Binary local descriptors based on robust hashing

Luca Baroffio, Matteo Cesana, Alessandro Redondi, Marco Tagliasacchi
Dipartimento di Elettronica e Informazione, Politecnico di Milano
Piazza Leonardo da Vinci, 32 - 20133 Milano - Italy
{surname}@elet.polimi.it

***Abstract*—A robust hash, or content-based fingerprint, is a succinct representation of the perceptually most relevant parts of a multimedia object. A key requirement of fingerprinting is that elements with perceptually similar content should map to the same fingerprint, even if their bit-level representations are different. In this work we focus on the construction of discriminative binary local descriptors exploiting a combination of content-based fingerprinting techniques and computationally efficient filters (box filters, Haar-like features, etc.) applied to image patches. In particular, we define a possibly large set of filters and iteratively select the most discriminative ones resorting to boosting techniques. The output values of the filtering process are quantized to one bit, leading to a very compact binary descriptor. Preliminary results show that such descriptor leads to compelling results, outperforming SIFT in terms of true-positive / false-positive rate when using as few as 64 bits.**

I. Introduction

Visual features provide a compact representation of the content of a given image patch that is robust and invariant to many global and local transformations. Binary descriptors have recently emerged as low-complexity alternatives to state-of-the-art descriptors such as SIFT [1]. The simplest descriptor of this class is BRIEF [2], which provides a binary description string whose entries are the result of different comparisons between couples of (smoothed) pixel values selected at random within a patch around the keypoint. BRISK [3] refines the process, introducing constraints about the pattern of pixel locations to be used for the comparisons and achieving rotation invariance. More recently, DBRIEF [4] was proposed. The elements of its description string are the result of the binarization of discriminative projections that can be computed fast.

In this paper we propose a novel binary descriptor, whose construction is based on robust hashing. A robust hash or content-based fingerprint of a media file is a compact signature of such file that preserves its semantic information under allowable changes made to it, while at the same time being as distinctive as possible. That is, a fingerprinting algorithm should be robust to a broad set of modifications (e.g., small rotations, compression, scaling, noise, etc.) and at the same time it should be able to discriminate between different (or even tampered) media contents. Specifically, the construction of our descriptor is based on the work by Ke et al. [5], that proposes a music identification system based on content-based fingerprinting, a set of Haar-like filters applied on the signal spectrogram and a pairwise boosting algorithm. The methods was later extended in Lee et al. [6] for content-based video retrieval based on spatio-temporal features.

MMSP'13, Sept. 30 - Oct. 2, 2013, Pula (Sardinia), Italy.
978-1-4799-0123-4/13/$31.00

II. Learning discriminative binary descriptors

In traditional digital fingerprinting methods, robust binary hashes are obtained starting from intermediate features, applying a set of heuristically chosen quantizers. Fingerprinting can be actually seen as a classification problem, since the goal is to assign matching multimedia content to the same class. Hence, any classification algorithm could be employed to obtain discriminative robust hashes. Among all the possible classification algorithm, AdaBoost [7] is a simple yet powerful technique that combines properly several weak learners to obtain a single strong classifiers whose performance is significantly better than the ones of any weak learner.

Ke et al. [5] propose a pairwise variant of the AdaBoost, which is the Pairwise Boosting algorithm for audio fingerprinting. Pairwise Boosting algorithm differs from AdaBoost in the fact that it does not directly exploit intermediate representations themselves but pairs of them, that are labeled as either matching or non-matching and used as training data.

Given an image intensity patch $x_n \in \mathbb{R}^{R \times C}$ we look for a a P-dimensional binary description string $D(x_n) = [D_1(x_n)\ D_2(x_n) \dots D_P(x_n)] \in \{-1, 1\}^P$ representing a robust hash for such an image patch. To this end, a training stage is performed resorting to a dataset of image patches extracted from a given collection of images. Let $L(x_n, x_m) \in \{-1, 1\}$ be a label describing the ground truth relationship between x_n and x_m. Such a label will assume a value equal to 1 if the two patches are matching whereas it will be equal to -1 in the case of two non-matching patches. Moreover, let $H = [h_1, h_2, \dots, h_F]$ define a set of image filters.

For each training image patch x_n, we obtain F scalar intermediate representations $[f_1(x_n), f_2(x_n), \dots, f_F(x_n)]$, where $f_i(x_n) = \langle x_n, h_i \rangle$, i.e. the result obtained by filtering the image patch x_n with the filter h_i. Then, such intermediate representations are fed to the Pairwise Boosting algorithm along with the ground truth relationship between each pair of training patches. The goal of such algorithm is to select the M most discriminative filters and the corresponding binarization thresholds. The flow of the Pairwise Boosting algorithm is depicted in Figure 1.

