Permutable Descriptors for Orientation-Invariant Image Matching

Gabriel Takacs[†] Vijay Chandrasekhar[†] Huizhong Chen[†]
David Chen[†] Sam Tsai[†] Radek Grzeszczuk[‡] Bernd Girod[†]

[†] Information Systems Laboratory, Stanford University, Stanford CA, USA
[‡] Nokia Research Center, Palo Alto CA, USA

ABSTRACT

Orientation-invariant feature descriptors are widely used for image matching. We propose a new method of computing and comparing Histogram of Gradients (HoG) descriptors which allows for re-orientation through permutation. We do so by moving the orientation processing to the distance comparison, rather than the descriptor computation. This improves upon prior work by increasing spatial distinctiveness. Our method method allows for very fast descriptor computation, which is advantageous since many mobile applications of HoG descriptors require fast descriptor computation on hand-held devices.

1. INTRODUCTION

As personal computing continues its migration from the desktop to the hand-held device, many new interesting applications in computer vision have appeared. Such applications include real-time image matching and Augmented Reality (AR). Unfortunately, these applications come with the challenges of running complex algorithms on small, under-powered devices. Though devices will continue to increase in clock speed, they will likely remain limited in power resources. Thus, light-weight algorithms will continue to be important. Real-time image matching and augmented reality on hand-held devices require fast algorithms that can surpass these challenges.

The dominant technique for image matching is to use local feature descriptors. Detecting and computing these descriptors can be time-consuming, and thus researchers have put effort into developing light-weight algorithms. This work demonstrates a new light-weight method for computing descriptors with state-of-the-art robustness.

1.1 Prior Work

Mobile vision applications require a descriptor that can be computed at video frame rates on a hand-held device. Therefore, we wish to have a very simple descriptor, while maintaining state-of-the-art robustness. The most popular descriptor, SIFT,[1] is robust, but far too slow for these real-time applications. Even the much faster SURF[2] is not up to the task of real-time feature extraction. People have proposed sufficiently fast descriptors,[3,4] however, the robustness of these descriptors is not satisfactory. We have recently proposed a Rotation-Invariant, Fast Feature (RIFF)[5] descriptor which builds on ideas from CHoG.[6] For speed, we have removed the orientation assignment phase of keypoint detection. Though RIFF is extremely fast and works well for small databases, the annular binning required for orientation invariance takes a toll on the distinctiveness of the descriptor.

There are two prominent techniques for rotation invariance in the current literature. The first uses steerable filters with descriptor permutation.[7,8] This method suffers from high computational overhead of computing many filter orientations. The second technique is to treat rotation as a circular shift and use the coefficients of the Fourier transform,[9] or the dual-tree complex wavelet transform.[10] This method is often not sufficiently robust to viewpoint variation, although, recent work by Nelson and Kingsbury[11] has improved this robustness. An additional technique has been proposed by Brasnett and Bober,[12] and included in the MPEG-7 Image Signatures standard.[13] Their method computes scalar statistics over circular regions. Though these signatures are rotation invariant, they are not robust to viewpoint variation.

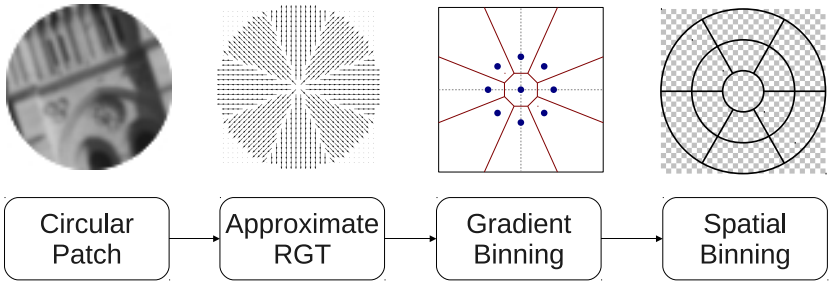


Figure 1. Descriptor computation pipeline. We start with a circular image patch and then compute the Approximate Radial Gradient Transform (ARGT) on the patch. Each gradient is then quantized before being placed in the appropriate spatial bin.

