Combining Fisher Vectors in Image Retrieval Using Different Sampling Techniques

Tomás Mardones[1], Héctor Allende[1] and Claudio Moraga[2]
[1]*Department of Informatics, Universidad Técnica Federico Santa María, Valparaíso, Chile*
[2]*European Centre for Soft Computing, Mieres, Spain*

Keywords: Fisher Vector, Image Retrieval, Feature Sampling Methods, Query by example.

Abstract: This paper addresses the problem of content-based image retrieval in a large-scale setting. Most works in the area sample image patches using an affine invariant detector or in a dense fashion, but we show that both sampling methods are complementary. By using Fisher Vectors we show how several sampling methods can be combined in a simple fashion inquiring only in a small fixed computational cost while significantly increasing the precision of the image retrieval system. As a second contribution, we show Fisher Vectors using their variance component, normally ignored in image retrieval applications, have better performance than their mean component under certain relevant settings. Experiments with up to 1 million images indicate that the proposed method remains valid in large-scale image search.

1 INTRODUCTION

Content-based image retrieval (CBIR) is an important area of research in Multimedia, since it is linked to numerous image applications. Given a query image, the problem consists in finding the most similar images in a database. In the last decade, the most popular method to face this problem with relative success is the Bag of Features (BoF, Bag of Words or Bag of Visual Words) representation (Nister and Stewenius, 2006; Philbin et al., 2007). It can handle up to a one million images database before the precision, time and memory constraints make it impractical for content-based image retrieval (Jégou et al., 2012; Nister and Stewenius, 2006). BoF first extracts local features, called descriptors, from an image and aggregates them into a histogram of "visual words", collecting 0-order statistics of the image descriptors.

To overcome the database size limitation, the Fisher Vector (FV) (Perronnin and Dance, 2007; Perronnin et al., 2010a) and Vector of Locally Aggregated Descriptors (VLAD) (Jégou et al., 2012) image representations replace the BoF histogram with descriptor's higher order statistics. The result in both cases is a single vector which dimension is related to the descriptor's dimension and a single parameter. In scenarios with more than one million images, using Scale Invariant Feature Transform (SIFT) descriptors (Lowe, 2004), it has been shown dimensionality reduction techniques can be used on FV and VLAD leading to very compact representations that preserve a high retrieval precision (Jégou et al., 2012; Perronnin et al., 2010a; Jégou et al., 2011; Gong et al., 2013).

The first contribution of this work is a simple technique to combine Fisher Vectors based on descriptors using different sampling methods, demonstrates under what assumption it does work and gives an intuitive insight on why combining "sparse" and dense descriptions of the image does improve performance. The second contribution is the elaboration of simple tools that allow us to measure the quality and potential of image representations combination methods. The final contribution corresponds to the reconsideration of the use of the Fisher Vector's variance component in image retrieval (Perronnin et al., 2010a).

The remainder of this paper is organized as follows. In the next section the related work is reviewed, next in section 3 brief review of the Fisher Vector representation is provided. In section 4 our contributions are described, and in section 5 their impact is evaluated through several experiments comparing them with other works.

Mardones T., Allende H. and Moraga C..
Combining Fisher Vectors in Image Retrieval Using Different Sampling Techniques.
DOI: 10.5220/0005179201280135
In *Proceedings of the International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2015), pages 128-135
ISBN: 978-989-758-077-2

2 RELATED WORK

Different descriptor sampling techniques can lead to distinct results, this is well known as there are published works that report different results just by changing the sampling method (Gordo et al., 2012). On the other hand, combining sampling techniques has been mostly overlooked in the context of CBIR, as most related works use different descriptors and the same sampling method (Gordo et al., 2012; Douze et al., 2011; Perronnin et al., 2010b; Zheng et al., 2014; Zhang et al., 2012; Wengert et al., 2011).

