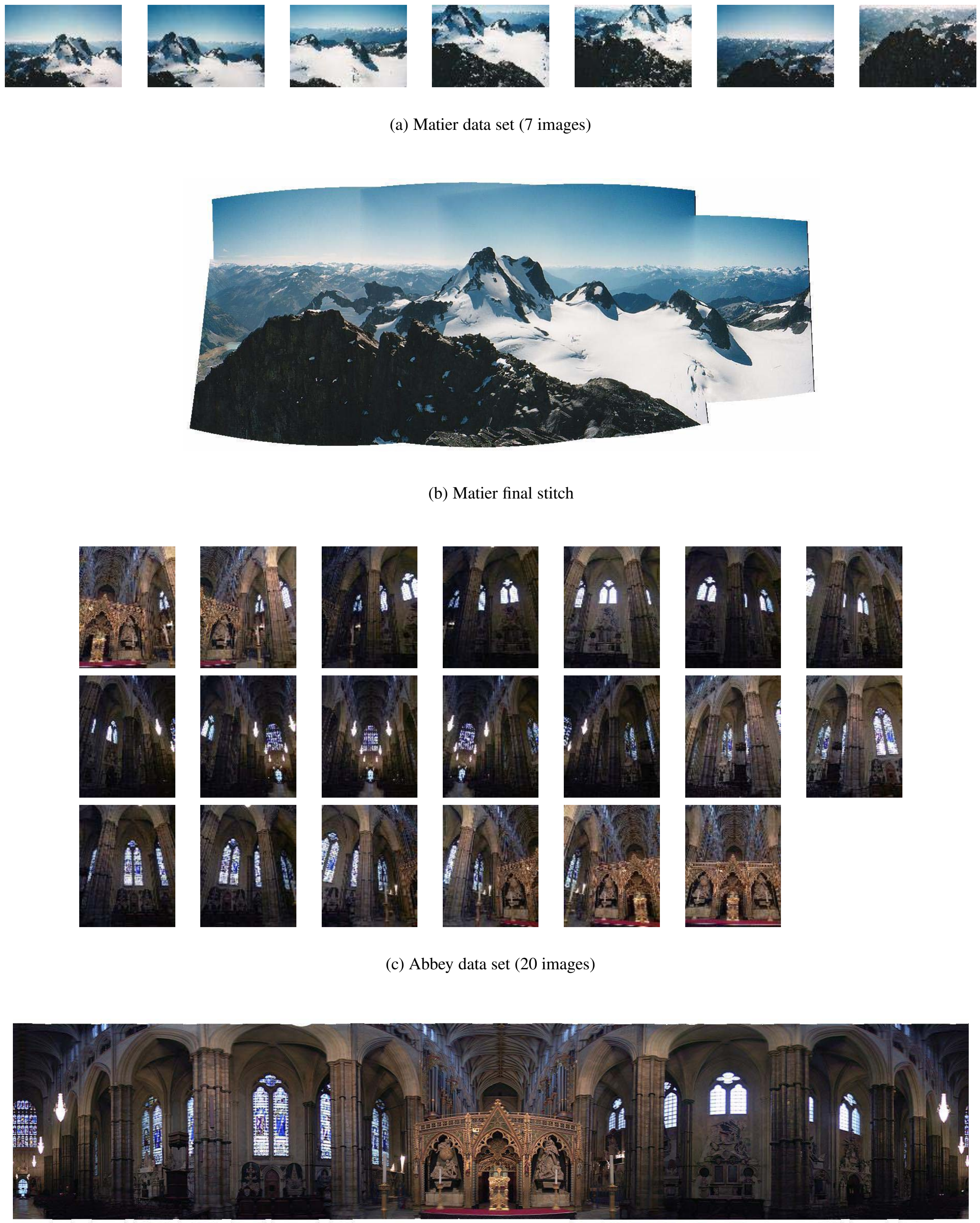
[12] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, June 2000.

[13] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proceedings of the Interational Conference on Computer Vision and Pattern Recognition (CVPR03)*, 2003.

[14] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organise my holiday snaps?". In *Proceedings of the 7th European Conference on Computer Vision (ECCV02)*, pages 414–431, 2002.

[15] M. Brown and D. Lowe. Recognising panoramas. In *Proceedings of the 9th International Conference on Computer Vision (ICCV03)*, volume 2, pages 1218–1225, Nice, October 2003.

[16] Patrice Simard, Yann LeCun, John S. Denker, and Bernard Victorri. Transformation invariance in pattern recognition-tangent distance and tangent propagation. In *Neural Networks: Tricks of the Trade*, pages 239–27, 1996.



(a) Matier data set (7 images)

(b) Matier final stitch

(c) Abbey data set (20 images)

(d) Abbey final stitch

*Figure 9. The stitched images used for the matching results found in this paper. We have successfully tested our multi-image matching scheme on a database containing hundreds of panoramic images. See http://www.research.microsoft.com/∼szeliski/StitchingEvaluation for more examples.*