Representing Local Binary Descriptors with BossaNova for Visual Recognition

Carlos Caetano[†], Sandra Avila[†], Silvio Guimarães[‡], Arnaldo de A. Araújo[†]
[†]Federal University of Minas Gerais, NPDI Lab — DCC/UFMG, Minas Gerais, Brazil
[‡]Pontifical Catholic University of Minas Gerais, VIPLAB — ICEI/PUC Minas, Minas Gerais, Brazil
{carlos.caetano, sandra}@dcc.ufmg.br, sjamil@pucminas.br, arnaldo@dcc.ufmg.br

ABSTRACT

Binary descriptors have recently become very popular in visual recognition tasks. This popularity is largely due to their low complexity and for presenting similar performances when compared to non binary descriptors, like SIFT. In literature, many researchers have applied binary descriptors in conjunction with mid-level representations (*e.g.*, Bag-of-Words). However, despite these works have demonstrated promising results, their main problems are due to use of a simple mid-level representation and the use of binary descriptors in which rotation and scale invariance are missing. In order to address those problems, we propose to evaluate state-of-the-art binary descriptors, namely BRIEF, ORB, BRISK and FREAK, in a recent mid-level representation, namely BossaNova, which enriches the Bag-of-Words model, while preserving the binary descriptor information. Our experiments carried out in the challenging PASCAL VOC 2007 dataset revealed outstanding performances. Also, our approach shows good results in the challenging real-world application of pornography detection.

Categories and Subject Descriptors

I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement—*Feature representation*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object recognition*

General Terms

Algorithms, Experimentation, Performance

Keywords

Visual recognition, local binary descriptors, feature extraction, mid-level representation, BossaNova representation

1. INTRODUCTION

The typical visual recognition pipeline is composed of the three steps: (i) extraction of local image descriptors; (ii) encoding the local features in a mid-level representation; and (iii) classification of the image descriptor. Usually, the image descriptors must be invariant to object translation, rotation, illumination, scale, among others. To cope with these properties, the most common local descriptors are SIFT [17] and SURF [4]. However, they are represented by high-dimensional real-valued vectors, which cause some performance problems in the encoding of the descriptors into a mid-level representation.

Regarding mid-level image representations, the Bag-of-Words (BoW) [23] is the most common approach for encoding the image descriptors. BoW models can be understood as the application of two critical steps [5]: coding and pooling. The coding step quantizes the image local features according to a codebook[1]. The pooling step summarizes the codes obtained into a single feature vector. In the classical BoW, the coding step associates the image local descriptors to the closest element in the codebook, and the pooling takes the average of those codes over the entire image.

Despite the fact that local image descriptors present good accuracy when used by classical BoW approaches, they have a high computational cost in computing the feature vectors, making it impossible extremely hard to use them in some real time applications.

To deal with these issues, the state-of-the-art takes into account binary descriptors, such as BRIEF [7] and ORB [21], instead of SIFT and SURF. According to [8], the methods for extracting these binary descriptors are faster than the methods for computing the SIFT. Moreover, thanks to their representation (*i.e.*, sequence of zeros and ones), the distance between two descriptors can be calculated by the Hamming distance instead of Euclidean distance, which is used for SIFT descriptor.

As reported by [10, 12, 13, 14, 24, 25], the use of binary descriptors in conjunction with mid-level representations is promising. This combination was first introduced by [12], using BRIEF as binary descriptor. The BRIEF descriptor plus FAST keypoint detection [20] are used to extract local invariant features for visual place recognition using the BoW model. They build a vocabulary tree that discretizes a binary descriptor space and use the tree to speed up correspondences for geometrical verification. Lately, in [13], the same authors enhanced the direct index technique and extended the experimental evaluation of their approach. Their main limitation is the use of features in which the rotation

[1]The codebook (or dictionary) is usually built by clustering a set of local descriptors. It can be defined by the set of visual words, corresponding to the centroids of clusters.

SAC'14 March 24-28, 2014, Gyeongju, Korea.

and scale invariance are missing.