A group of recent works (Wengert et al., 2011; Zhang et al., 2012; Zheng et al., 2014; Gordo et al., 2012) combined colour and SIFT descriptors using the same keypoints. In (Wengert et al., 2011), using the Bag of Features framework, a global colour descriptor and a local counterpart that couples itself with the SIFT features improving the precision of the retrieval system (Hessian-affine detector) were proposed. In (Zheng et al., 2014) a multi-dimensional inverted index to perform feature fusion in the BoF framework is described, specifically of SIFT and Colour Names (Shahbaz Khan et al., 2012) features (Hessian-affine detector). Gordo et al. (Gordo et al., 2012) used the Fisher Kernel framework, concatenating two Fisher Vectors, one for SIFT features and the other for colour features (both using dense sampling) obtaining a significant performance boost. None of these works employ different sampling methods with the same descriptor.

Douze et al. (Douze et al., 2011) combined several different image representations to build a new representation. Two of them were Fisher Vectors and a histogram of oriented gradients, which used SIFT descriptors extracted from Hessian-affine interest points and the histogram of oriented gradients (densely sampled) respectively, therefore more than one sampling technique was involved, but the complexity of the model makes very hard to know if these elements did improve the precision of the final representation.

3 FISHER VECTOR IMAGE REPRESENTATION

In this work we choose to work with Fisher Vectors because they have analytical properties that are easier to work with compared to VLAD and BoF (Sánchez et al., 2013; Jégou et al., 2012). This representation is based on the work of Jaakola and Haussler (Jaakkola and Haussler, 1999) on Fisher kernels as a method to compare aggregated data combining a generative and discriminative approach. Perronnin and Dance (Perronnin and Dance, 2007) adopted this framework, building Fisher Vectors using SIFT descriptors modelling their probability density function using a Gaussian Mixture Model (GMM).

Let $X = \{x_n, n = 1,...,N\}$ be the set of D-dimensional local descriptors extracted from an image with N descriptors. Let u_λ be a GMM, with parameters $\lambda = \{w_k, \mu_k, \sigma_k, k = 1,...,K\}$ where K is the number of Gaussians and w_k, μ_k and σ_k stand for the mixture weight, the mean vector and the diagonal covariance matrix of Gaussian k respectively. The GMM models the generative process of any descriptor, assuming independency between the descriptors generation. The Fisher Vector mean and variance component corresponding to the k-th Gaussian correspond to:

$$\mathcal{G}^X_{\mu_k} = \frac{1}{\sqrt{w_k}} \sum_{n=1}^{N} \gamma_n(k) \left(\frac{x_n - \mu_k}{\sigma_k} \right), \quad (1)$$

$$\mathcal{G}^X_{\sigma_k} = \frac{1}{\sqrt{w_k}} \sum_{n=1}^{N} \gamma_n(k) \frac{1}{\sqrt{2}} \left[\frac{(x_n - \mu_k)^2}{\sigma_k^2} - 1 \right]. \quad (2)$$

where $\gamma_n(k)$ corresponds to the soft assignment of descriptor x_n to the k-th Gaussian:

$$\gamma_n(k) = \frac{w_k u_k(x_n)}{\sum_{j=1}^{K} w_j u_j(x_n)}, \quad (3)$$

with $\sum w_i = 1, w_i \in [0,1], i = \{1,\cdots,K\}$ and $n = \{1,\cdots,N\}$.

The final FV corresponds to the concatenation of every component. To avoid the dependence on the sample size the resulting FV is divided by the sample size N. Finally two usual normalization steps are used: power normalization (Perronnin et al., 2010a) and L_2 normalization (Perronnin et al., 2010a; Sánchez et al., 2013). For further details the reader may refer to (Sánchez et al., 2013).

4 COMBINING FISHER VECTORS

SIFT descriptors extracted from regions found by interest point detectors and those extracted by using dense sampling obey to different generation processes. This occurs because most interest point detectors center the descriptor in a high contrast area like an edge and rotates it following some criteria to make it invariant to several transformations. These characteristics make these descriptors unlikely to describe plain regions like the sky and to differentiate edges with the same aspect, but different rotations in the

same image. Therefore, the probability density functions related to each set of descriptors are not equal, so their respective associated GMMs and Fisher Vectors are different.

In the following some simple tools that will be useful to estimate the expected improvement of combining different Fisher Vectors will be described. After that we will provide a proof that shows concatenating Fisher Vectors can be a good solution under certain circumstances.

