RATE-ACCURACY OPTIMIZATION OF BINARY DESCRIPTORS

Alessandro Redondi[⋆], *Luca Baroffio*[⋆], *Joao Ascenso*[†], *Matteo Cesana*[⋆], *Marco Tagliasacchi*[⋆]

[⋆] Dipartimento di Elettronica e Informazione, Politecnico di Milano
[†]Instituto Superior de Engenharia de Lisboa - Instituto de Telecomunicações, Lisbon

ABSTRACT

Binary descriptors have recently emerged as low-complexity alternatives to state-of-the-art descriptors such as SIFT. The descriptor is represented by means of a binary string, in which each bit is the result of the pair-wise comparison of smoothed pixel values properly selected in a patch around each keypoint. Previous works have focused on the construction of the descriptor neglecting the opportunity of performing lossless compression. In this paper, we propose two contributions. First, design an entropy coding scheme that seeks the internal ordering of the descriptor that minimizes the number of bits necessary to represent it. Second, we compare different selection strategies that can be adopted to identify which pair-wise comparisons to use when building the descriptor. Unlike previous works, we evaluate the discriminative power of descriptors as a function of rate, in order to investigate the trade-offs in a bandwidth constrained scenario.

Index Terms— Visual features, coding.

1. INTRODUCTION

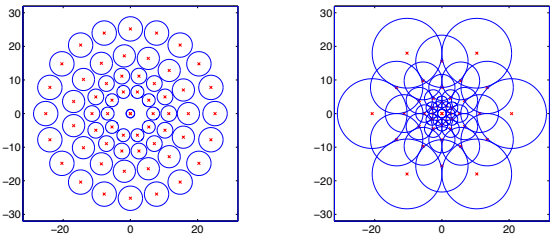
Visual features provide a concise representation of the underlying content that are robust and invariant to many global and local transformations. In recent years, the focus has been on the design of local features [1][2][3]. That is, salient keypoints are identified by means of a detector, and a descriptor is computed from the pixel values belonging to the image patch around the keypoint. Although the design of a descriptor assume different forms, in [3] it was shown that descriptors are typically computed by processing an image patch according to three modules: pre-smoothing, transformation and spatial pooling. For example, the state-of-the-art SIFT descriptor [1] consists of local histograms (pooling) of gradients (transformation) of patches smoothed with a Gaussian kernel.

Extracting local features from visual content can be computationally demanding, due to the computation of the detector and the descriptor. In recent years, several works have addressed the problem of reducing the complexity while retaining the desirable invariance and robustness properties of state-of-the-art visual features. As for detectors, several low-complexity algorithms have been proposed, in the case of both corner (e.g., FAST [4], AGAST [5], etc.) and blob (fast Hessian [2], CenSurE [6], etc.) detectors. Similarly, in the case of descriptors, there are two approaches which have been pursued in the literature. Some descriptors, e.g., SURF [2], are designed as a fast approximation of SIFT. Instead, a different line of investigation includes the case of binary descriptors. The simplest descriptor of this class is BRIEF [7], which generates a binary string whose bits are the result of the comparison of pairs of (smoothed) pixel values selected at random within a patch around the keypoint. BRISK [8] constraints the pixel locations to be used for the comparisons, so as to limit the memory accesses. Long pairs, i.e., those whose pixel locations are further apart than a threshold, are used to determine the orientation of the patch, thus achieving rotation invariance. Short pairs are instead used to build the descriptor. The choice of the pairs to be included in the descriptor is further investigated by FREAK [9], which proposes a heuristic rule that tries to maximize the discriminative power of the descriptor. Recently, DBRIEF was proposed [10]. The elements are the result of the binarization of discriminative projections that can be computed fast.

