COMPARISON OF LOCAL FEATURE DESCRIPTORS FOR MOBILE VISUAL SEARCH

Vijay Chandrasekhar[1], David M. Chen[1], Andy Lin[1], Gabriel Takacs[1], Sam S. Tsai[1], Ngai-Man Cheung[1], Yuriy Reznik[2], Radek Grzeszczuk[3], Bernd Girod[1]

[1]Information Systems Laboratory, Stanford University, Stanford, CA 94305
[2]Qualcomm Inc., San Diego, CA 92121
[3]Nokia Research Center, Palo Alto, CA 94304

ABSTRACT

We evaluate the performance of MPEG-7 image signatures, Compressed Histogram of Gradients descriptor (CHoG) and Scale Invariant Feature Transform (SIFT) descriptors for mobile visual search applications. We observe that SIFT and CHoG outperform MPEG-7 image signatures greatly in terms of feature-level Receiver Operating Characteristic (ROC) performance and image-level matching. Moreover, CHoG descriptors demonstrate such gains while being comparable with MPEG-7 image signatures in bit-rate.

Index Terms— MPEG-7 Image Signature, feature descriptor, mobile visual search, image signature.

1. INTRODUCTION

Mobile phones have evolved into powerful image and video processing devices, equipped with high-resolution camera, color displays, and hardware-accelerated graphics. They are also equipped with location sensors, GPS receivers, and connected to broadband wireless networks allowing fast transmission of information. This enables a class of applications which use the camera phone to initiate search queries about objects in visual proximity to the user. Such applications can be used for identifying products, comparison shopping, finding information about movies, CDs, real estate or products of the visual arts. Google Goggles [1], Nokia Point and Find [2] and Snaptell [3] are examples of recently developed commercial applications. For these applications, a query photo is taken by a mobile device and compared against previously stored database photos. A set of image feature descriptors is used to assess the similarity between the query photo and each database photo. This feature set needs to be robust against geometric and photometric distortions encountered when the user takes the query photo at an arbitrary viewpoint in an unknown lighting environment.

The size of the data sent over the network needs to be as small as possible to reduce latency and improve user experience. One approach to the problem is to transmit the JPEG compressed query image over the network, but this might be prohibitively expensive at low uplink speeds. An alternate approach is to extract feature descriptors on the phone, compress the descriptors and transmit them over the network as illustrated in Figure 1. Such an approach has been demonstrated to reduce the amount of transmission data significantly [4, 5]. Furthermore, feature extraction can be carried out quickly ($<$ 1 second) on current generation phones making this approach feasible [6]. In this work, we focus on the latter approach.

1.1. Prior Work

SIFT [7], Speeded Up Robust Features (SURF) [8], Gradient Location and Orientation Histogram (GLOH) [9], CHoG [4] are some examples of feature descriptors proposed in the literature. The review paper by Mikolajczyk *et al.* [9] compares the performance of several descriptors. However, Mikolajczyk *et al.* do not take the bit-rate of descriptors into account in their comparisons.

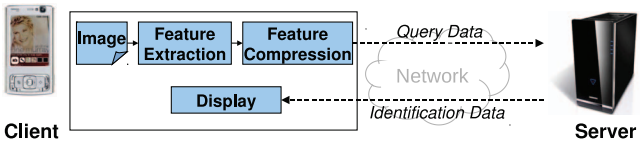


Fig. 1. A mobile CD cover recognition system where the server is located at a remote location. Feature descriptors are extracted on the mobile-phone and query feature data is sent over the network. Once the CD cover is recognized on the server, identification data is sent back to the mobile-phone.

Low bit-rate feature descriptors are of increasing interest to the computer vision community. Often, feature vectors are reduced by decreasing the dimensionality of descriptors via Principle Component Analysis (PCA) or Linear Discriminant Analysis (LDA) [10, 11, 12]. In [13], we have studied dimensionality reduction and entropy coding of SIFT and SURF descriptors. Yeo *et al.* [14] and Shakhnarovich *et al.* [15] reduce the bit-rate of descriptors by using projections on SIFT descriptors to build binary hashes. As part of the MPEG-7 standard, Brasnett and Bober [16] propose a 60-bit feature descriptor, which will be the focus of this evaluation.