4.1 Combination Performance Tools

A question that we asked ourselves was: how can we know if Fisher Vectors based on descriptors sampled differently can improve the precision of our information retrieval system. For simplicity, let FV_1 and FV_2 be Fisher Vectors based on different sampling strategies. A requisite is that they must have complementary information: if FV_1 obtains the same results as FV_2 in every query , it is unlikely that their combination would improve the performance of the system. On the other hand, if by using FV_1 we obtain good results for a set of queries and FV_2 provides good results in a complementary set where FV_1 does not perform well, we may be able to combine both to increase the precision of the system.

To test if two sampling methods are able to work together we propose to use a histogram of the difference of average precision between both. Let $AP_{FV_s}(q)$ be the average precision for a query q using a Fisher Vector with the s sampling strategy, the AP difference between the use of sampling method 1 and 2 is:

$$diff_{AP}(q) = AP_{FV_1}(q) - AP_{FV_2}(q). \qquad (4)$$

By using several queries it is possible to build a histogram. To use this formula a benchmark dataset is needed. This can be useful in some situations, because different sampling techniques are more appropriate for certain image types (e.g. nature, buildings, sculptures, medical).

In Figure 1 it is shown the $diff_{AP}$ histogram obtained from Fisher Vectors based on Hessian-affine and dense sampled descriptors respectively. It is important to notice that $diff_{AP}$ values lie in the $[-1,1]$ range and to know what those values mean. Positive values of Eq.4 represent queries where FV_1 performed better than FV_2. The most critical case is when FV_1 obtains the maximum AP 1 and FV_2 the minimum 0 and viceversa; in Figure 1 it is possible to see that this accounts approximately for 10% of the queries. Negative values represent queries where FV_2 obtained a better precision. Both positive and negative values of

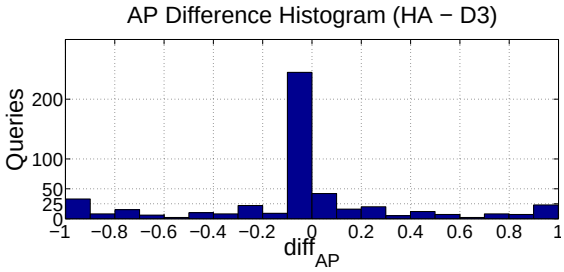


Figure 1: Average precision difference histogram from Holidays dataset (500 queries). FVs are computed using Hessian-affine (HA) and dense (D3) sampled descriptors. They are compared using Eq.(4) in each one of the 500 queries.

$diff_{AP}$ are necessary to know that two image representations are complementary. Values near to zero of Eq.4 represent queries where both methods performed equally good or bad. For these queries the combination of different methods is less prone to obtain better results.

It is also trivial to expand this tool to any image representation, as long as it is possible to obtain the average precision of a query.

To have a rough idea of the potential of combining several representations, the oracle combination function max_{AP} is introduced. By oracle we mean this function needs to know the AP obtained by every image representation beforehand:

$$max_{AP}(q,R) = max_{r\in R}\left(AP_r(q)\right), \qquad (5)$$

where q is a query and R the set of possible image representations. Using max_{AP} the mean AP (MAP) is obtained by averaging $max_{AP}(q,R)$ over all q.

Basically max_{AP} chooses the best representation for each query procuring in the worst case the best MAP obtained with an individual method.

The main goal of using this function is to know a soft upper limit that we can reach by improving combination methods. In section 5 max_{AP} will be used to compare the results obtained by us.

4.2 Concatenation as a Combination Method

Fisher Vectors concatenation is not a new idea, this simple technique has been used before, mainly to combine SIFT and color descriptors (Gordo et al., 2012; Perronnin et al., 2010b), but its use has not been justified, just has been part of the experimental setup.

In the remainder of this section we will show what implicit assumption is done when Fisher Vectors are concatenated.

Let $u_1 = \sum_{k=1}^{K} w_{1,k}\mathcal{N}(X_1|\mu_{1,k},\sigma^2_{1,k})$ be a mixture of K Gaussians that models the X_1 descriptors probability distribution. On the other hand, let $u_2 = \sum_{k=1}^{K} w_{2,k}\mathcal{N}(X_2|\mu_{2,k},\sigma^2_{2,k})$ be another mixture of K Gaussians representing the X_2 descriptors distribution. Without loss of generality we suppose that both GMMs have the same number of Gaussians. X_1 and X_2 are D dimensional and the covariance matrices of u_1 and u_2 are diagonal following the Fisher Vector framework.

