A Study on Low-Cost Representations for Image Feature Extraction on Mobile Devices

Ramon F. Pessoa, William R. Schwartz, and Jefersson A. dos Santos

Department of Computer Science, Universidade Federal de Minas Gerais (UFMG),
Belo Horizonte, Minas Gerais, Brazil, 31270-901
{ramon.pessoa,william,jefersson}@dcc.ufmg.br

Abstract. Due the limited battery life and wireless network bandwidth limitations, compact and fast (but also accurate) representations of image features are important for multimedia applications running on mobile devices. The main purpose of this work is to analyze the behavior of techniques for image feature extraction on mobile devices by considering the *triple trade-off problem* regarding effectiveness, efficiency, and compactness. We perform an extensive comparative study of state-of-the-art binary descriptors with bag of visual words. We employ a dense sampling strategy to select points for low-level feature extraction and implement four bag of visual words representations which use hard or soft assignments and two most commonly used pooling strategies: average and maximum. These mid-level representations are analyzed with and without lossless and lossy compression techniques. Experimental evaluation point out ORB and BRIEF descriptors with soft assignment and maximum pooling as the best representation in terms of effectiveness, efficiency, and compactness.

1 Introduction

In 2014, the number of smartphone users worldwide achieved around 1.75 billion. Recent forecasts indicate that the growth of smart mobile devices usage should increase even more the next years [1]. Many challenges and opportunities have emerged concerning image/video processing tasks, such as annotation, categorization, detection and retrieval. The challenges in image processing in mobile devices include constraints such as memory and computing resources which may be very limited [2]. Due the limited battery life of mobile devices, energy usage is also a critical issue [2]. Regarding feature extraction from images, those constraints configure a *trade-off* among effectiveness, efficiency and compactness.

In this work, we deal with the *feature extraction triple trade-off problem* in mobile devices by evaluating low-cost feature representations. We concentrate our efforts in three main fronts: (1) binary low-level descriptor selection; (2) mid-level representations; and (3) feasibility analysis of data compression techniques. Binary descriptors are interesting options because they provide effective and compact representation [3]. Mid-level representations based on Bag of Visual Words (BoVW, or just BoW) are also good alternative since they provide suitable representation for the amount of local features extracted. Finally, we analyze the use of well-known compression algorithms to reduce as most as possible the size of the final feature representation. To study low-cost representations for image feature extraction on mobile devices, we have adopted a content-based image retrieval (CBIR) application protocol.

Content-Based Image Retrieval (CBIR) applications on mobile devices have been typically modeled using a client-server architecture [2]. Girod et al. [2] present a low latency interactive visual search system. They use interest point detection, compressed histogram of gradients (CHoG) descriptor and a mid-level representation. However, due to the complexity of spatial sub-block assignment scheme, the extraction of the CHOG is not fast enough. In addition, the quality of features is also influenced by the detection of interest points, which does not receive much attention by the CHOG. In [3], lossless compression of binary image features is proposed to be used in a mobile CBIR enviroment. The coding solution exploits the redundancy between descriptors of an image by sorting the descriptors and applying differential pulse coded modulation (DPCM) and arithmetic coding. They do not use mid-level representation, just apply lossless compression on binary features before sending them to a server side. Each binary descriptor is computed from a patch around a detected keypoint. They propose a lossless predictive coding scheme for binary features.

Our work differs from the literature in several aspects. First, to the best of our knowledge, there is no works in the literature that evaluate (or use) low-cost mid-level representation based on dense sampling in mobile applications and it has been shown that dense sampling is more accurate than interest point detection to compute bag-of-visual-word features [4]. Second, there is no work that evaluates the state-of-the-art binary descriptor called BinBoost [6] for image feature extraction on mobile devices. Finally, we evaluate many compression techniques and different assignment/pooling strategies to obtain compact representations.

2 Evaluation Methodology

We aim at evaluating binary low-level descriptors in different mid-level representations to find the most suitable setups in terms of effectiveness, efficiency, and compactness. As mentioned earlier, we have adopted a CBIR process which is composed by offline and online phases.

In the *Offline phase*, after extracting local feature vectors from the image dataset, the feature space is quantized and each region corresponds to a visual word. We use the codebook to create bag-of-word representations for all images in the database. Whereas, in the *Online phase*, given a query image, its local feature vectors are computed and then assigned to the visual words in the dictionary. Finally, the local assignment vectors are summarized by a pooling strategy, which creates the bag-of-visual-words representation. A compression step may be processed to reduce the feature vector size.