1.2 Contributions

In this work we present the RIFF-Polar descriptor, which provides state-of-the-art robustness and speed. We improve upon our previous work on RIFF by increasing the distinctiveness of the descriptor with more spatial bins. To do so, we propose computing upright descriptors and using a distance metric which minimizes over possible orientations. This way, we are able to achieve state-of-the-art robustness and distinctiveness for image matching tasks. We also propose a novel distance metric which provides increased angular resolution without suffering from decreased robustness. The proposed schemes are compared with a variety of prior work to demonstrate the performance in image matching tasks.

The structure of the remainder of the paper is as follows. We first introduce and evaluate the RIFF-Polar descriptor in Section 2. We then evaluate the Receiver Operating Characteristic (ROC) and image matching performance in Section 3.

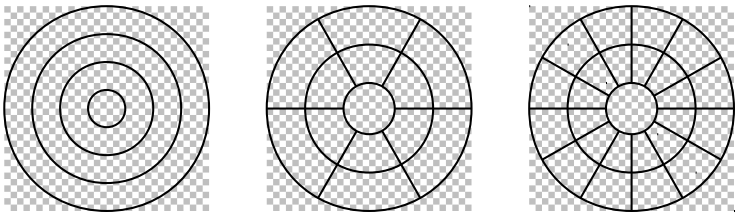
2. PERMUTABLE DESCRIPTOR

The RIFF-Polar descriptor pipeline is similar in nature to many Histogram of Gradients (HoG) descriptors. In particular, the gradients from a local image patch are quantized and binned and then placed in spatial bins to form a set of histograms. As in CHoG, we prefer to maintain the compressed version of the histogram of gradients, rather than statistics of the histogram, as in SIFT and SURF. Descriptors are compared by using a symmetric KL-divergence or an L_1-norm, both of which are meaningful metrics over probability mass functions.

As with the previously proposed annular RIFF,[5] we obtain orientation invariance of the gradient histograms with the Radial Gradient Transform (RGT). The RGT takes gradients along radial and tangential directions, relative to the center of the feature. For speed with no loss in robustness, the RGT is approximated by taking gradients along only eight directions. These directions naturally align with the image pixel grid. Once the Approximate RGT is computed, the radial gradients are quantized using either a vector or scalar quantizer. Following prior investigations,[5,6] we use a 9-bin quantizer as a good balance between low dimensionality and high performance. The vector quantizer is shown in Figure 1. For speed, this quantizer can be well approximated by a scalar quantizer. The quantized gradients are then binned into the appropriate spatial cell. This pipeline is illustrated by Figure 1. For additional speed it is possible to skip pixels in the local patch. Skipping pixels reduces the number of samples in the histogram, but does not change the underlying distribution.

2.1 Spatial Binning

The key to HoG descriptors is the balance between robustness to viewpoint change and distinctiveness. Robustness comes from having large enough spatial bins such that small offset errors have a small impact on the resulting histogram. This is because the number of pixels that are put in the wrong bin is small relative to the total number of pixels in the bin. However, distinctiveness is obtained by having many segregated spatial bins. Thus, there is a trade off between robustness and distinctiveness.



Annuli 4 *Polar 3×6* *Polar 3×12*

Figure 2. Cell configurations for spatial binning. More cells increases distinctiveness, while larger cells increases robustness to scale and offset errors.

Figure 2 shows the spatial binning configurations that we use in this paper. All configurations are based on a 40×40 pixel patch extracted around the interest point. The *Annuli 4* configuration provides orientation invariance without a minimizing distance function, but suffers low distinctiveness due to having a small number of large bins. Unfortunately, as the number of annuli increases, the width of each bin decreases, reducing robustness. Therefore adding more annular bins does not necessarily improve the descriptor.