Zhang *et al.* [25] introduced an approach to use local binary descriptors for the visual object categorization task. They proposed a new encoding method to address the high dimensionality issue of the traditional binary bitstring encoding. A comparison between Hamming and Euclidean distances was made, proving the benefits of using Hamming distance with binary descriptors. The proposed approach was validated by applying LBP features [18], however, the original LBP algorithm presents some limitations such as noise sensitivity and also, the rotational invariance is missing. Grana *et al.* [14], combined the ORB descriptor and a BoW model for image classification. To deal with the binary string nature of the ORB descriptors, the authors suggested a variation of k-means, called k-majority, replacing Euclidean distance by Hamming distance and majority selected vector as the new cluster center. However, the ORB descriptor suffers from partial scale-invariance.

Whiten *et al.* [24] presented an extension of the FREAK descriptor in conjunction with the BoW model for action recognition. The first bytes of the descriptor encode the appearance and some implicit motion and the remaining bytes strengthen the motion model by building a binary string through local motion patterns. According to Whiten *et al.*, throughout the construction of this descriptor, emphasis is placed on ensuring the entire descriptor remains binary, gifting it with highly optimized processing and feature matching. The authors yield significant computational gains in approaches such as standard BoW models, where thousands of matches must often be made at each frame.

On a mobile environment, Chatzilari *et al.* [10] attempted to examine the visual recognition by simultaneously evaluating the performance and the computational cost of state-of-the-art keypoint detection, feature extraction and encoding algorithms. They seek to balance the system so as to select a configuration that is able to run in an acceptable time frame and, at the same time, can provide satisfactory results for the specific application.

In the face of the good results for the combination of binary descriptors with a classical BoW for visual recognition, we propose in this paper to replace the classical BoW by a recent mid-level representation, named BossaNova [1], which enriches the BoW representation by keeping a histogram of distances between the descriptors found in the image and each codebook element. As shown in the experiments, our approach outperforms previous methods that apply binary descriptors on PASCAL VOC 2007 dataset [11]. Moreover, we explore our method in the challenging real-world application of pornography detection. We evaluate our approach on Pornography dataset [1]. Our result is comparable to the best one published, which is obtained by using HueSIFT descriptors (SIFT + color information) and BossaNova representation.

The remainder of this paper is organized as follows. In Section 2, we survey some binary descriptors and their properties. In Section 3, we provide a brief description of the BossaNova image representation. In Section 4, we introduce our approach for visual recognition. In Section 5, we analyze our experimental results on two challenging datasets. Finally, in Section 6, we present our concluding remarks and discuss future work directions.

2. BINARY DESCRIPTORS

Binary descriptors have received considerable attention having a similar or better recognition performance when compared the BRIEF descriptor [7] to the SURF descriptor [4], for example. The main feature of this kind of descriptor is the time required for extracting it. Even being faster than non binary descriptors, their recognition performance may be comparable to more complex descriptors. In this section, we explore some binary descriptors and their properties.

2.1 BRIEF

The BRIEF descriptor [7] (Binary Robust Independent Elementary Features) describes features using simple binary tests among pixels from a smoothed image (*e.g.*, using a Gaussian kernel with variance equal to 2 and size equal to 9×9 pixels). By itself BRIEF is neither scale nor rotation invariant. Nevertheless, its performance is similar to a more complex local descriptor, the SURF, when compared to its robustness to illumination, blur, and perspective distortion.

The BRIEF descriptor is represented by a binary string in which each bit represents a simple comparison between two elements inside a patch. The keypoint is the center of this patch. The bit is set to '1' if the first point is more *intense* than the other one, otherwise it is set to '0'. Despite the several ways, presented in [7], to perform the selection of points that will be compared, the most common strategy for choosing these points is based on a randomly way according to a Gaussian distribution with respect to the keypoint of the patch. An important observation is that the number of selected points leads to the descriptor size (*e.g.*, 128, 256 and 512).