In the similarity search, a distance function (Euclidean) is used as similarity measure to rank the database images.

2.1 Low-level Feature Extraction

Five well-known binary descriptors are used to encode low-level local properties: (1) BRIEF [7]; (2) ORB [8]; (3) BRISK [9]; (4) FREAK descriptor [10]; and (5) BinBoost descriptor [6]. We used a dense sampling (6 pixels, as in [5]) strategy to select points for low-level feature extraction.

2.2 Mid-Level Representation

We evaluated two codeword assignment strategies with two different pooling approaches (average and maximum). To summarize, we use *Hard assignment* where the local feature descriptors of the image are matched with visual words of the vocabulary (the nearest one). A histogram of the visual descriptors is populated by the corresponding bins. We also use *Soft assignment*. In this case, instead of assigning a descriptor to a single corresponding visual word, we assign it to k bins in a soft manner. More specifically, for every descriptor, we add a quantity q to the bins of the k top nearest visual words. This quantity q is the Gaussian kernel (Radial Basis Function) distance of the descriptor and the visual word.

2.3 Data Compression

Data compression is classified into two categories: lossless and lossy.

In this paper, we have used two approaches of lossless compression: Huffman Encoding and Error Enconding + Huffman. The first one, *(Huffman encoding)* is based on frequency of occurrence for each possible value of common symbols which are generally represented using fewer bits than less common symbols. We use the variation of Huffman called byte-oriented Huffman Code [11] where a sequence of bytes is assigned to bin values of a BoW representation. In the second approach, we use *Error Enconding* witch is the difference between bin values. The first bin is the maximum error and have their own value. In this paper, we applied this technique to try get repeated differences of bin values and get a better compression using Huffman encoding.

Two approaches of lossy compression were tested in the Bag of Words using Soft-Assignment with Maximum Pooling approach (BoW Soft-MAX). In *Soft-MAX Truncated*, instead of sending float values, all numbers expressed in floating-point were truncated and transformed in integer values. As a result, we use the approach *Soft-MAX using Ranges*. In this lossy compression, we created fixed features vectors with values in ranges of 5 or 10 or 15 or 20 or 25 or 30. We choose values after looking for the minimum and maximum in all values of the features of Soft-MAX. The idea is activate a bin if the value of the Soft-MAX Truncated is next to the central point in the range.

3 Experimental Setup

The experiments were conducted in two public available image datasets: Caltech-101 [12] and The PASCAL VOC 2007 [13]. Caltech-101 dataset contains 9,145 images. The complete dataset size is 138.6 MB. The PASCAL VOC 2007 dataset consists of 9,963 images. It has multiple objects per image. The complete PASCAL VOC 2007 size is 875.5 MB.

We used the $P@10$ metric (a well known and widely used metric of information retrieval) to evaluate *effectiveness*. This measure is called precision at N or $P@N$. The *efficiency* was evaluated by computing the feature extraction and representation time, in seconds. Finally, we have used the representation size (in bytes) and the Compression Ratio (CR) as measures for evaluating the *compactness*. In this aspect, our baseline is the size of all images in the datasets. Thus, we always divide images size per image representation size. If CR is high, the compression ratio is better because the resulting image representation is smaller.

4 Results and Discussion

In this section, we present the experimental results and the discussion. Our analysis is organized in three main parts. Section 4.1 presents the effectiveness evaluation, Section 4.2 presents the efficiency evaluation of each representation approach and Section 4.3 evaluates different compression techniques.

4.1 Effectiveness Evaluation

Table 1 (a) and (b) show the P@10 for each binary descriptor with four different BoW-based mid-level representations. Regarding the Caltech 101 dataset, the Soft-Assignment with Max Pooling (Soft-MAX) achieved the best results for all tested descriptors. For the VOC Pascal 2007 dataset, Soft-MAX achieved high P@10 values for BinBoost, BRIEF and ORB descriptors. Soft-AVG and Hard-AVG mid-level representation also yield very high results for some descriptors.