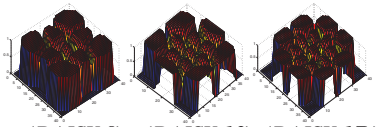
In our work [4, 5], we propose a framework for computing low bit-rate feature descriptors called CHoG. Gradient histograms are quantized using Huffman trees, Type Quantization or Lloyd Max Vector Quantization and compressed efficiently using fixed and variable length codes. CHoG descriptors can be compared directly in the compressed domain eliminating the need for decompression in the descriptor matching process. In [5], we provide a comprehensive comparison of several low bit-rate descriptors proposed in the literature and show that CHoG outperforms all other schemes.

1.2. Outline

In this work, we compare the performance of MPEG-7 image signature tools [16], SIFT and CHoG in the context of mobile visual search applications. In Section 2, we review MPEG-7 image signatures and CHoG descriptors. In Section 3, we discuss feature level Receiver Operating Characteristic (ROC) experiments and pairwise image-matching experiments for the different descriptors.

2. BACKGROUND

In Section 2.1, we review MPEG-7 image signatures and the matching pipeline used for comparing two sets of image signatures. In Section 2.2, we discuss CHoG descriptors and the compression method used to achieve low bit-rates.



(*DAISY-9*) (*DAISY-13*) (*DAISY-17*))

Fig. 2. The DAISY spatial binning configurations used for $n = 9, 13, 17$ spatial bins. We use soft spatial binning where each pixel contributes to multiple spatial bins.

2.1. MPEG-7 Image Signature

As part of the MPEG-7 standard, Brasnett and Bober [16] propose a pipeline for detecting near-duplicates and images with very similar content. The pipeline is shown to be highly effective in detecting scaled, rotated, flipped and compressed variants of a query image.

The matching pipeline relies on 512-bit global and 60-bit local descriptors. Both global and local descriptors are based on the multi-resolution Trace transform which constructs a set of 1-D representations of an image [16]. A binary identifier is extracted from each representation using a Fourier transform. The 512-bit global descriptor is computed on the entire image. Up to 80 Difference of Gaussian (DoG) interest points and Harris corners are detected in the image scale-space. The 60-bit local desciptors are used to describe patches extracted around these interest points. The 512-bit global image signature is effective for detection of near-duplicates but is not effective when comparing images of the same object taken at different perspectives and lighting conditions. Hence, we do not use the 512-bit global signature in our comparisons here.

Next, we discuss how image matching is done with two sets of local descriptors. Descriptors are compared using Hamming distance. A 4-stage geometric matching scheme is proposed in the standard for matching two sets of local descriptors. Hypotheses are formed in stages one and three. A series of geometric tests are performed in stages two and four. The tests in stages two and four must be passed in order for a hypothesis to progress to the next stage. The stages become increasingly computationally complex so that each stage aims to minimise the number of hypotheses that are accepted for subsequent processing. The geometric matching pipeline is robust to affine transforms.

2.2. CHoG Descriptor

Lowe [7], Bay *et al.* [8], Dalal and Triggs [17], Freeman and Roth [18] and Winder *et al.* [12] have proposed histogram of gradient based descriptors. The CHoG [4] descriptor also falls in this category.

2.2.1. Descriptor Computation

The patch extracted around the interest point is first divided into localized cells. The CHoG descriptor uses polar spatial binning configurations [19, 12] as shown in Figure 2. Next, we quantize the gradient histogram in each spatial bin. Let $P_{D_x,D_y}(d_x, d_y)$ be the normalized joint (x, y)-gradient histogram in each spatial bin. We coarsely quantize the 2D gradient histogram and capture the histogram directly into the descriptor. We approximate $P_{D_x,D_y}(d_x, d_y)$ as $\hat{P}_{\hat{D}_x,\hat{D}_y}(\hat{d}_x, \hat{d}_y)$ for $(\hat{d}_x, \hat{d}_y) \in S$, where S represents a small number of quantization centroids or bins as shown in Figure 3. The histogram binning schemes exploit the underlying gradient statistics observed in patches extracted around interest points, as shown in Figure 3. We perform a Vector Quantization (VQ) of the gradient distribution into a small set of bin centers, S, shown in Figure 3. We call these bin configurations VQ-3, VQ-5, VQ-7 and VQ-9. As we increase the number of bin centers, we obtain a more accurate approximation of the gradient distribution and the performance of the descriptor improves [4].