In some applications, both complexity and bandwidth represent a scarce resource. This is the case, for example, of wireless multimedia sensor networks, in which battery-operated nodes are used to sense the visual scene. Instead of acquiring, compressing and transmitting data in the pixel domain, an alternative approach consists of computing a feature-based representation, in which features can be conveniently compressed and sent to a remote node for further processing [11]. In our previous work [12], we investigated the rate-accuracy trade-offs that can be achieved for the case of SURF descriptors, adopting different lossy coding schemes, inspired by recent works on mobile visual search [13]. For the first time, the problem of lossless coding of binary descriptors is addressed in this paper. Depending on the available bit budget, the number of pair-wise comparisons can be varied accordingly. In this context, we study the discriminative power of the descriptors as a function of rate, when varying the strategy used to select the most suitable comparisons. The proposed selection methods are general, and can be applied to any binary descriptor. In this paper, we present our results in the case of BRISK and FREAK descriptors. The evaluation methodology takes advantage of an annotated database of image patches extracted from different views of the same scene [3]. In this way, we are able to focus on the design of the descriptor independently from the underlying detector. Our results show that an appropriate selection of the pairs, together with a suitable reordering of them before lossless coding, enables to significantly improve the discriminative power. For example, in the case of BRISK, the false positive rate is reduced from 76% to 58%, when the true positive rate is equal to 95% and 128 pairwise comparisons are used, which are lossless compressed using about 80 bits on average.

The rest of this paper is organized as follows. Section 2 introduces the necessary background and formulates the problem addressed in this paper. Section 3 describes the proposed lossless coding strategy, and Section 4 the different selection strategies that can be used to build the descriptor. Experimental results are reported in Section 5 and Section 6 concludes the paper.

The project GreenEyes acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number:296676.



(a) (b)

Fig. 1. (a) The BRISK sampling pattern with $N = 60$ points and (b) the FREAK sampling pattern with $N = 43$ points. The red crosses denote the position of the sampling points; the blue circles represent the patch of pixels used to smooth the intensity of a particular sampling point.

2. PROBLEM STATEMENT

Let $\mathcal{I}$ denote an image that is processed to extract a set of local features $\mathcal{D}$. First, a scale invariant detector is applied, to identify stable keypoints in the scale-space domain. The number of detected keypoints $M = |\mathcal{D}|$ depends on both the image content and on the type and parameters of the adopted detector. Then, the (oriented) patches around the detected keypoints are further processed to compute the corresponding descriptors. Hence, $d_m \in \mathcal{D}$ is a descriptor, which consists of two components: i) a 4-dimensional vector $\mathbf{c}_m = [x_m, y_m, \sigma_m, \theta_m]^T$, indicating the position (x_m, y_m), the scale σ_m of the detected keypoint, and the orientation angle θ_m of the image patch; ii) a D-dimensional vector $\mathbf{d}_m$, which represents the descriptor associated to the keypoint $\mathbf{c}_m$.

In this paper, we study the class of binary descriptors, i.e., $\mathbf{d}_m \in \{0,1\}^D$. Each descriptor element is a bit, which represents the result of a binary test evaluated based on the content of the patch associated to the keypoint $\mathbf{c}_i$. In particular, we consider two state-of-the-art descriptors, namely BRISK [8] and FREAK [9], which follow a similar construction. Indeed, in both cases, each binary test compares the (smoothed) intensity values of a pair of pixels, whose possible locations within the patch are illustrated, respectively, by the sampling patterns in Figure 1(a) and Figure 1(b). More formally, let $\mathbf{p}_m^i \in \mathbb{R}^2$, $i = 1, \ldots, N$, denote the position of a sampling point defined in a coordinate system centered at (x_m, y_m), rotated with an angle θ_m, and properly scaled according to σ_m. Let $\mathcal{I}(\mathbf{p}_m^i, \rho_i)$ denote an intensity value obtained by averaging the pixel values at locations around $\mathbf{p}_m^i$. Although different averaging filters can be used, the publicly available implementation of BRISK and FREAK adopt a simple box mean filter with floating point boundaries and side length equal to ρ_i. The value of ρ_i depends on the distance from the center of the sampling pattern. In BRISK it is chosen so as to avoid overlap between neighboring sampling points, while in FREAK it allows for sampling points to overlap (so that less points are used).