Let X be a sample of X_1 and X_2 descriptors. If we assume that the feature space spanned by X_1 is very different from X_2's feature space, the following should be true:

$$P(X \in X_1) = \sum_{k=1}^{K} w_{2,k}\mathcal{N}(X|\mu_{2,k},\sigma^2_{2,k}) \approx 0 \quad (6)$$

$$P(X \in X_2) = \sum_{k=1}^{K} w_{1,k}\mathcal{N}(X|\mu_{1,k},\sigma^2_{1,k}) \approx 0 \quad (7)$$

Equations 6 and 7 are only approximately zero because X_1 and X_2 are in the same feature space. This implies that any D-dimensional Gaussian has a chance (in this case, near to zero) to generate a descriptor, even when it is very far from its mean.

Leveraging the Gaussians independence, it is possible to join both mixtures, so one "big" mixture can represent descriptors sampled from X_1 and X_2:

$$u = \sum_{k=1}^{2K} w_k\mathcal{N}(X|\mu_k,\sigma^2_k), \quad (8)$$

where $w_k = [w_{1,1},\cdots,w_{1,K},w_{2,1},\cdots,w_{2,K}]^T$, $\mu_k = [\mu_{1,1},\cdots,\mu_{1,K},\mu_{2,1},\cdots,\mu_{2,K}]^T$ and $\sigma_k = [\sigma_{1,1},\cdots,\sigma_{1,K},\sigma_{2,1},\cdots,\sigma_{2,K}]^T$.

If we use this mixture to compute a FV, using a set X of N descriptors sampled from X_1 and X_2, the contribution of the ith Gaussian is depicted by Eq. 1 and 2. What changes is the $\gamma_n(k)$ function:

$$\gamma_n(k) = \frac{w_k\mathcal{N}(x_n|\mu_k,\sigma^2_k)}{\sum_{j=1}^{2K} w_j\mathcal{N}(x_n|\mu_j,\sigma^2_j)}, \quad (9)$$

with $\sum w_i = 1, w_i \in [0,1], i = \{1,\cdots,2K\}$ and $n = \{1,\cdots,N\}$.

If $x_n \in X_1$ and $i = \{1,\cdots,K\}$, by making use of Eq.6 we can approximate to zero all the terms related to the u_2 Gaussians in the denominator of Eq.9. If $x_n \in X_1$, but $i = \{K+1,\cdots,2K\}$, $\gamma_n(i) \approx 0$.

This implies that the Fisher Vector components starting from the $KD+1$ to the $2KD$ are approximately zero if the descriptors are sampled from X_1. Hence these descriptors can only contribute up to the KD component. Analogously X_2 descriptors contribute only in the $KD+1$, $2KD$ range of the FV.

Therefore if the X descriptors are divided into two groups S_1 and S_2 depending on whether they belong to X_1 or X_2 respectively and only the relevant Gaussians are taken into account (e.g. u_1 for S_1) the FV can be decomposed as $\mathcal{G}^X_\lambda = [\mathcal{G}^{S_1}_{\lambda_1}\mathcal{G}^{S_1}_{\lambda_1}]^T$, where λ, λ_1 and λ_2 correspond to the parameters of u, u_1 and u_2 respectively. This vector is equivalent to the one we obtain by concatenating the Fisher Vectors produced independently using the initial Gaussian mixtures and their respective descriptors.

When using PCA or other dimension reduction technique on the concatenated Fisher Vectors, they should target each FV individually to benefit from the knowledge that each FV comes from a different distribution.

One important advantage of using concatenated Fisher Vectors, compared to the standard approach, is that additional FVs provide extra information, while there is only a fixed computational cost overhead when extracting additional features of the image and the process of learning the parameters of an additional GMM. In the next section it is shown that this method can obtain better precision with the same memory usage.