2.2 ORB

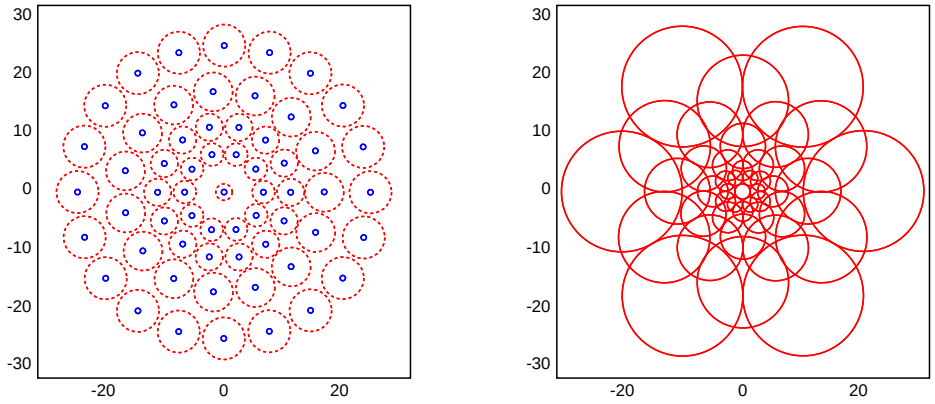
The ORB descriptor [21] (Oriented FAST and Rotated BRIEF) can be considered as an alternative for SIFT and SURF being two times faster than SIFT and one time faster than SURF. The ORB descriptor is robust to noise and invariant to rotation, solving the invariance problem of BRIEF. Despite this improvement, the ORB descriptor is partially invariant to scale.

According to [8], the invariance to rotation is done by estimating the patch rotation using the intensity centroid. Patch moments are used to compute the intensity centroid and outperform gradient-based approaches.

The sampling pattern is steered estimating the orientation, and usual binary tests are used for computing the descriptor. Furthermore, for selecting a couple of points, a k-nearest neighborhood strategy based on error-prone is done. The random sampling has been replaced to a sampling scheme that uses machine learning for de-correlating BRIEF features under rotational invariance. Unlike BRIEF, ORB's descriptor size is fixed to a 256 bitstring.

2.3 BRISK

The BRISK descriptor [16] (Binary Robust Invariant Scalable Keypoints) provides scale and rotation invariance, however it is very sensitive to light intensity variations. As illustrated in Figure 1(a), the BRISK descriptor computes a weighted Gaussian average over a selected pattern of points that are close to the keypoint. For comparing the points, Gaussian windows are used to set the bit to '1' or '0'. Due to its size, the BRISK descriptor is represented by a 512 bits, which is more greater that BRIEF and ORB, and consequently, more computation and storage are required.



(a) BRISK [16] (b) FREAK [19]

Figure 1: Sampling patterns of two local binary descriptor. In (a) it is illustrated the sampling pattern of BRISK descriptor which is based on 60 points: the small blue circles denote the sampling locations; the bigger red dashed circles are drawn at a radius, which corresponds to the standard deviation of the Gaussian kernel used to smooth the intensity values at the sampling points. In (b) it is illustrated the sampling pattern of FREAK descriptor in which each circle represents a receptive field where the image is smoothed with its corresponding Gaussian kernel.

2.4 FREAK

The FREAK descriptor [19] (Fast Retina Keypoint) also provides scale and rotation invariance, however its pattern is based on Gaussians and it is biologically-inspired on the retinal pattern of the human eye.

In practice, FREAK improves upon the sampling pattern and method of pair selection that BRISK uses. Thus, for computing this descriptor, 43 weighted Gaussians at locations around the keypoint are evaluated. As can be see in Figure 1(b), overlappings are considered in order to compute average values related to some points. Moreover, the patterns are much more concentrated near the keypoint that leads to a more accurate description of the keypoint.

To speed up the matching process, the actual FREAK algorithm also uses a cascade for comparing these pairs, and puts the 64 most important bits in the beginning of the descriptor. Just like BRISK, FREAK leads to a 512 bit binary descriptor.