Consider the set $\mathcal{A}$ of all sampling point pairs

$$\mathcal{A} = \{(\mathbf{p}_m^i, \mathbf{p}_m^j) \in \mathbb{R}^2 \times \mathbb{R}^2 | i < N \wedge j < N \wedge i, j \in \mathbb{N}\}. \quad (1)$$

Given a patch corresponding to the detected keypoint $\mathbf{c}_m$, it is possible to compute up to $N(N-1)/2$ binary comparisons, i.e., one for each pair in $\mathcal{A}$. That is,

$$b = \begin{cases} 1, & \mathcal{I}(\mathbf{p}_m^j, \rho_j) > \mathcal{I}(\mathbf{p}_m^i, \rho_i) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The total number of possible binary tests depends on the configuration of the sampling pattern. The standard BRISK sampling pattern consist of $N = 60$ points, thus $|\mathcal{A}| = 1770$. Instead, the proposed FREAK sampling pattern employs $N = 43$ points, thus $|\mathcal{A}| = 903$.

In this paper, we study different strategies that can be used to select a subset of the pairs, subject to a constraint on the target size of the descriptor D. In addition, we propose to lossless code the descriptor, in such a way that it can be represented with $R \leq D$ bits, on average. Our main contribution consists in the proposal of novel selection strategies that take explicitly into account: i) the discriminative power of each pair, which can be obtained by considering sets of matching and non-matching patches; ii) the cost, in bits, of encoding the descriptor.

3. CODING THE DESCRIPTOR

Regardless of the specific selection strategy, further detailed in Section 4, the descriptor will consist of a string of D bits, each representing the outcome of a binary test. Since the binary tests are not statistically independent, it is possible to model the descriptor as a binary source with memory and perform lossless coding using a number of bits $R \leq D$. Let $H(\pi_n)$, $n = 1, \ldots, D$, denote the entropy of the n-th element of the descriptor, which is computed as

$$H(\pi_n) = -p_n(0) \log_2 p_n(0) - p_n(1) \log_2 p_n(1) \quad (3)$$

In a similar way, it is possible to compute the conditional entropy $H(\pi_{n_1}|\pi_{n_2})$. The statistics used to compute $H(\pi_n)$ and $H(\pi_{n_1}|\pi_{n_2})$ can be obtained by analyzing a training set of descriptors extracted from an image collection. Let $\tilde{\pi}_n$, $n = 1, \ldots, D$, denote a permutation of the D selected pairs, which indicates the sequential order used for encoding the descriptor. The average code length needed to lossless code the descriptor is lower bounded by

$$R = \sum_{n=1}^{D} H(\tilde{\pi}_n | \tilde{\pi}_{n-1}, \ldots, \tilde{\pi}_1) \quad (4)$$

In order to optimize the coding efficiency, it is useful to find the permutation that minimizes the lower bound in (4). In our work, we adopted a greedy strategy, which assumes that the descriptor can be modeled as a Markov source of the first order, i.e., $H(\tilde{\pi}_n|\tilde{\pi}_{n-1}, \ldots, \tilde{\pi}_1) = H(\tilde{\pi}_n|\tilde{\pi}_{n-1})$. Therefore, we propose to reorder the descriptor selecting the elements iteratively. Specifically, the n-th element is chosen as the one that minimizes the conditional entropy with respect to the previously selected element

$$\tilde{\pi}_n = \arg\min_{\pi_n} H(\pi_n|\tilde{\pi}_{n-1}) \quad (5)$$

As for the first element, we opted for selecting the one with the lowest entropy, although we verified that this heuristic choice does not significantly affect the coding rate.