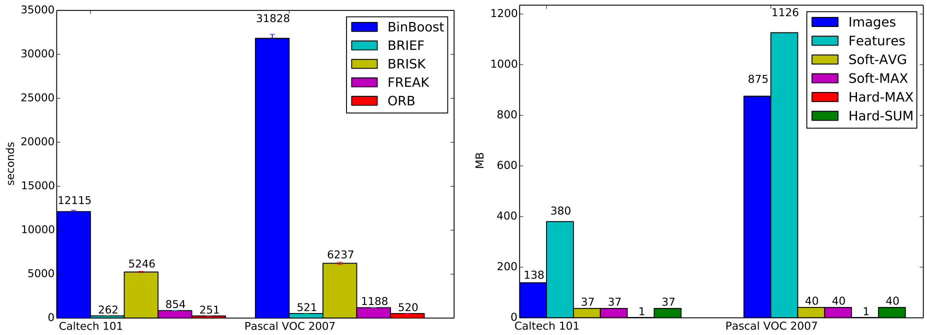
Although, in the Caltech 101 datase, FREAK descriptor with Hard-MAX has similar P@10 as in FREAK with Soft-AVG or Soft-MAX, all mid-level representations with FREAK descriptor yields low precision (P@10) comparing with the other binary descriptors.

Table 1: P@10 for each descriptor with different mid-level representations (H = Hard, S = Soft, BB = BinBoost, BF = BRIEF, BK = BRISK, FK = FREAK, OB = ORB). P@10 reported with a confiance of 95%.

	(a) Caltech 101				(b) Pascal VOC 2007			
	S-AVG	**S-MAX**	**H-MAX**	**H-AVG**	**S-AVG**	**S-MAX**	**H-MAX**	**H-AVG**
BB	15.8 ± 0.3	**24.4 ± 0.4**	17.6 ± 0.3	**23.0 ± 0.4**	40.2 ± 0.2	**44.9 ± 0.2**	42.4 ± 0.2	43.6 ± 0.2
BF	16.0 ± 0.3	**26.3 ± 0.4**	18.2 ± 0.3	18.6 ± 0.3	39.8 ± 0.2	**44.6 ± 0.2**	39.6 ± 0.2	**44.6 ± 0.2**
BK	19.7 ± 0.3	**26.7 ± 0.4**	18.1 ± 0.3	20.0 ± 0.3	**43.3 ± 0.2**	42.0 ± 0.2	39.0 ± 0.2	42.2 ± 0.2
FK	**18.5 ± 0.3**	**18.5 ± 0.3**	**18.5 ± 0.3**	17.2 ± 0.3	**41.1 ± 0.2**	37.7 ± 0.2	37.9 ± 0.2	**41.1 ± 0.2**
OB	14.1 ± 0.2	**26.2 ± 0.4**	17.1 ± 0.3	17.7 ± 0.3	39.7 ± 0.9	**45.7 ± 0.2**	38.8 ± 0.2	42.3 ± 0.2

4.2 Efficiency Evaluation

Figure 1(a) presents the computational time required for each descriptor to extract features for all images in Caltech 101 and Pascal VOC 2007 combined. According to the results, while BinBoost and BRISK are the most expensive descriptors, BRIEF, FREAK and ORB can be considered good choices since they are very fast. For this experiment, we extracted each descriptor five times per image and the results are reported with a confidence of 95% (α=0.05).



(a)

(b)

Fig. 1: (a) Time (in seconds) spent for feature extraction of all images of Caltech 101 and VOC 2007 using different descriptors. (b) Size (MB) of the Images, Features and ORB' Mid-Level Representations in the datasets Caltech 101 and Pascal VOC 2007.

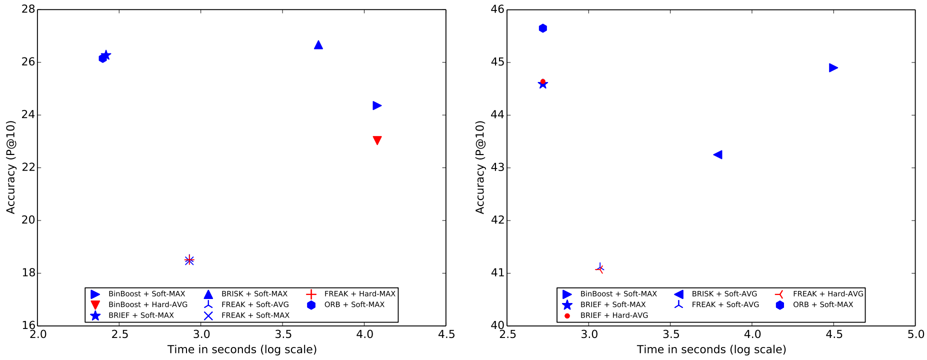
To evaluate each descriptor in both effectiveness and efficiency aspects, we present the scatter plot in Figure 2, which shows the relation between P@10 and time in seconds (in log scale) for the best descriptors in the Caltech 101 (Figure 2(a)) and VOC 2007 (Figure 2(b)) datasets. In this scenario, the most suitable representations are "BRIEF + Soft-MAX" and "ORB + Soft-MAX" for Caltech 101 and "ORB + Soft-MAX", "BRIEF + Soft-MAX", and "BRIEF + Hard-AVG" for VOC 2007.