To improve the distinctiveness without reducing robustness, we must introduce angular cell divisions. The *Polar 3×6* and *Polar 3×12* configurations are instances of such cell divisions, with 3 radial divisions, and 6 and 12 angular divisions, respectively. The *Polar 3×6* configuration provides a good balance between the number of cells and the cell size. Unfortunately, such angular divisions destroy the orientation invariance of the descriptor. In the next section, we will show how to regain orientation invariance, while maintaining the robustness and distinctiveness of the descriptor.

In addition to the hard cell boundaries shown in Figure 2, we also use a soft binning technique, similar to DAISY.[7] With soft binning, each gradient is weighted and distributed to each spatial bin according to the distance from the center of that bin. For this descriptor, the gradient binning is also soft in the same way. Such soft binning provides more robustness, as boundary pixels are less affected by an offset.

2.2 Distance Metric

For applications, such as SIFT, SURF, and CHoG where the patches are oriented before computing descriptors, a simple, fixed, distance metric suffices. However, for speed at the descriptor extraction stage, we wish to compute upright feature descriptors as they do not require resampling. Such descriptors can be later reoriented if rotation-invariance is desired. To reorient a RIFF-Polar descriptor, all we need to do is permute the histograms, which yields the same result as having rotated the image patch by the width of an angular bin. We define a permutation distance, $D_{\texttt{ori}}(p, q)$, as the minimum over all permutations,

$$D(p, q) = \min_{\theta} D(p, q_\theta) \tag{1}$$

where $D(p, q)$ is the underlying distance function, p and q are the descriptors, and q_θ is q permuted by the angle θ. This yields a minimum distance and associated angle of rotation. However, since the possible permutations represent discrete samples of θ, we can get a better estimate of the distance and angle by interpolating between the samples. We do so by fitting a parabola to the minimum-distance sample and its left and right neighbors. The minimum value of this parabola is the estimated distance, and its location is the associated angle.

In general, any underlying distance function, $D(p, q)$, can be used. However, since our descriptors are histograms of gradients, we prefer a distance that is well-suited for probability distributions, such as L_1-norm or symmetric KL-divergence. For all subsequent experiments we use symmetric KL-divergence, since we have previously shown that it performs well for CHoG.[6]

For descriptors computed on oriented patches, as few as three or four angular divisions typically suffice. However, to be able to reorient the descriptor after computation, more angular divisions, and thus resolution, are required. As we will show later, the 60 degree resolution provided by *Polar 3×6* is insufficient, therefore we introduce the *Polar 3×12* configuration which provides an angular resolution of 30 degrees.

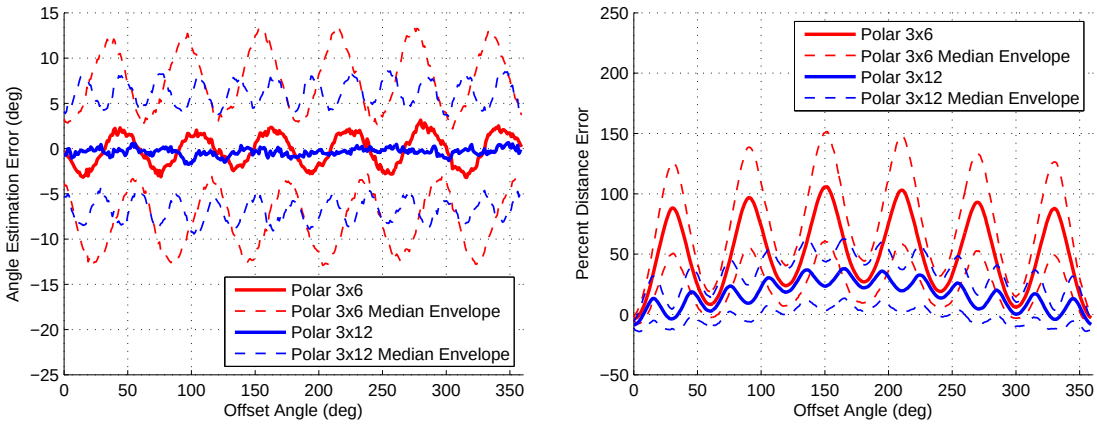


Figure 3. Evaluation of the permutation distance function for angle estimation (*left*) and distance estimation (*right*). The average error for both cell configurations are plotted (*solid*), along with the median absolute error envelope (*dashed*).