4. BUILDING THE DESCRIPTOR

In this section we consider different selection strategies. We start considering the methods adopted in the reference implementations of BRISK and FREAK, i.e., `brisk` and `freak`, respectively. Then, we describe the proposed strategies, i.e., `matching-based`, `coding-based` and `hybrid`, which are then evaluated in Section 5.

4.1. `brisk`

The construction of the BRISK descriptor proceeds by identifying the subset of short-distance pairs $\mathcal{S}$ and long-distance pairs $\mathcal{L}$:

$$\mathcal{S} = \{(\mathbf{p}_m^i, \mathbf{p}_m^j) \in \mathcal{A} | \|\mathbf{p}_m^j - \mathbf{p}_m^i\| < \delta_{max}\} \quad (6)$$

$$\mathcal{L} = \{(\mathbf{p}_m^i, \mathbf{p}_m^j) \in \mathcal{A} | \|\mathbf{p}_m^j - \mathbf{p}_m^i\| > \delta_{min}\} \quad (7)$$

The long-distance pairs are used in BRISK to estimate the orientation of the patch θ_m. Instead, the descriptor is obtained by concatenating the binary tests corresponding to the short-distance pairs as in (6), such that $(\mathbf{p}_m^i, \mathbf{p}_m^j) \in \mathcal{S}$. Hence, the number of elements D of the descriptors depends on the value of the threshold distance δ_{max}. In [8], δ_{max} was set equal to $13.67\sigma_m$, so as to achieve a descriptor with $D = |\mathcal{S}| = 512$ elements. In our experimental evaluation we will vary δ_{max} to test different sizes of the descriptor.

4.2. `freak`

The FREAK descriptor [9] introduced a heuristic algorithm to select the set of binary tests used to construct the descriptor. During a training phase, FREAK analyses all the $N(N-1)/2$ possible pairs in the set $\mathcal{A}$, for a large number of patches extracted from several images. Let $\mathbf{D} \in \{0,1\}^{N(N-1)/2 \times Q}$ a matrix containing the result of the binary tests for all Q patches. FREAK computes the variance of each pair (row), and selects as first pair the one with the largest variance. This is equivalent to selecting the pair for which the occurrences of zeros and ones are more evenly distributed, i.e., the row with the largest entropy. Then, the other pairs are iteratively selected, by choosing the row that minimizes the correlation with the previously selected one. Hence, the algorithm is greedy in nature and, at each step, tries to select a pair so as to maximize diversity. The algorithm terminates when the budget D is exhausted.

4.3. `matching-based`

The selection strategy adopted in FREAK considers the statistical distribution of binary tests computed on a large number of patches. However, it does not consider how good the selected pairs are when matching descriptors extracted from different images of the same scene. We propose a novel selection strategy, named `matching-based`, that explicitly considers the joint distribution of binary tests computed in the case of matching and non-matching patches. Specifically, we exploit the availability of the dataset introduced in [3], which includes a large set of patches extracted from several images of the same scene and acquired from different viewpoints. In addition, the dataset provides the information about which patches correspond to the same physical keypoint, i.e., the matching patches. As for FREAK, let $\mathbf{D} \in \{0,1\}^{N(N-1)/2 \times Q}$ a matrix containing the result of the binary tests for all Q patches. Let $\mathcal{M}$ denote the set of indexes of matching pairs, i.e.,

$$\mathcal{M} = \{(q_1, q_2) | \, d_{q_1} \text{ and } d_{q_1} \text{ are matching keypoints}\} \quad (8)$$