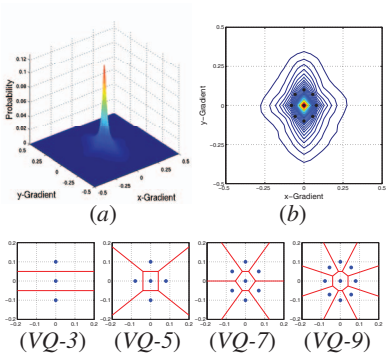


Fig. 3. The joint (d_x, d_y) gradient distribution (a) over a large number of cells, and (b), its contour plot. The greater variance in y-axis results from aligning the patches along the most dominant gradient after interest point detection. The quantization bin constellations VQ-3, VQ-5, VQ-7 and VQ-9 and their associated Voronoi cells are shown at the bottom.

2.2.2. Descriptor Compression

We quantize the gradient histogram in each cell individually and map it to an index. The indices are then encoded with fixed-length or entropy codes, and the bitstream is concatenated together to form the final descriptor. In prior work [4, 5], we have explored several schemes for histogam compression. One such histogram compression scheme is type coding [5], which we describe here.

Let m represent the number of gradient bins. m varies from 3 to 9 for the VQ bin configurations shown in Figure 3. Let $P = [p_1, p_2, ...p_m] \in R_+^m$ be the original distribution as described by the gradient histogram, and $Q = [q_1, q_2, ...q_m] \in R_+^m$ be the quantized probability distribution defined over the same sample space. For type coding, given a parameter n, we first construct a lattice of distributions (or *types*) $Q = Q(k_1, \ldots, k_m)$ with probabilities

$$q_i = \frac{k_i}{n}\,, \;\; k_i,\, n \in Z_+\,, \;\; \sum_i k_i = n \qquad (1)$$

We then pick and transmit the index of the type that is closest to the original distribution P. The quantization scheme used for finding values $\{k_i\}$, given P and n as input parameters is described in [5]. The paramater n controls the fidelity of quantization. The higher the value of n parameter, higher the fidelity. The total number of types $K(m, n)$ is the number of partitions of n into m terms $k_1 + \ldots + k_m = n$

$$K(m, n) = \binom{n+m-1}{m-1}\,, \qquad (2)$$

implying that the rate $R(m, n)$ needed for type encoding is upper-bounded according to $R(m, n) \leq \log_2 K(m, n) \sim (m-1)\log_2 n$. The paramater $n \sim m$ typically provides good trade-off between bit-rate and feature error rate [5]. Feature-level ROC performance and image-matching performance obtained by varying m and n are discussed in Section 3. Next, we map the quantized type to an index. The algorithm that maps a type to its index $f_n : \{k_1, \ldots, k_m\} \rightarrow [0, K(m, n) - 1]$ is described in [5]. In the final step, we encode the index in each spatial cell with fixed-length or entropy codes.

Patch Modification	Description
Rotate 180^o	Rotate patch by 180^o
Rotate 90^o	Rotate patch by 90^o
Brightness	Increase all pixel intensity by 25%
Blur	Blur each patch with a Gaussian filter of σ=2
Shift	Circular shift each patch by 1 pixel
Winder-Brown	Patches that correspond to same 3-D point at different scales, orientation and lighting.

Table 1. Description of patch modifications applied to *Liberty* data.