Though *Polar 3×12* provides better resolution it may suffer a loss in robustness due to small cells. This problem motivates the use of a *Hybrid* descriptor and distance function. The hybrid distance operates on the *Polar 3×12* configuration, but aggregates adjacent angular histogram bins. This allows for 30 degree resolution, but the robustness of *Polar 3×6*. Other variants of this scheme are possible, but have not been evaluated.

We have evaluated the distance function by measuring the distance between 1000 matching pairs of image patches from the *Liberty* dataset,[8] and then rotating one patch and using the permutation distance to determine the distance and angle of rotation. We can then compute the average error between the true angle and estimated angle, as well as the true and estimated distances. These errors are plotted in Figure 3 as a function of angle. For this experiment we use a symmetric KL-divergence as the distance.

Figure 3 (*left*) shows how well the angle can be estimated using our permutation distance. We can see that *Polar 3×12* is better for angle estimation, yielding a flatter mean error and a tighter median envelope. The median envelope shows that 50% of distances are estimated to within about 6 degrees. The oscillations are also smaller than *Polar 3×6*, and at half the period.

Angle estimation is important, but evaluation of the distance estimation will give us a better clue to how the distance function will perform in image matching. Figure 3 (*right*) shows the mean error for distance estimation, given as a percent of the correct distance. At its peaks, *Polar 3×6* can be off by 100%, while *Polar 3×12* is off by half of that at 50%. The error for *Polar 3×12* is also significantly flatter versus angle. This error in distance is important since nearest neighbor searching is used for image matching. If the distance between the correct descriptor match is increased then it may no longer be the nearest match. As we will see in Section 3.2, the large errors of *Polar 3×6* significantly effect image matching, while the smaller errors of *Polar 3×12* are well tolerated.

3. DESCRIPTOR EVALUATION

Having designed our descriptor and distance metric, we now present an evaluation and comparison of the proposed schemes. We evaluate our descriptor using two methods, Receiver Operating Characteristic (ROC) and pairwise image matching. The ROC curve gives a good indication of the average performance of the descriptor, while pairwise image matching shows the performance in a real matching system.

3.1 Receiver Operating Characteristic

The best way to compare feature descriptors without the rest of the image matching apparatus is to use the ROC curve. We generate ROC curves using the methods of Winder *et al.*,[8] using their *Liberty* dataset. The dataset consists of two sets of image patches from photos of the Statue of Liberty that have been matched and reconstructed in 3D. The first set has matching and oriented pairs, and the second set has non-matching pairs. To test the rotation-invariance property of descriptors, we modify the dataset by rotating one patch in each of