3. BOSSANOVA REPRESENTATION

In this section, we only provide a brief introduction to the BossaNova mid-level image representation, which offers more information-preserving pooling operation based on a distance-to-codeword distribution. More details can be found in [1, 2].

Let $\mathcal{X}$ be an unordered set of binary descriptors extracted from an image. $\mathcal{X} = \{\mathbf{x}_j\}$, $j \in \{1, \ldots, N\}$, where $\mathbf{x}_j \in \mathbb{R}^D$ is a binary descriptor vector and N is the number of binary descriptors in the image. Let $\mathcal{C}$ be a visual codebook obtained by an unsupervised learning algorithm (*e.g.*, k-medians clustering algorithm). $\mathcal{C} = \{\mathbf{c}_m\}$, $m \in \{1, \ldots, M\}$, where $\mathbf{c}_m \in \mathbb{R}^D$ is a codeword and M is the number of visual codewords. The BossaNova representation of the image $\mathbf{z}$ that is used for classification is defined as follows.

The BossaNova approach follows the BoW formalism (coding/pooling), but proposes an image representation which keeps more information than BoW during the pooling step. Thus, in BossaNova coding, [1] proposed a soft-assignment strategy considering only the k-nearest codewords for coding a local descriptor. Mathematically speaking, the BossaNova coding step can be modeled by a function f as follows:

$$\begin{aligned} f : \mathbb{R}^D &\longrightarrow \mathbb{R}^M, \\ \mathbf{x}_j &\longrightarrow f(\mathbf{x}_j) = \alpha_{\mathbf{j}} = \{\alpha_{m,j}\}, \\ \alpha_{m,j} &= \frac{exp^{-\beta_m d_2(\mathbf{x}_j, \mathbf{c}_m)}}{\sum_{m'=1}^{K} exp^{-\beta_m d_2(\mathbf{x}_j, \mathbf{c}_{m'})}} \end{aligned}$$

where $d_2(\mathbf{x}_j, \mathbf{c}_m)$ is the distance between $\mathbf{c}_m$ and $\mathbf{x}_j$. The parameter β_m regulates the softness of the soft-assignment (the bigger it is, the hardest the assignment).

The BossaNova pooling function g estimates the probability density function of α_m: $g(\alpha_m) = \text{pdf}(\alpha_{\mathbf{m}})$, by computing the following histogram of distances $z_{m,b}$:

$$\begin{aligned} g : \mathbb{R}^N &\longrightarrow \mathbb{R}^B, \\ \alpha_{\mathbf{m}} &\longrightarrow g(\alpha_m) = z_m, \\ z_{m,b} &= \text{card}\left(\mathbf{x}_j \mid \alpha_{m,j} \in \left[\frac{b}{B}; \frac{b+1}{B}\right]\right), \\ &\quad \frac{b}{B} \geq \alpha_m^{min} \quad \text{and} \quad \frac{b+1}{B} \leq \alpha_m^{max}, \end{aligned}$$

where B denotes the number of bins of each histogram z_m, and $[\alpha_m^{min}; \alpha_m^{max}]$ limits the range of distances for the descriptors considered in the histogram computation.

After computing the local histograms z_m for all the $\mathbf{c}_m$ centers, the BossaNova vector $\mathbf{z}$ [1] can be written as:

$$\mathbf{z} = [[z_{m,b}], st_m]^{\mathrm{T}}, \quad (m, b) \in \{1, \ldots, M\} \times \{1, \ldots, B\},$$

where $\mathbf{z}$ is a vector of size $M \times (B+1)$, s is a nonnegative constant and t_m is a scalar value for each codeword, counting the number of binary descriptors $\mathbf{x}_j$ close to that codeword.