Spatial bins	Gradient bins	Type param n	Bit-rate
DAISY-9	VQ-5	5	56
DAISY-9	VQ-7	7	81
DAISY-13	VQ-5	5	82
DAISY-13	VQ-7	7	119

Table 2. CHoG descriptor parameters

Fixed-length encoding provides the benefit of compressed domain matching at the cost of a small performance hit.

We use symmetric Kullback Leibler (KL) divergence for comparing CHoG descriptors as it is shown to perform better than using L_1 or L_2 norm [4]. For matching sets of descriptors, we use the ratio test scheme proposed in [7] followed by a RANSAC affine consistency check. We use a threshold of 0.9 for the ratio test as it gives a good tradeoff between false-acceptance and false-rejection rates.

3. RESULTS

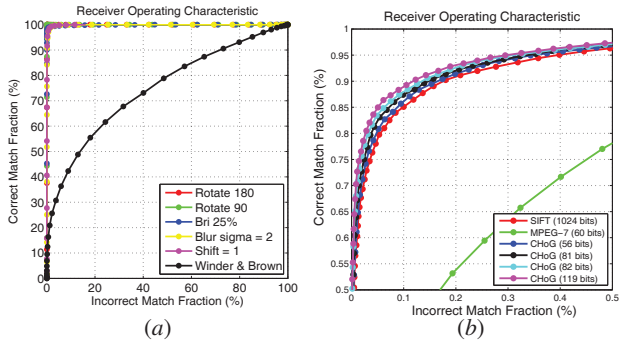
In Section 3.1, we discuss feature level ROC experiments and in Section 3.2, we discuss pairwise image-matching experiments for the different descriptors.

3.1. Feature Level Experiments

Feature level experiments are evaluated using the data sets provided by Winder and Brown [12]. We randomly select 10,000 matching pairs and 10,000 non matching pairs from the *Liberty* set for testing purposes. We use the method proposed by Winder and Brown [12] for descriptor evaluation. We compute a distance between each pair of descriptors. From these distances, we form two histograms, one for matching pairs and one for non-matching pairs. From the two histograms we obtain a ROC curve which plots correct match fraction against incorrect match fraction.

First, we perform a control experiment to validate the effectiveness of MPEG-7 signatures for the simple modifications they were designed for. We apply the set of modifications shown in Table 1 to the *Liberty* patches. Note that the *Winder-Brown* modification in the table refers to matching pairs obtained from the data sets provided by the authors. For simple modifications, we observe that the MPEG-7 signature performs well as seen from the high ROC performance in Figure 4. We observe a large gap in performance when we plot the ROC performance for the *Winder-Brown* patch modification. From this, we conclude that the MPEG-7 descriptor is robust to simple image modifications like scaling, rotation, cropping and compression, but is not robust to the kinds of scale, orientation, perspective and photometric distortions present in the Winder and Brown data set.

Next, the 60 bit MPEG-7 image signature is compared to low bit rate CHoG descriptors, and the 128-dimensional SIFT descriptor. The 128-dimensional SIFT descriptor is quantized to 8 bits in each dimension resulting in a 1024 bit descriptor. We use Hamming



(*a*) (*b*)

Fig. 4. Figure (*a*) shows the performance of MPEG-7 signature for different kinds of patch modifications. The MPEG-7 signature is robust to simple modifications but is not robust to the distortions present in the Winder-Brown data set. Figure (*b*) shows comparison of 60 bit MPEG-7 image signature to CHoG descriptors for the Winder-Brown data set. We observe that the 56 bit CHoG descriptor outperforms the 60 bit MPEG-7 signature by a big margin. Note that the low bit-rate CHoG descriptors perform on par with the 1024 bit SIFT descriptor.

distances for comparing MPEG-7 signatures, L2 norm for comparing SIFT descriptors and symmetric KL divergence for comparing CHoG descriptors. The parameters and bit-rates for the CHoG descriptors are listed in Table 2. We observe in Figure 4 that low bit-rate CHoG descriptors perform on par with SIFT and outperform MPEG-7 image signatures by a significant margin.