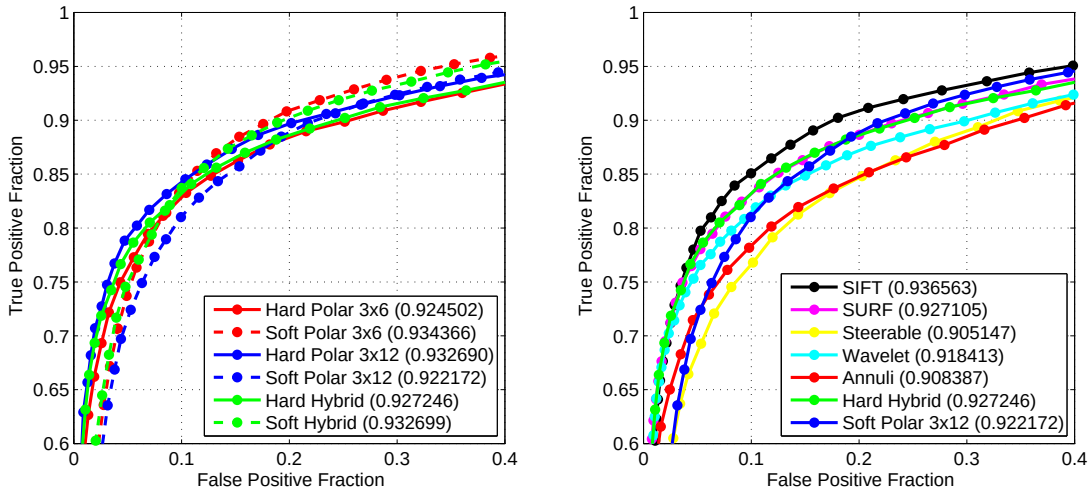


Figure 4. Receiver operating characteristics for the *Liberty* dataset, showing a comparison of the proposed schemes (*left*), and a comparison with prior work (*right*). The value in parentheses next to the label is the area under the ROC curve. Higher curves and areas are better.

the matching pairs by a uniform random amount. By measuring a histogram for the distance between pairs in each set, we can evaluate the effect of a threshold on the fraction of true-positive and false-positive matches.

We perform two ROC comparisons. First, we compare the proposed descriptor configurations to each other. These configurations include hard and soft versions of *Polar 3×6*, *Polar 3×12*, and *Hybrid*. Second, we compare our proposed schemes to other state-of-the art methods, including SIFT, SURF, steerable,[8] and wavelet[11] descriptors. These results are shown in Figure 4 (*left*) and (*right*), respectively. The value in parentheses next to the label is the area under the ROC curve.

In Figure 4 (*left*), we can see that our proposed schemes all perform similarly in an ROC test. Many of the curves cross, making it difficult to distinguish an overall best configuration. The best configuration depends on the allowable fraction of false positives. Again, by looking at the area under the curves we see that they are all within about one percent of each other. Further, as we will see later, the angular resolution of the descriptor is not well accounted for by the ROC. This means that to truly compare these descriptors for orientation invariance, we must use a real matching system. Such matching comparisons are presented in the next section.

In Figure 4 (*right*), we demonstrate the ROC performance of our *Hard Hybrid* and *Soft Polar 12* schemes, compared to other prior work. We choose these two schemes because of their superior pairwise image matching performance, as shown below. The first two comparisons, to SIFT and SURF, are performed with the upright *Liberty* dataset. This dataset must be used since these descriptors are not rotation invariant, but instead rely on the interest point detector for orientation. SIFT outperforms all other schemes, and sets the performance benchmark. SURF is a widely used descriptor that works very well in diverse applications, and has the benefit of being faster than SIFT. Therefore, though it does not match the performance of SIFT, SURF's ROC performance is quite respectable, as the descriptor has proven itself.

Almost exactly matching the performance of SURF, is our *Hard Hybrid* scheme. Following that, the next best performance is achieved by both the *Soft Polar 12* and *Wavelet* schemes, whose ROC curves cross. Such performance with the wavelet-based scheme is impressive because of the robustness challenges inherent with this type of descriptor. The *Annuli* scheme under-performs all except the steerable filter descriptor. As previously mentioned, this relatively low performance is due to low distinctiveness caused by the small number of spatial bins. The *Steerable* scheme is similar to that described by Winder *et al.*,[8] except that we use our *Polar 3×6* spatial binning configuration. To achieve orientation invariance we also use our proposed distance function. We see that the *Steerable* scheme performs the worst under these conditions, however, its performance could be improved by proper tuning.

Figure 5. Example image pairs from the *CD* dataset. A clean database picture (*top*) is matched against a real-world picture (*bottom*) at various orientations.