Similarly, it is possible to define a set $\mathcal{N}$ of indexes of non-matching pairs. Then, for each of the pairs in $\mathcal{A}$, it is possible to compute the mutual information $I^{\mathcal{M}}(\pi_n)$, $n = 1, \ldots, N(N-1)/2$,

$$I^{\mathcal{M}}(\pi_n) = \sum_{x \in 0,1} \sum_{y \in 0,1} p_n^{\mathcal{M}}(x,y) \log_2 \frac{p_n^{\mathcal{M}}(x,y)}{p_n(x) \cdot p_n(y)}, \quad (9)$$

where $p_n(0)$ and $p_n(1)$ are the probabilities of observing, respectively, zero or one, as the output of the binary test involving the n-th pair in $\mathcal{A}$. The probability $p_n^{\mathcal{M}}(x,y)$ measures the joint occurrences of zeros and ones in pairs of matching descriptors. For example, $p_n^{\mathcal{M}}(0,0)$ is the probability that both descriptors in a matching pair contain a zero in the element corresponding to the n-th binary test. Similarly, it is possible to define the mutual information for non-matching pairs, i.e., $I_n^{\mathcal{N}}$. For each pair in $\mathcal{A}$, we compute the following scoring function

$$J_n = I^{\mathcal{M}}(\pi_n) - I^{\mathcal{N}}(\pi_n), \quad (10)$$

and we rank pairs in decreasing order of J_n. The proposed selection strategy chooses the top-D pairs with the highest value of J_n.

4.4. `coding-based`

The `coding-based` selection strategy proceeds by building a descriptor of D elements, with the goal of minimizing the number of bits necessary for coding it. The selection strategy follows the same approach already described in Section 3 but, at each iteration of the greedy algorithm, is considers all the possible $N(N-1)/2$ pairs rather than a subset of D pairs. The algorithm terminates when D pairs are selected.

4.5. `hybrid`

The `hybrid` approach combines the `matching-based` with the `coding-based` approach. The D elements of the descriptors are selected by means of the following greedy strategy

$$\tilde{\pi}_n = \arg\max_{\pi_n} \alpha(I^{\mathcal{M}}(\pi_n) - I^{\mathcal{N}}(\pi_n)) - (1-\alpha) H(\pi_n | \tilde{\pi}_{n-1}) \quad (11)$$

As for the first pair $\tilde{\pi}_1$, the term $H(\pi_n|\tilde{\pi}_{n-1})$ is replaced by $H(\pi_n)$. The parameter α enables to trade-off the goodness of matching with the cost of coding the descriptor.

5. EXPERIMENTS

To compare the performance of the proposed methods we adopted the data set described in [3], which contains patches extracted from images representing different views of the same scene. The data set is divided in three sets of patches, each corresponding to a scene, namely, *Yosemite*, *Liberty* and *Notredame*. In our experiments, one was used for training, and a different one for testing. Each set contains approximately 250k patches. From each set, we extracted two matrices $\mathbf{D} \in \{0,1\}^{N(N-1)/2 \times Q}$, using, respectively, the BRISK and FREAK patterns shown in Figure 1. In addition, the data set provides the ground truth regarding the pairs of matching patches. We created an equal number of non-matching pairs by sampling pairs at random. We considered different target lengths of the descriptors in the set $D \in \{16, 32, 64, 128, 256, 512\}$. For each value of D, we used the descriptors extracted from the training set to learn the selected pairs using one of the methods illustrated in Section 4. In our results, we adopt the notation PATTERN-`selection` to denote both the sampling pattern used (i.e., BRISK or FREAK) and the selection method. For the hybrid scheme we set $\alpha = 0.75$.

Results were evaluated on descriptors extracted from a test set. For each pair of descriptors, we computed the Hamming distance between them. The descriptors were considered to be matching if the Hamming distance was below a given threshold. Due to the availability of the ground truth, we were able to trace the Receiver Operating Characteristic (ROC) curve, as illustrated in Figure 2 by varying such a threshold. The true positive rate indicates the fraction of matching descriptors that were correctly identified to be so. The false