3.2. Image Level Experiments

The Zurich Building Database [20] is used for pairwise image matching experiments. The database consists of 1005 building images and 115 query images. The database is small enough for pairwise image matching to be feasible. We match each query image with each database image and declare the database image with the highest number of feature correspondences to be the matching candidate. Matching accuracy is defined as the number of correctly identified query images. The original database images have resolution 640×480, while the original query images have resolution 320×240. In our experiments, we upsampled the query images to resolution 640×480 as it improves image matching performance. Pairwise image matching allows fair comparison of different feature descriptors independent of the techniques used for large-scale retrieval [21]. The performance of CHoG descriptors in a large-scale retrieval system with a million images is discussed in [5].

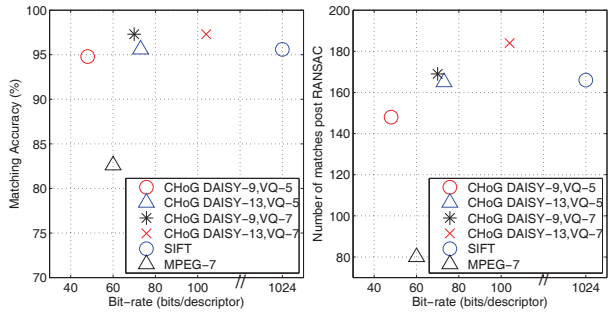
Pairwise image matching results are summarized in Table 3. We use a threshold of 0.9 for the ratio test prior to RANSAC geometric matching for all descriptors. The number of feature correspondences after the geometric consistency step for different descriptors is shown in Figure 5. We make the following observations from Table 3.

Comparing schemes (1) and (2), we observe that the RANSAC based approach provides a higher matching accuracy than the geometric matching scheme proposed in the MPEG-7 standard. By visually inspecting matching feature correspondences, we observe that the MPEG-7 standard produces more false positives than the ratio test/RANSAC based approach. Furthermore, we observe that the MPEG-7 geometric matching scheme is less robust to challenging affine geometric distortions.

We note that the matching accuracy for schemes (1) and (2) at 70% and 76% is low. The matching accuracy of these two schemes

#	Descriptor Type	Keypoint Type	Geometric Matching	Maximum Keypoints	Acc (%)
1	MPEG-7	MPEG-7	MPEG-7	80	70.4
2	MPEG-7	MPEG-7	RANSAC	80	76.5
3	MPEG-7	DoG	RANSAC	600	82.6
4	SIFT	DoG	RANSAC	600	95.6
5	CHoG	DoG	RANSAC	600	97.4

Table 3. Image matching performance of different schemes



(*a*) (*b*)

Fig. 5. Figure (*a*) compares the matching accuracy of MPEG-7, CHoG and SIFT descriptors. Note the gap in performance between CHoG and MPEG-7 descriptors. CHoG descriptors perform on par with SIFT while being comparable in bit-rate to MPEG-7 descriptors. Figure (*b*) shows the average number of feature matches post RANSAC for the different descriptors. There is a $2\times$ gap in performance between CHoG and MPEG-7 descriptors at a comparable bit-rate.

is low primarily because the maximum number of keypoints/image is limited to 80 in the MPEG-7 standard. We conclude that 80 keypoints/image is not sufficient for visual search applications.

In schemes (3)-(5), we do not restrict the number of keypoints to 80 per query image. For schemes (3)-(5), we use the DoG keypoint detector available online [22]. We extract upto 600 descriptors from each image and use the same matching pipeline for the different descriptors. The CHoG descriptor used for comparison here has a bit-rate of 73 bits/descriptor. Increasing the maximum number of keypoints from 80 to 600 improves the matching accuracy of the MPEG-7 scheme from 76% to 82%. However, note that we can achieve a higher matching accuracy using CHoG or SIFT descriptors with 600 keypoints. The CHoG descriptor performs the best with an accuracy of 97.4%. This indicates that CHoG and SIFT descriptors are more discriminative than MPEG-7 image signatures, as also inferred previously from the feature-level experiments in Figure 4. Note, however, that the CHoG descriptor achieves the performance of the SIFT descriptor at a bit-rate comparable to the MPEG-7 image signature.