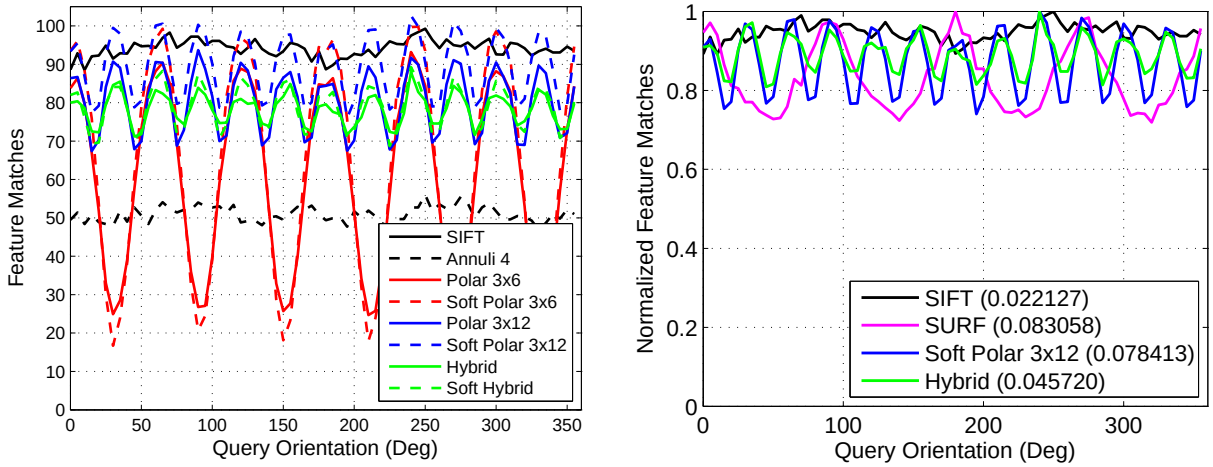


Figure 6. Pairwise image matching results comparing our proposed schemes (*left*), and normalized matching results comparing the variance of our schemes with prior work (*right*). Higher and flatter are better. The values in parentheses are standard deviations.

One distinct advantage of the *Hard Hybrid* RIFF scheme is that it can be computed in the same manner as RIFF-Annuli,[5] and thus is extremely fast. Though it is just as fast as RIFF-Annuli, *Hard Hybrid* provides about twice the distinctiveness and robustness. Additionally, the same descriptor can be used for either upright or rotation invariant matching applications. Therefore, for tracking applications, where adjacent frames are very similar, a simple upright distance function suffices. This means that the only additional computational overhead incurred by our scheme is during image matching, which is typically not done at video frame rates.

3.2 Pairwise Image Matching

In this section, we show how ROC performance translates into pairwise image matching performance. The experiments use images from a database of CD/DVD/book cover images.[14] Sample images from this *CD* dataset are shown in Figure 5. Note that the 500×500-pixel database images are clean, while the 640×480-pixel queries contain perspective distortion, background clutter, and glare.

We wish to test not only the matching performance, but also the rotational invariance of our descriptor. To do so, the query images are rotated in 5° increments, and matched against the corresponding upright database images. We use a ratio-test followed by RANSAC to ensure that the resulting feature matches are correct. SIFT and RIFF schemes use the same Difference of Gaussian (DoG) interest points, while SURF uses fast Hessian points. We first compare our proposed schemes to each other in Figure 6 (*left*), where we plot the average number of feature matches versus the amount of rotation, averaged over 10 image pairs. We then compare the rotation invariance of our best schemes to SIFT and SURF in Figure 6 (*right*), where we plot the normalized average number of matches versus rotation angle.

We saw in the Section 3.1 that all of the proposed schemes perform similarly with an ROC metric. So, to distinguish them and to select the best, we turn to the image matching results in Figure 6 (*left*). Here we can clearly see the effects of the angular resolution of the spatial binning. For a benchmark, we have also included results for SIFT and our previously proposed RIFF-Annuli. First, we note that the the angle resolution of *Polar 3×6* leads to large, periodic drops in the number of matches. These dead-zones occur when the query descriptors are offset by half of a spatial bin. Simply by increasing the number of angular divisions to 12, the *Polar 3×12* schemes eliminate these drastic dips. The flatness of the response can be further enhanced by using the *Hybrid* schemes. Second, we note that the soft schemes increase the distinctiveness of the descriptor, thereby raising the peaks, but lowering the troughs. This means that the soft version of a scheme has higher variance than the hard version. The *Soft Polar 3×12* scheme has the overall highest performance, due to another relationship between 3×12 and 3×6, which is that the 3×12 curve follows the 3×6 curve near its peaks. This allows for *Soft Polar 3×12* to have similar variance as the hard version, with an higher average performance.