In brief, the BossaNova representation is defined by three parameters: the number of codewords M, the number of bins B in each histogram, and the range of distances $[\alpha_m^{min}, \alpha_m^{max}]$ – the minimum distance α_m^{min} and the maximum distance

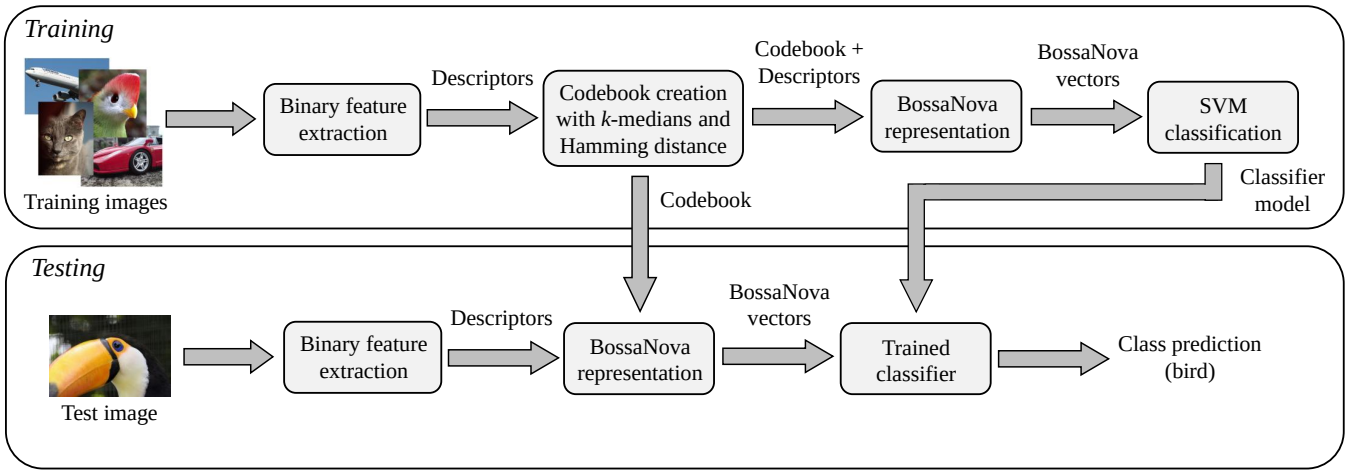


Figure 2: Overview of our approach using local binary descriptors and BossaNova representation.

α_m^{max} in the $\mathbb{R}^D$ descriptor space that define the bounds of the histogram. As in [1], we set up the bounds as $\alpha_m^{min} = \lambda_{min} \cdot \sigma_m$ and $\alpha_m^{max} = \lambda_{max} \cdot \sigma_m$, where σ_m is the standard deviation of each cluster $\mathbf{c}_m$ obtained by k-medians clustering algorithm.

Avila *et al.* [1] applied their representation in the context of visual recognition. In comparison to the BoW representation, BossaNova significantly outperforms it. Besides, the BossaNova approach were ranked at the second place, considering only visual-based approaches, in the ImageCLEF 2012 challenge [3]. Furthermore, by using a simple histogram of distances to capture the relevant information, the method remains very flexible and keeps the representation compact. For those reasons, we choose to apply the BossaNova approach for mid-level features.

4. VISUAL RECOGNITION USING LOCAL BINARY DESCRIPTORS

Recognizing categories of objects and scenes is a fundamental human ability and an important, yet elusive, goal for computer vision research. Images consist of pixels that have no semantic information by themselves, making the task very challenging.

Over the last decade, progress in visual recognition tasks has been quantifiable thanks to (i) the design of discriminative low-level local descriptors, such as SIFT, and (ii) the emergence of mid-level representations based on the Bag-of-Words (BoW) model.

On the subject of low-level features, binary descriptors have recently emerged as low-complexity alternatives to state-of-the-art descriptors. Despite the promising results obtained in visual recognition tasks, the previous methods have employed binary descriptors, in which the invariance of rotation and scale are missing, as well as simple mid-level image representations.

In order to address those problems, we propose combining more robust binary descriptors, such as BRISK [16], with a recent mid-level image representation, namely BossaNova [1], which enriches the BoW representation, while preserving the local descriptor information. As shown in the experiments, our approach outperforms previous methods in literature that use binary descriptors as low-level features.