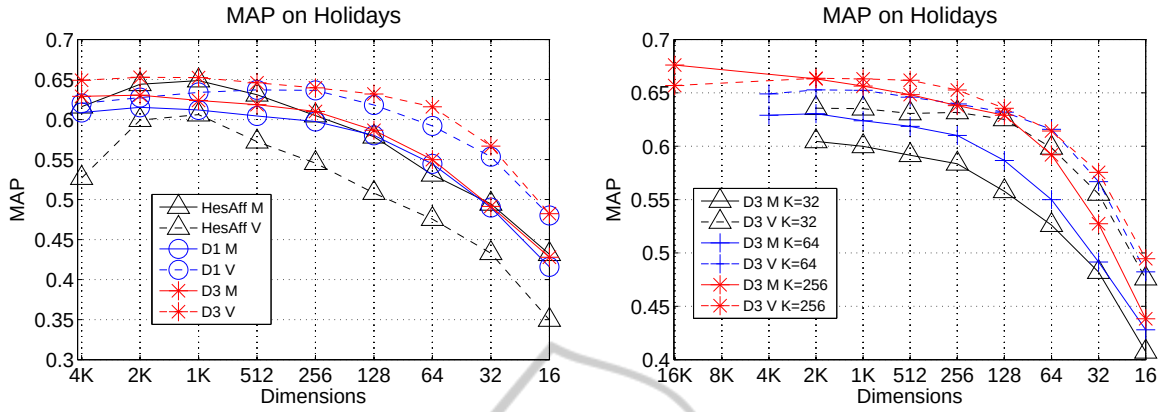
5 EXPERIMENTS AND RESULTS

First the datasets and features used in the experiments are described. Then results for individual and concatenated representations are provided for several settings. Finally, the results are compared with other recent works.

5.1 Datasets and Features

Datasets. The following two public benchmarks are employed. INRIA Holidays (Jégou et al., 2008) consists of 1,491 images of 500 scenes and objects. Each scene / object has a query image and accuracy is measured as the Mean Average Precision (MAP) (Manning et al., 2008). The University of Kentucky Benchmark (UKB) (Nister and Stewenius, 2006) consists of 10,200 images of 2,550 objects. Each image is used alternatively as a query to search within the 10,200 images (including itself) and the performance is measured as 4×recall@4 (called Kentucky Score sometimes) averaged over the 10,200 queries. Therefore, the score goes from 0 to four on this dataset.

The MIRFLICKR-25K dataset (Huiskes and Lew, 2008) is used to learn the GMM parameters and the PCA matrices. For the large-scale experiments reported in Section 5.5, the MIRFLICKR-1M dataset images are used as distractors (Huiskes and



(a) (b)

Figure 2: Comparing the mean (M) and variance (V) component using different sampling techniques. (a) Hessian affine based sampling and dense sampling at two scale settings are tested . (b) The effect of using different number of Gaussians with 3 scale dense sampling.

Lew, 2008), except for the 25K images repeated in MIRFLICKR-25K.

Features. 128-dimensional RootSIFT descriptors (Arandjelović, 2012) will be used as the local descriptors as they have become an increasingly popular choice that performs better than SIFT in image retrieval. Three sampling methods are employed to extract descriptors in most experiments. The first corresponds to dense sampling: 24 pixel length square regions every 8 pixels at 3 scales (D3). The second and third sampling methods are the Hessian affine (HA or HesA) and Hessian Laplace (HL or HesL) interest point detectors respectively (Tuytelaars and Mikolajczyk, 2008). Additionally two other interest point based sampling methods were tested: Harris Affine (HarA) and Harris Laplace (HarL) (Tuytelaars and Mikolajczyk, 2008), but the chosen two performed better or similarly than the rest when concatenated most of the time. Also 1 and 2 scales at dense sampling were tested (D1 and D2 respectively), but using 3 scales works best when concatenated. The extracted features are reduced to 64 dimensions with PCA. The GMM used have 64 Gaussians. Each Fisher Vector is computed separately, then power and L2 normalized (Jégou et al., 2012; Perronnin et al., 2010b). Fisher Vector's mean component is used to represent the interest point based methods, but the variance component is used for the dense sampling method as it steadily attains better precision. To reduce Fisher Vector dimensionality PCA is used. In the rest of the section we will loosely refer to the Fisher Vectors based on the descriptors sampled with the previously mentioned methods as HA, HL and D3V (V stands for variance component).