Given the previous comparison we select two schemes; the one with the highest average performance, and the one with the lowest variance. We compare the variance of these schemes to SIFT and SURF in Figure 6 (*left*). The results in this plot are normalized by their peak value which allows for a direct, fair comparison of the schemes. We see that the flattest response is that of SIFT, followed in order by *Hybrid*, *Soft Polar 3×12*, and SURF. The large anisotropy of SURF is caused by box filtering in the interest point detector. Though the proposed schemes have a larger variance than SIFT, the fact that their response is flatter than SURF demonstrates that the variance is in an acceptable range for most applications.

4. CONCLUSIONS

We have presented a novel, state-of-the-art method for computing permutable descriptors. The goal of the method is to remove orientation assignment from the descriptor computation. We are able to do so and still obtain rotation-invariance by moving the orientation assignment to the distance metric. We propose two such schemes, *Soft Polar 3×12* which has a high average performance, and *Hard Hybrid* which has better isotropy and is faster to compute. Both of our proposed schemes are more isotropic than SURF and perform comparably to SIFT. With no additional overhead, the design of *Hard Hybrid* fits well with the previously proposed RIFF tracking scheme. This RIFF-Polar descriptor improves upon the distinctiveness of the previously proposed RIFF-Annuli. Mobile applications in augmented reality and real-time image matching can benefit from such a robust, light-weight local feature descriptor.

REFERENCES

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

[2] Bay, H., Tuytelaars, T., and Gool, L. V., "SURF: Speeded Up Robust Features," in [*Proc. of European Conference on Computer Vision (ECCV)*], (May 2006).

[3] Michael Calonder, Vicent Lepetit, Pascal Fua, Kurt Konolige, James Bowman, Patrick Mihelich, "Compact Signatures for High-speed Interest Point Description and Matching," in [*Int. Conf. on Computer Vision (ICCV)*], (2009).

[4] Simon Taylor, Edward Rosten, Tom Drummond, "Robust Feature Matching in 2.3μs," in [*Conference on Computer Vision and Pattern Recognition*], (June 2009).

[5] Gabriel Takacs, Vijay Chandrasekhar, D. Chen, S. Tsai, R. Grzeszczuk, B. Girod, "Unified real-time tracking and recognition with rotation-invariant fast features," in [*IEEE Conference on Computer Vision and Pattern Recognition*], (June 2010).

[6] Vijay Chandrasekhar, Gabriel Takacs, and David Chen, Sam Tsai, Radek Grzeszczuk, Bernd Girod, "CHoG: Compressed Histogram of Gradients, a Low Bitrate Descriptor," in [*Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*], (2009).

[7] Engin Tola, Vincent Lepetit, Pascal Fua, "A Fast Local Descriptor for Dense Matching," in [*CVPR*], (2008).

[8] Simon Winder, Gang Hua, Matthew Brown, "Picking the best DAISY," in [*Conference on Computer Vision and Pattern Recognition*], (2009).

[9] Ahonen T, Matas J, He C, Pietikinen M, "Rotation invariant image description with local binary pattern histogram fourier features," in [*Image Analysis, SCIA 2009 Proceedings, Lecture Notes in Computer Science 5575*], (2009).

[10] Nick Kingsbury, "Rotation-Invariant Local Feature Matching with Complex Wavelets," in [*Proc. European Conf. Signal Processing (EUSIPCO)*], (2006).

[11] Nelson, J. D. B. and Kingsbury, N. G., "Enhanced shift and scale tolerance for dual-tree complex wavelet rotation invariant matching," in [*IEEE Transactions on Image Processing*], (2010).
[12] P. Brasnett and M. Bober, "A Robust Visual Identifier Using the Trace Transform," in [*Visual Information Engineering Conference*], (Jul 2007).
[13] ISO/IEC 15938-3:2009 - Information technology – Multimedia content description interface – Part 1: Visual (2009).
[14] Chen, D. M., Tsai, S. S., Vedantham, R., Grzeszczuk, R., and Girod, B., *CD Cover Database: Query Images* (April 2008).