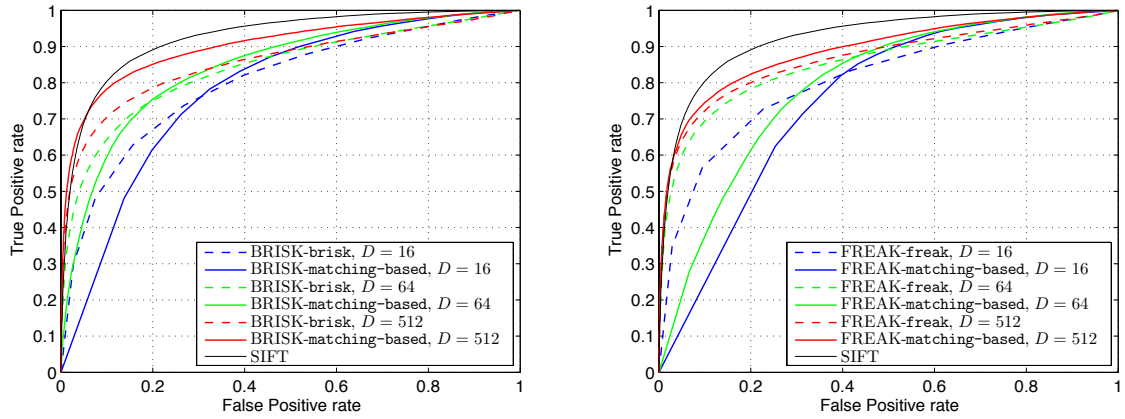


Fig. 2. ROC curves for a) BRISK and b) FREAK. We show the original implementations at different bitrates versus the `matching-based` method. The black curve shows the performance of SIFT on the same dataset.

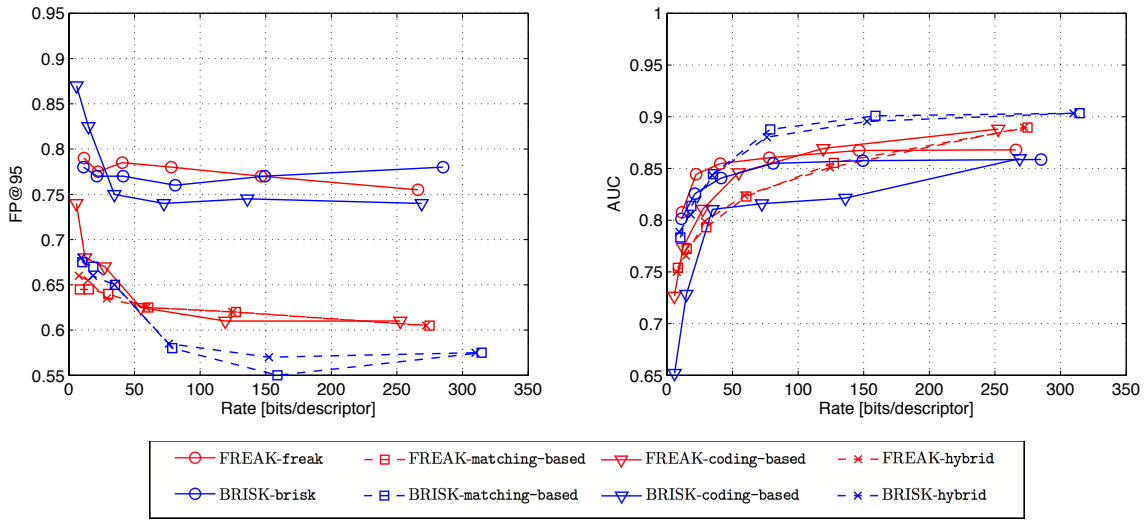


Fig. 3. a) False positive rate at true positive rate equal to 95%. b) Area under the curve.