4.3 Compactness Evaluation

In this section, we evaluate the feature representation compactness aiming at finding the most suitable descriptors concerning their feature vector size. As it can be seen in Figure 1(b), it is better to transfer mid-level representation to be processed in the server side instead of images or low-level features because mid-level representation are more compact.

Lossless Compression: Figures 3(a) and 3(b) present the relation between P@10 and Compression Ratio (CR) for the most suitable feature representations in the Caltech 101 and Pascal VOC 2007 datasets, respectively.

According to the results "BRIEF + Soft-MAX" and "ORB + Soft-MAX" can be considered the most suitable approaches for the Caltech 101 dataset. For

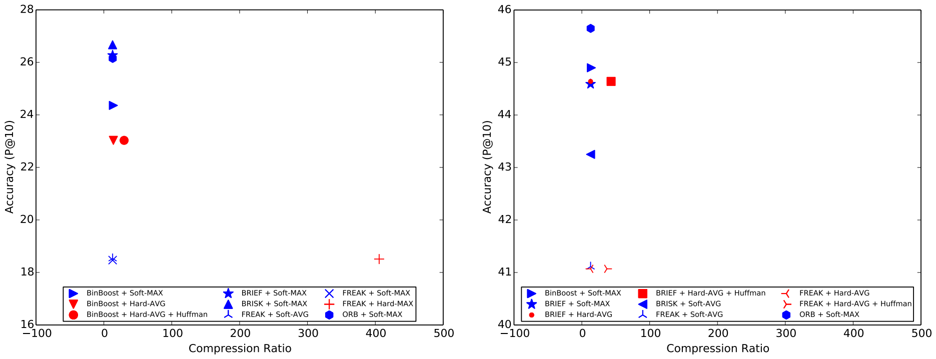


(a) Caltech 101 dataset

(b) Pascal VOC 2007 dataset

Fig. 2: Relation of Accuracy (P@10) versus Time in seconds (log scale) for the most accurate descriptors (See Table 1).

the Pascal VOC 2007 dataset, the best ones are "ORB + Soft-MAX", "BRIEF + Soft-MAX", "BRIEF + Hard-AVG" and "BRIEF + Hard-AVG + Huffman". BinBoost and BRISK are time consuming and have been discarded.



(a) Caltech 101 dataset

(b) Pascal VOC 2007 dataset

Fig. 3: Relation between P@10 and Compression Ratio (CR) for the most suitable feature representations in Table 1 and/or using lossless compression.

Lossy Compression: Figure 4(a) and 4(b) present the relation between P@10 and Compression Ratio (CR) for the Lossy compression of Soft-MAX representation in the Caltech 101 and Pascal VOC 2007 datasets, respectively.

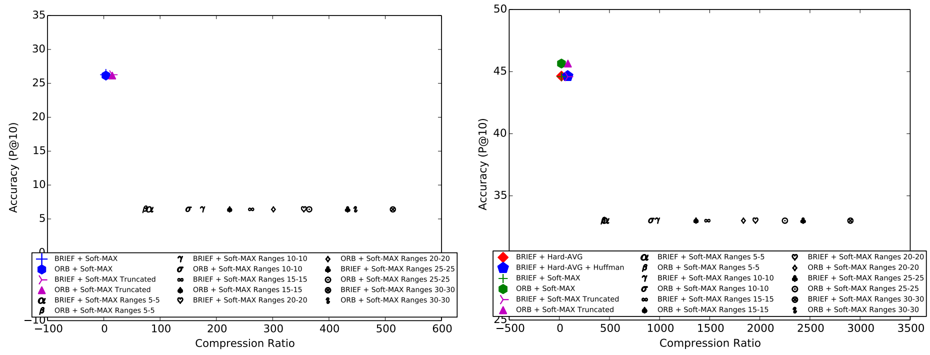
In the Caltech 101 dataset, we have included the best Soft-MAX representation to compare its performance with the compact ones. In the Pascal VOC 2007 dataset, we also have included the "BRIEF + Hard-AVG" and "BRIEF + Hard-AVG + Huffman". In both datasets, "ORB + Soft-MAX" and "BRIEF + Soft-MAX" have better P@10 compared with the lossy compression of Soft-MAX representations.