In Figure 2, we illustrate the overview of our approach pipeline, which involves two phases: training and testing. In training phase, we first extract binary image descriptors on a dense spatial grid. As discussed in [10], that setup for binary descriptors extraction proves to give very good performances in standard image datasets. Next, following the visual recognition strategy, the local descriptors must be encoded into a mid-level representation to be used for the classification task. However, a visual codebook must be created before the encoding. Thus, we apply a k-medians clustering algorithm instead of the classical k-means method. The main reason is the type of descriptors employed there, *i.e.*, as the binary descriptors are represented by binary values, the k-medians clustering performs better. Additionally, the Euclidean distance is replaced by Hamming distance in order to compute distance between the descriptors and the centroids. After, for each image, we extract the BossaNova mid-level feature vector. Finally, once we obtained the BossaNova vectors, one-versus-all classification is performed by non-linear SVM classifiers. The kernel matrices are computed as $exp(-\gamma d(x, x'))$ with d being the distance and γ being fixed to the inverse of the pairwise mean distances.

In testing phase, a new/test image is classified by applying the trained classifier obtained during the training phase. Thus, for the test image, the binary descriptors are extracted on a dense spatial grid. Next, the BossaNova mid-level feature vector is generated using the visual codebook previously created. After, that feature vector is given as input to the trained classifier to predict the class label of the test image.

5. EXPERIMENTAL RESULTS

In this section, we present some experimental results for visual recognition task. We assess our approach on two challenging datasets: (i) PASCAL VOC 2007 [11] (visual object categorization); and (ii) Pornography [1] (video classification). Each dataset is described at the moment of its use.

In order to study the behavior of binary descriptors and their mid-level representation, for both datasets, four binary descriptors (BRIEF, ORB, BRISK, FREAK) are densely extracted (every 6 pixels). We obtained the code of binary descriptors from OpenCV's repository [6], one of the most popular libraries for computer vision. All binary descriptors

are extracted with their default parameters. Also, we used the BossaNova code made available at `http://www.npdi.dcc.ufmg.br/bossanova`.

All experiments were conducted on a 64-bit Linux machine (Ubuntu 12.04) powered by Intel® Xeon® CPU X5670 @ 2.93 GHz CPUs with 24 cores and 70 GB RAM. Despite the large computational power available, we do not require that power to process our experimental results.

5.1 Results for PASCAL VOC 2007 dataset

The PASCAL VOC 2007 dataset [11] consists of 9,963 images collected from Flickr photo-sharing website. The goal of this challenge is to recognize 20 visual object classes in realistic scenes (*aeroplane*, *bicycle*, *bird*, *boat*, *bottle*, *bus*, *car*, *cat*, *chair*, *cow*, *dining table*, *dog*, *horse*, *motorbike*, *person*, *potted plant*, *sheep*, *sofa*, *train*, *tv/monitor*). Those images are split into three subsets: *train* (2,501 images), *val* (2,510 images) and *test* (4,952 images). Our experimental results are obtained on *train+val/test* sets.

In order to learn the codebook, we apply the k-medians clustering algorithm with Hamming distance over five hundred thousand descriptors randomly sampled. For the BossaNova representation, we also incorporate spatial information using the standard spatial pyramidal matching scheme [15]. In total, four spatial cells are extracted (1×1; 3×1).

As described in Section 3, the BossaNova representation has three parameters. Here, we kept the BossaNova parameter values the same as in [1] ($B = 2$, $\lambda_{min} = 0.4$ and $\lambda_{max} = 2$, $s = 10^{-3}$), except for the number of visual codewords, we set $M = 1024$.

The classification performance is measured by the mean Average Precision (mAP) across all classes.

Table 1 shows the results of our experiments over PASCAL VOC 2007 dataset. We can notice that our approach (binary descriptor and BossaNova representation) outperforms the previous methods, which used the classical BoW representation. In our results, it should be noted that BRISK gives the best result (mAP = 38.00%), while FREAK gives the lowest result (mAP = 33.32%) on PASCAL VOC 2007.