initialize: $w_i = \frac{1}{n}, i = 1..n$

for m = 1..M

1. find the hypothesis $h_m(x_1, x_2)$ that minimizes weighted error over distribution w, where $h_m(x_1, x_2) = sgn[(f_m(x_1) - t_m)(f_m(x_2) - t_m)]$ for filter f_m and threshold t_m
2. calculate weighted error: $err_m = \sum_{i=1}^{n} w_i \cdot \delta(h_m(x_{1i}, x_{2i}) \neq y_i)$
3. assign confidence to h_m: $c_m = \log(\frac{1-err_m}{err_m})$
4. update weights for matching pairs: if $y_i = 1$ and $h_m(x_{1i}, x_{2i}) \neq y_i$, then $w_i \leftarrow w_i \cdot exp[c_m]$
5. normalize weights such that $\sum_{i:y_i=-1}^{n} w_i = \sum_{i:y_i=1}^{n} w_i = \frac{1}{2}$.

final hypothesis:

$H(x_1, x_2) = sgn(\sum_{m=1}^{M} c_m h_m(x_1, x_2))$

Fig. 1. Flow of the Pairwise Boosting algorithm employed to select the most discriminative filters along with the binarization thresholds.

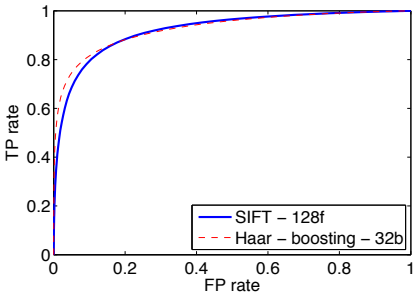

Fig. 2. Receiver Operating Characteristic for the *Notredame* dataset. 32 bits Pairwise Boosted descriptor based on Haar features (red dashed line) outperforms the SIFT descriptor (solid blue line).

III. Experiments

We conducted some tests to evaluate the performance of this novel binary feature construction approach, as well as to compare the results with the ones obtained with other popular descriptors. Brown et al. [8] provide a dataset of image patches of size 64×64 pixels, along with the ground truth relationship between each pair of patches. Such data is employed to train the Pairwise Boosting algorithm, as mentioned in Section II. In particular, we select the first 100k patches from the *Notredame* collection, whereas we test the performance on the whole *Liberty* dataset, composed by 450k patches.

As for the set of filters employed to obtain the intermediate representations, we apply a Discrete Wavelet Transform resorting to Haar wavelets. In particular, we compute the first 5 levels of the transform, retaining all the coefficients except the ones corresponding to the low-pass subband, resulting in a dictionary consisting of 252 different filters.

Figure 2 shows the results of the test conducted on the *Notredame* dataset in terms of a Receiver Operating Characteristic. The blue solid curve represents the performance of SIFT algorithm, consisting in 128 double precision floating point elements. The red dashed curve refers to the descriptor obtained applying Pairwise Boosting to the dictionary of Haar features, retaining the 64 most discriminant elements. Note that the latter approach outperforms the state-of-the-art SIFT descriptor with as few as 64 bits.

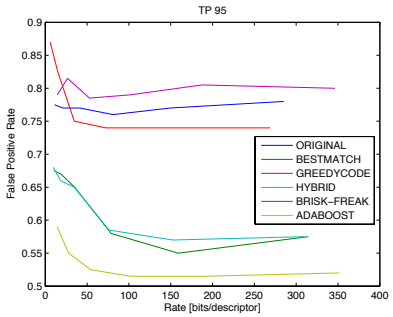

Fig. 3. Results in terms on False Positive rate at .95 True Positive rate for different BRISK dexel selection approaches.

In addition, we exploited the Pairwise Boosting scheme to select the more significant components of the BRISK descriptor. In particular, the BRISK sampling pattern is composed by 60 points, leading to 1770 possible comparisons. We employed the Pairwise Boosting algorithm to select the 64 most significant entries. Figure 3 compares the results of such approach with respect to other feature selection schemes [9]. Pairwise Boosted BRISK achieves the best results in terms of False Positive rate @ .95 True Positive rate.

IV. Conclusions

We introduce a new method to construct discriminative binary descriptors, inspired to fingerprinting techniques. Tests show that such approach achieves better results than SIFT with as few as 64 bits. Future work will address the construction of a more complete set of filters and testing on a full image retrieval pipeline.

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, Nov. 2004.

[2] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: binary robust independent elementary features," in *Proceedings of the 11th European conference on Computer vision: Part IV*, ser. ECCV'10, 2010.

[3] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[4] T. Trzcinski and V. Lepetit, "Efficient Discriminative Projections for Compact Binary Descriptors," in *European Conference on Computer Vision*, 2012.

[5] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer vision for music identification," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2005.

[6] S. Lee, C. D. Yoo, and T. Kalker, "Robust video fingerprinting based on symmetric pairwise boosting," *IEEE Trans. Cir. and Sys. for Video Technol.*, Sep. 2009.

[7] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proceedings of the Second European Conference on Computational Learning Theory*, London, UK, UK, 1995.

[8] M. A. Brown, G. Hua, and S. Winder, "Discriminative learning of local image descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 1, January 2011.

[9] A. Redondi, L. Baroffio, J. Ascenso, M. Cesana, and M. Tagliasacchi, "Rate-accuracy optimization of binary descriptors," in *submitted to International Conference on Image Processing*, sept. 2013.