positive rate indicates the fraction of non-matching descriptors that were considered to be matching. In Figure 2 training is performed on *Liberty* and testing on *Notredame*. We observe that, for a given descriptor, e.g., BRISK-`brisk` and BRISK-`matching-based`, the discriminative power decreases when decreasing D and, consequently, the rate R needed to encode the descriptor. Overall, BRISK-`matching-based` tends to outperform BRISK-`brisk` for low-positive rates and large descriptor lengths. In order to summarize the ROC curve in a single value, we considered two options: i) the value of false positive rate when the true positive rate is equal to 95% (FP@95), as in [3][10]; ii) the area under the curve (AUC). Figure 3(a) and 3(b) show the discriminative power of the different methods using the same training and test sets as for Figure 2. Similar results were obtained for different combinations of training and testing, but were omitted due to space limitations. By fixing the sampling pattern, we observed that the `matching-based` and `hybrid` selection methods performed best, with an improvement as high as 20% in terms of FP@95 and 5% in terms of AUC. Between the sampling patterns, BRISK obtained better results at high bitrates, while at low bitrates (i.e., less than 50 bits/descriptor) FREAK was preferable. Note that the coding scheme described in Section 3 enabled to significantly reduce the bit budget. For example, the original BRISK descriptor with $\{512, 256, 128, 64\}$ bits was encoded using, respectively, approximately $\{285, 150, 80, 40\}$ bits on average.

6. CONCLUSIONS

In this paper we studied lossless coding for binary descriptors and its implication on their discriminative power. We presented an entropy coding scheme that operates on binary descriptors so as to minimize the number of bits necessary to represent them. We also presented methods to select only those descriptor elements which maximize the discriminative power. Experiments on two state-of-the-art binary descriptors show that substantial improvements can be achieved. In our future investigations we plan to study the effects of the proposed methods on a complete image retrieval pipeline, as well as using the proposed schemes for compressing binarized SIFT descriptors [14].

7. REFERENCES

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

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[3] Matthew Brown, Gang Hua, and Simon A. J. Winder, "Discriminative learning of local image descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 43–57, 2011.

[4] Edward Rosten, Reid Porter, and Tom Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 105–119, 2010.

[5] Elmar Mair, Gregory D. Hager, Darius Burschka, Michael Suppa, and Gerd Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," in *ECCV (2)*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios, Eds. 2010, vol. 6312 of *Lecture Notes in Computer Science*, pp. 183–196, Springer.

[6] Motilal Agrawal, Kurt Konolige, and Morten Rufus Blas, "Censure: Center surround extremas for realtime feature detection and matching," in *ECCV (4)*, David A. Forsyth, Philip H. S. Torr, and Andrew Zisserman, Eds. 2008, vol. 5305 of *Lecture Notes in Computer Science*, pp. 102–115, Springer.

[7] Michael Calonder, Vincent Lepetit, Mustafa Özuysal, Tomasz Trzcinski, Christoph Strecha, and Pascal Fua, "Brief: Computing a local binary descriptor very fast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, 2012.

[8] Stefan Leutenegger, Margarita Chli, and Roland Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *ICCV*, Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc J. Van Gool, Eds. 2011, pp. 2548–2555, IEEE.

[9] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst, "Freak: Fast retina keypoint," in *CVPR*. 2012, pp. 510–517, IEEE.

[10] Tomasz Trzcinski and Vincent Lepetit, "Efficient discriminative projections for compact binary descriptors," in *ECCV (1)*, Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, Eds. 2012, vol. 7572 of *Lecture Notes in Computer Science*, pp. 228–242, Springer.

[11] A. Redondi, M. Cesana, and M. Tagliasacchi, "Rate-accuracy optimization in visual wireless sensor networks," in *International Conference on Image Processing*, oct. 2012, pp. 124 – 129.

[12] A. Redondi, M. Cesana, and M. Tagliasacchi, "Low bitrate coding schemes for local image descriptors," in *International Workshop on Multimedia Signal Processing*, sept. 2012, pp. 124 –129.

[13] B. Girod, V. Chandrasekhar, D.M. Chen, Ngai-Man Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S.S. Tsai, and R. Vedantham, "Mobile visual search," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 61–76, july 2011.

[14] Christoph Strecha, Alexander A. Bronstein, Michael M. Bronstein, and Pascal Fua, "Ldahash: Improved matching with smaller descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 66–78, 2012.