Table 1 also shows the comparison with published results. In comparison to Zhang et al. [25], all our results outperform the Zhang's result (mAP = 33.24%) using LBP + BoW (one scale) with a codebook of 1,200 words. Their best published result is 35.17% for LBP + BoW (multi-scale). Regarding this result, we can see that, even without using multi-scale descriptors, our results outperform the Zhang's result, except for FREAK descriptor.

The comparison to Chatzilari *et al.* [10] is particularly relevant, because we employ the same binary extraction as them (BRIEF and ORB descriptors with default parameters). We can observe that our results, BRIEF + BossaNova (mAP = 36.22%) and ORB + BossaNova (mAP = 37.14%), are much better than Chatzilari's results, BRIEF + BoW (mAP = 21.54%) and ORB + BoW (mAP = 21.62%). Furthermore, they used a codebook size of 2,000 visuals words, while we only used 1,024 visual words. In view of that, our results are remarkably good, since it is well known that larger codebooks lead to higher accuracy [9].

5.2 Results for Pornography dataset

Pornography consumption has increased in recent years, which is due in large part to the availability and anonymity provided by the Internet [22]. Pornographic material, however, is often unwelcome is certain environments (*e.g.*, schools, workplaces), channels (*e.g.*, general-purpose social networks), or for certain publics (*e.g.*, children). That raises the need to detect and filter such content.

In this section, we explore our approach in the real-world application of pornography detection. We evaluate our approach on Pornography dataset [1], which contains nearly 80 hours of 400 pornographic and 400 non-pornographic videos. The pornographic class consists of several genres of pornography and depicts actors of many ethnicities, including multi-ethnic ones. The non-pornographic class is divided in two subsets: *easy*, with 200 videos randomly selected from the Internet; and *difficult*, with 200 videos selected from textual search queries like "beach", "wrestling" and "swimming", which is particularly challenging for the detector.

The dataset comes already separated into 16,727 video shots. As in [1], we select the middle frame of each video shot. The experimental evaluation is a 5-fold cross-validation. The video classification performance is reported by accuracy rate, where the final video label is obtained by majority voting over the images.

In our experiments, we apply the same experimental setup proposed by [1]. Our goal is to compare our approach to previous methods that employed this dataset in conjunction with BossaNova representation. In order to learn the visual codebook, we create a vocabulary by using a k-medians clustering algorithm with Hamming distance over one million randomly sampled descriptors. For the BossaNova representation, we kept the parameter values the same as in [1]: $M = 256$, $B = 10$, $\lambda_{min} = 0$, $\lambda_{max} = 3$ and $s = 10^{-3}$. Also, in the interest of a fair comparison, we do not incorporate

Table 1: Image classification mAP (%) results of our approach and published results on PASCAL VOC 2007 dataset [11].

	Approach	mAP (%)
Published results	BoW (LBP, one scale) [25]	33.24
	BoW (LBP, multi-scale) [25]	35.17
	BoW (BRIEF) [10]	21.54
	BoW (ORB) [10]	21.62
Our results	BossaNova (BRIEF)	36.22
	BossaNova (ORB)	37.14
	BossaNova (BRISK)	**38.00**
	BossaNova (FREAK)	33.32

Table 2: Video classification (%) results (and standard deviations) of our approach and published results on Pornography dataset [1].

	Approach	Acc. (%)
Published results	BoW (HueSIFT) [1]	83.0 ± 3
	BossaNova (HueSIFT) [1]	**89.5** ± 1
Our results	BossaNova (BRIEF)	86.3 ± 3
	BossaNova (ORB)	86.5 ± 3
	BossaNova (BRISK)	88.6 ± 2
	BossaNova (FREAK)	86.9 ± 3

spatial information to BossaNova representation.