It is worth to observe that Soft-MAX Truncated achieves similar P@10 compared with raw Soft-MAX, but it is more compact (transformation of interger to float values). For example, the precision (P@10) of "ORB + Soft-MAX" is 45,6471% (≈ 45.65%) with CR = 21.41 and the "ORB Soft-MAX Truncated" is 45.6492% (≈ 45.65%) with CR = 85.83. In this case, the highest CR values were observed with the "ORB Soft-MAX Truncated" approach, which are more compact (See Table 2).

Even though the compression approaches that use ranges (lossy compression) are extremely compact, they produce low precision rates, which invalidates their use.

Table 2: Compression Ratio (CR) of "BRIEF + Soft-MAX", "BRIEF + Soft-MAX Truncated", "ORB + Soft-MAX" and "ORB + Soft-MAX Truncated" in the datasets Caltech 101 and Pascal VOC 2007.

	Caltech 101	**Pascal VOC 2007**
BRIEF + Soft-MAX	3.69	21.3
BRIEF + Soft-MAX Truncated	14.9	85
ORB + Soft-MAX	3.75	21.41
ORB + Soft-MAX Truncated	15.07	85.83



(a) Caltech 101 dataset

(b) Pascal VOC 2007 dataset

Fig. 4: Relation between P@10 and Compression Ratio (CR) for the lossy compression of Soft-MAX representation. The best representations in Figure 3 have been included only for comparison.

5 Conclusions and Future Work

In this paper, we conducted extensive evaluation of low-cost mid-level representation approaches by exploiting binary local descriptors in the context of feature extraction on mobile devices. The experimental results pointed out that the most suitable representations in terms of effectiveness, efficiency, and compactness are ORB and BRIEF descriptors with Soft assignment and MAX pooling. In addition, even though the BinBoost is an accurate descriptor, it produces a larger feature vector and together with BRISK descriptor spends much more time to extract features. Another good alternative to have an acceptable accuracy rate gaining a better compression ratio is to use the descriptors ORB or BRIEF with Soft-MAX Truncated instead of Soft-MAX. As future work, we intend to investigate the use of algorithms for detection of interest points. We also plan to use more datasets and more mid-level representations and test the impact of using different distance metrics on the final CBIR ranking.

6 Acknowledgments

This work was partially financed by Brazilian National Research Council – CNPq (Grant 477457/2013-4 and 449638/2014-6), Minas Gerais Research Foundation – FAPEMIG (Grants APQ-01806-13, APQ-00567-14 and APQ-00768-14), and Coordination for the Improvement of Higher Education Personnel – CAPES.

References

1. Cisco, ”Cisco visual networking index: Global mobile data traffic forecast update, 20142019,” Tech. Rep., 2015.
2. B. Girod, V. Chandrasekhar, D. M. Chen, N.-M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham, ”Mobile visual search,” Signal Processing Magazine, vol. 28, no. 4, pp. 6176, 2011.
3. J. Ascenso and F. Pereira, ”Lossless compression of binary image descriptors for visual sensor networks,” in DSP, 2013, pp. 18.
4. K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, ”The devil is in the details: an evaluation of recent feature encoding methods,” 2011.
5. O. Penatti, F. B. Silva, E. Valle, V. Gouet-Brunet, and R. da S. Torres, ”Visual word spatial arrangement for image retrieval and classification,” Pattern Recognition, vol. 47, no. 2, pp. 705 720, 2014.
6. T. Trzcinski, M. Christoudias, P. Fua, and V. Lepetit, ”Boosting binary keypoint descriptors,” in CVPR, 2013, pp. 28742881.
7. M. Calonder, V. Lepetit, C. Strecha, and P. Fua, ”Brief: Binary robust independent elementary features,” 2010, pp. 778792.
8. E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, ”Orb: an efficient alternative to sift or surf,” in ICCV, 2011, pp. 25642571.
9. S. Leutenegger, M. Chli, and R. Y. Siegwart, ”Brisk: Binary robust invariant scalable keypoints,” in ICCV, 2011, pp. 25482555.
10. A. Alahi, R. Ortiz, and P. Vandergheynst, ”Freak: Fast retina keypoint,” in CVPR, 2012, pp. 510517.
11. E. Silva de Moura, G. Navarro, N. Ziviani, and R. Baeza-Yates, ”Fast and flexible word searching on compressed text,” ACM Transactions on Information Systems, vol. 18, no. 2, pp. 113139, 2000.
12. L. Fei-Fei, R. Fergus, and P. Perona, ”Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories,” Computer Vision and Image Understanding, vol. 106, no. 1, pp. 5970, 2007.
13. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, ”The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results,” http://www.pascalnetwork.org/challenges/VOC/voc2007/workshop/index.html.