5.2 Fisher Vector Variance Component and Different Sampling Methods

In most image retrieval works using Fisher Vectors (Perronnin et al., 2010a; Jégou et al., 2012; Gong et al., 2013; Gordo et al., 2012) the variance component is ignored, since in (Jégou et al., 2012) it was reported that using both component (using interest point detectors) did not provide any significant improvement over using just the mean component and doubling the number of Gaussians. Even the "non-probabilistic version" of the Fisher Vectors, VLAD (Jégou et al., 2012), used mainly in image retrieval, does not have a variance component. On the other hand, Sánchez et al.(Sánchez et al., 2013) saw an improvement in image classification by using dense sampling and the variance component for low values of K, compared to the use of the mean component. This evidence was enough to experiment with the variance component.

On Figure 2(a) it is possible to see that by using the Hessian affine interest point detector and the variance component, the performance degrades and is quite unstable. This behaviour was similar in other experiments when using the HesL, HarA and HarL detectors. On the other hand, the results of using dense sampling and the variance component is positive. The first fact which is noticed is the superior MAP obtained by the variance component after every dimensionality reduction. And more importantly for image retrieval is that it maintains its precision much better at lower dimensionalities (at least when using PCA).

Sánchez et al. (Sánchez et al., 2013) mentioned that the difference between using the mean and vari-

Table 1: MAP on Holidays.

Detectors			MAP / maxAP on Holidays				
HA	HL	D3V	D' = 2048	D' = 512	D' = 256	D' = 128	D' = 32
×			64.4	63.1	60.5	57.9	49.4
	×		67.0	65.2	63.1	59.5	51.8
		×	65.3	64.6	64.0	63.2	56.7
×	×		67.7 / 70.9	64.6 / 67.0	61.9 / 64.2	58.2 / 59.9	47.8 / 50.3
×		×	74.3 / 77.4	**73.7** / 75.2	71.6 / 74.0	**69.7** / 71.6	58.7 / **59.1**
	×	×	74.9 / 78.0	72.8 / 75.7	**71.7** / 73.8	**69.7** / **72.8**	**60.0** / 59.0
×	×	×	**75.7** / **78.9**	73.1 / **76.8**	71.1 / **75.1**	68.8 / 72.0	58.6 / 56.2

Table 2: KS on UKB.

Detectors			KS / maxKS on UKB				
HA	HL	D3V	D' = 2048	D' = 512	D' = 256	D' = 128	D' = 32
×			3.30	3.31	3.23	3.14	2.82
	×		3.39	3.33	3.24	3.14	2.79
		×	2.58	2.58	2.55	2.52	2.38
×	×		3.50 / 3.53	3.40 / 3.42	3.32 / 3.34	3.22 / 3.23	2.77 / 2.73
×		×	3.38 / 3.53	3.27 / 3.46	3.21 / 3.39	3.13 / 3.29	2.84 / 2.80
	×	×	3.30 / 3.53	3.17 / 3.43	3.10 / 3.35	3.02 / 3.25	2.76 / **2.77**
×	×	×	**3.53** / **3.62**	**3.41** / **3.52**	**3.34** / **3.45**	**3.25** / **3.33**	**2.88** / 2.69