Table 2 shows our results and the ones reported on the literature over the Pornography dataset. Again, we note that, in our results, BRISK descriptor gives the best result (88.6%). Also, we can observe that our best result is close to the best one reported. Here, it is important to notice that, the best published result is obtained by using HueSIFT descriptors, a SIFT variant including color information, which is particularly relevant for this dataset. Furthermore, we can show the advantage of our approach (BossaNova + binary descriptors) when compared to the classical BoW approach, which also employed HueSIFT descriptors.

6. CONCLUSION

In this paper, we proposed an approach for the visual recognition tasks, which employs local binary image descriptors in conjunction with the recent mid-level image representation, namely BossaNova.

We experimentally compared the performances of our approach with the published results on PASCAL VOC 2007 dataset, a benchmark in visual object category recognition, as well as on a real-world application of pornography detection. As shown in our experimental results, our approach surpassed the previous methods (up to nearly 16.5%) on PASCAL VOC 2007. For pornography video classification, our approach yielded results comparable to the state-of-the-art result, which employs HueSIFT descriptors.

From evaluation results obtained, BRISK is recommended as the best binary descriptor for visual recognition tasks. In both datasets, it gives the best results.

Possible directions for future works include to evaluate the most recent binary descriptors, regarding the accuracy and processing time. Also, we hope to improve our results by exploiting further parameters, as in our experiments we used default parameter values.

7. ACKNOWLEDGMENTS

The authors are thankful to InWeb, CAPES, CNPq and FAPEMIG, Brazilian Research and Development agencies, for the support to this work.

8. REFERENCES

[1] S. Avila, N. Thome, M. Cord, E. Valle, and A. Araújo. Pooling in image representation: the visual codeword point of view. *CVIU*, 117(5):453–465, 2013.

[2] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo. BOSSA: Extended bow formalism for image classification. In *ICIP*, pages 2909–2912, 2011.

[3] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo. Bossanova at imageclef 2012 flickr photo annotation task. In *Working Notes of the CLEF*, 2012.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-up robust features (SURF). *CVIU*, 110(3):346–359, 2008.

[5] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, pages 2559–2566, 2010.

[6] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[7] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: binary robust independent elementary features. In *ECCV*, pages 778–792, 2010.

[8] A. Canclini, M. Cesana, R. A., M. Tagliasacchi, J. Ascenso, and C. R. Evaluation of low-complexity visual feature detectors and descriptors. In *Int. Conf. on Digital Signal Processing*, 2013.

[9] K. Chatfield, V. Lemtexpitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.

[10] E. Chatzilari, G. Liaros, S. Nikolopoulos, and Y. Kompatsiaris. A comparative study on mobile visual recognition. In *MLDM*, volume 7988, pages 442–457. Springer Berlin Heidelberg, 2013.

[11] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007.

[12] D. Gálvez-López and J. Tardós. Real-time loop detection with bags of binary words. In *IROS*, pages 51–58, 2011.

[13] D. Gálvez-López and J. Tardós. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012.

[14] C. Grana, D. Borghesani, M. Manfredi, and R. Cucchiara. A fast approach for integrating ORB descriptors in the bag of words model. In *SPIE Conference Series*, volume 8667, 2013.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[16] S. Leutenegger, M. Chli, and R. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *ICCV*, pages 2548–2555, 2011.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[18] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[19] R. Ortiz. FREAK: Fast retina keypoint. In *CVPR*, pages 510–517, 2012.

[20] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *ECCV*, 2006.

[21] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski. ORB: An efficient alternative to SIFT or SURF. In *ICCV*, pages 2564–2571, 2011.

[22] M. Short, L. Black, A. Smith, C. Wetterneck., and D. Wells. A review of internet pornography use research: methodology and content from the past 10 years. *Cyberpsychology, Behavior, and Social Networking*, 15(1):13–23, 2012.

[23] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[24] C. Whiten, R. Laganiere, and G.-A. Bilodeau. Efficient action recognition with MoFREAK. In *CRV*, pages 319–325, 2013.

[25] Y. Zhang, C. Zhu, S. Bres, and L. Chen. Encoding local binary descriptors by bag-of-features with hamming distance for visual object categorization. In *ECIR*, pages 630–641, 2013.