Finally, we compare MPEG-7, CHoG and SIFT descriptors using the same matching pipeline described in schemes (3)-(5) in Table 3. We compare matching accuracy and average number of matching feature correspondences post RANSAC for MPEG-7, SIFT and the different CHoG descriptors listed in Table 2. We observe a significant gap in matching accuracy between MPEG-7 ($\sim$80%) and the different CHoG descriptors ($> 95\%$) at a comparable bit-rate. Furthermore, there is a $2\times$ gap in the number of feature matches between CHoG and MPEG-7 descriptors. From Figure 5, we conclude that low bit-rate CHoG descriptors perform on par with SIFT while being comparable in bit-rate to the MPEG-7 descriptor.

4. CONCLUSION

We have evaluated the performance of MPEG-7 image signatures, CHoG, and SIFT descriptors for mobile visual search. We conclude that both SIFT and CHoG outperform MPEG-7 image signatures significantly in terms of both feature-level ROC performance and image-level matching. Moreover, CHoG descriptors demonstrate such gains while being comparable with MPEG-7 image signatures in bit-rate. Based on this comparison, we conclude that the CHoG descriptor is a better alternative to MPEG-7 image signatures and SIFT for mobile visual search applications.

5. REFERENCES

[1] *Google Goggles*, `http://www.google.com/mobile/goggles/`.

[2] *Nokia Point and Find*, `http://www.pointandfind.nokia.com`.

[3] *SnapTell*, `http://www.snaptell.com`.

[4] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, R. Grzeszczuk, and B. Girod, "CHoG: Compressed Histogram of Gradients - A low bit rate feature descriptor," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, Florida, June 2009.

[5] V. Chandrasekhar, Y. Reznik, G. Takacs, D. M. Chen, S. S. Tsai, R. Grzeszczuk, and B. Girod, "Study of Quantization Schemes for Low Bitrate CHoG descriptors," in *Proceedings of IEEE International Workshop on Mobile Vision (IWMV)*, San Francisco, California, June 2010.

[6] G. Takacs, V. Chandrasekhar, D. M. Chen, S. S. Tsai, R. Grzeszczuk, and B. Girod, "Unified Real-time Tracking and Recognition with Rotation Invariant Fast Features," in *Accepted to IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, SFO, California, June 2010.

[7] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

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

[9] K. Mikolajczyk and C. Schmid, "Performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[10] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," in *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*. 2004, vol. 02, pp. 506–513, IEEE Computer Society.

[11] G. Hua, M. Brown, and S. Winder, "Discriminant Embedding for Local Image Descriptors," in *Proc. of International Conference on Computer Vision (ICCV)*, 2007.

[12] S. Winder, G. Hua, and M. Brown, "Picking the best daisy," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, Miami, Florida, June 2009.

[13] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, and B. Girod, "Transform coding of feature descriptors," in *Proc. of Visual Communications and Image Processing Conference (VCIP)*, San Jose, California, January 2009.

[14] C. Yeo, P. Ahammad, and K. Ramchandran, "Rate-efficient visual correspondences using random projections," in *Proc. of IEEE International Conference on Image Processing (ICIP)*, San Diego, California, October 2008.

[15] G. Shakhnarovich and T. Darrell, "Learning Task-Specific Similarity," *Thesis*, 2005.

[16] P. Brasnett and M.Z.Bober, "Robust visual identifier using the trace transform," in *Proc. of IET Visual Information Engineering Conference (VIE)*, London, UK, July 2007.

[17] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2005.

[18] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *Proc. of International Workshop on Automatic Face and Gesture Recognition*, 1994, pp. 296–301.

[19] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[20] L. Van Gool H.Shao, T. Svoboda, "Zubud-Zürich buildings database for image based recognition.," Tech. Rep. 260, ETH Zürich, 2003.

[21] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, USA, June 2006.

[22] *SIFT code*, `http://www.vlfeat.org/~vedaldi/code/code.html`.