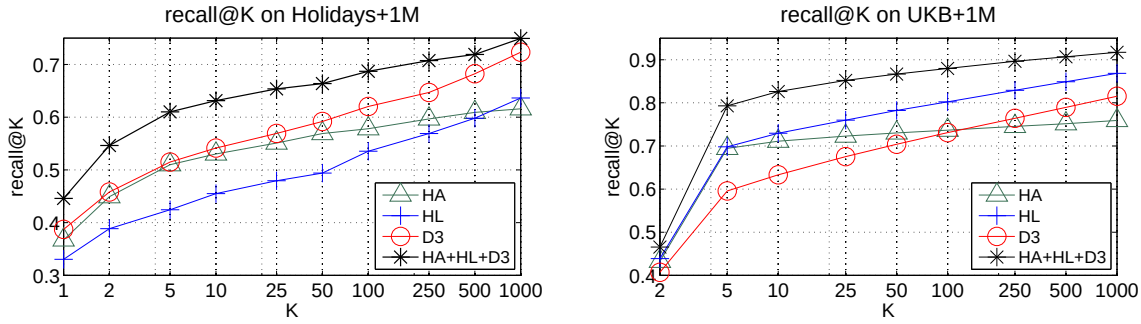
ance component fades as the number of Gaussian increases. In image retrieval it is very important to know how does this behave as the dimensionality decreases. In Figure 2(b) it can be seen that by increasing the number of Gaussians and using the mean component the accuracy increment is substancial and it does not show signs of stopping. Still, the accuracy decreases at a faster pace and at lower dimensions the representations using the variance component do have the advantage and the K selection is less relevant. In additional experiments it was observed that using interest point detectors and the mean component of the Fisher Vector is a better choice independently of the K parameter.

5.3 Concatenate Fisher Vectors

In Table 1 and 2 we can see the results of the baseline methods in both datasets against their combinations at several memory usage scenarios. Note that 512 dimensions for a concatenated representation means that each component uses 256 dimensions (if 2 representations are being used). On Holidays the results are promising, the combination of dense and interest point sampling achieves a MAP increase, ranging from 3.3% to 11.3%, using the same number of dimensions. The max_{AP} results are slightly better all the time, except in some cases where the dimensionality is very small and the precision of individual representations tend to get worse very fast.

On UKB we used the max_{KS} function, analog to max_{AP}, but using the Kentucky Score (KS) instead of AP. The results are much more mixed than in Holidays. UKB is a dataset that focuses just on object recognition, whereas Holidays is a mix of scenes, landmarks and objects (simply holiday pictures). This characteristic allows scale and rotation invariant detectors to perform on UKB particularly well most of the time. This is reflected on the fact that concatenating the FVs based on HA and HL sampling methods produce better results, even if they are methods that detect similar regions. It is interesting to see that max_{KS} has similar results for every dual combination, this led us to think that despite the bad results of the dense sampling method it does provide important information for certain queries. To test this idea, we weighted the HA FV by two and concatenated it to the D8 FV obtaining a score of 3.50 at 2048 dimensions. Adding D8 to the mix of HL and HA representations results in a slightly better representation for UKB, and its difference with max_{KS} is still significant, so better results could be attained with a better combination method.

When seeing the previous results, it is clear that not every sampling method combination will be adequate for every database, nevertheless the combination of HA, HL and D3V should be a good option for most databases containing natural images, since it was able to get good results on both benchmarks.



(a) (b)

Figure 3: recall@K using almost 1M distractor images and 128-D representations.

5.4 Comparison with State of the Art

The results presented in Table 3 show how do other state of the art methods perform against our method. It was hard to decide which works to choose, since normally incompatible methods are compared and in this case most of the ideas presented in the other works are applicable to our work or viceversa, since they improve other points in the pipeline. We omitted methods that used other descriptors (e.g. colour descriptors), query expansion, spatial information and any type of information that makes the representation unsuitable for very large scale retrieval. The work of Jégou et al.(Jégou et al., 2012) serves as a baseline comparison, since it uses the most usual parameters and processing steps. An important difference is that we were not able to reproduce the results in UKB using only the Hessian Affine detector (see Table 2) where Jégou et al. obtained a KS of 3.35 using the full vector and 3.33 with a 128-D representation. The main difference should be the training set used to obtain the PCA matrix. The other works used for comparison improve the normalization step (Delhumeau et al., 2013; Arandjelović and Zisserman, 2013), the learned visual words (Arandjelović and Zisserman, 2013; Jégou and Chum, 2012) and the dimensionality reduction (Delhumeau et al., 2013; Jégou et al., 2012). In general, the results are very favorable for our method in Holidays and in UKB it does a good job at higher dimensions. Still it is fair to emphasize the higher (but fixed) computational overhead present in our algorithm given the use of several detectors.

5.5 Large-scale Experiments

In Figure 3 the *recall@K* is shown for both datasets using MIRFLICRK-1M distractor images. Following the experimental setup of (Delhumeau et al., 2013; Arandjelović and Zisserman, 2013) for large-scale retrieval, 128-D representations were used (43-D$\times$3 for the concatenated one). The proposed method obtains a remarkable advantage on both datasets, disregarding the irregular performance of some sampling methods. The biggest advantage is obtained when using a K from 5 to 50, a very important segment for image retrieval engines. The MAP for Holidays+1M was 56.5% for the proposed method, 12.3% less than the initial MAP, compared to the 15.4% average loss of the individual methods. On UKB+1M the KS was 3.09 for the proposed method, 0.16 less than than the initial KS, compared to the 0.34 average loss of the individual representations.

Table 3: Comparison with the State of the Art.

Method	K	D	Holidays	UKB
FV (Jégou et al., 2012)	64	8192	60.5	3.35
VLAD (Jégou et al., 2012)	64	8192	55.6	3.28
(Arandjelović et al., 2013)	256	32768	64.6	-
(Arandjelović et al., 2013)	256	128	62.5	-
(Delhumeau et al., 2013)	64	8192	65.8	-
(Jégou and Chum, 2012)	64	128	61.4	3.36
HA+HL+D3V	64	8192	75.4	3.53
HA+HL+D3V	64	128	68.8	3.25

6 CONCLUSIONS AND FUTURE WORK

In this work it was primarily shown that the combination of different descriptor sampling methods can be very beneficial in the task of image retrieval. To justify the use of concatenation as a combination method, some of its theoretical implications were treated in the case of using Fisher Vectors. Also a couple of simple tools were presented to help with the task of analyzing the potential of coupling pairs of representations and to have an idea of the performance attainable when combining a group of representations. Furthermore it was shown that the variance component of Fisher Vectors can be very infor-

mative depending on the descriptors probability distribution.

The results obtained encourage further work in this direction. Concatenation is a very simple and fast (practically zero cost) method of combination, but it does not make any distinction between the different representations involved, even if they perform badly with certain kind of images. A deeper research on ensemble methods could prove to be fruitful.

Other way to look the results is that there are several families of descriptors that can contribute with rich information, but the a specific sampling method detects only a few of these families. To identify these sets of complementary descriptors and to develop methods to extract them is another rich field of research.

ACKNOWLEDGEMENTS

This work was supported by the following research and fellowship grants: Fondecyt 1110854, DGIP-UTFSM, MECESUP and CONICYT. The work of C. Moraga was partially supported by the Foundation for the Advance of Soft Computing, Mieres, Spain, and by the CICYT Spain, under project TIN 2011-29827-C02-01.

REFERENCES

Arandjelović, R. (2012). Three things everyone should know to improve object retrieval. In *Proc. CVPR*, pages 2911–2918.

Arandjelović, R. and Zisserman, A. (2013). All about VLAD. In *Proc. CVPR*, pages 1578–1585.

Delhumeau, J., Gosselin, P.-H., Jégou, H., and Pérez, P. (2013). Revisiting the VLAD image representation. In *Proc. ACM Int. Conf. on Multimedia*, pages 653–656.

Douze, M., Ramisa, A., and Schmid, C. (2011). Combining attributes and Fisher vectors for efficient image retrieval. In *Proc. CVPR*, pages 745–752.

Gong, Y., Lazebnik, S., Gordo, A., and Perronnin, F. (2013). Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *Pattern Analysis and Machine Intelligence*, 35(12):2916–2929.

Gordo, A., Rodriguez-Serrano, J. A., Perronnin, F., and Valveny, E. (2012). Leveraging category-level labels for instance-level image retrieval. In *Proc. CVPR*, pages 3045–3052.

Huiskes, M. J. and Lew, M. S. (2008). The MIR Flickr retrieval evaluation. In *Proc. ACM Int. Conf. on Multimedia Information Retrieval*, pages 39–43.

Jaakkola, T. S. and Haussler, D. (1999). Exploiting generative models in discriminative classifiers. In *Proc. Conf. on Advances in Neural Information Processing Systems II*, pages 487–493.

Jégou, H. and Chum, O. (2012). Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In *Proc. ECCV*, pages 774–787.

Jégou, H., Douze, M., and Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV*, volume I, pages 304–317.

Jégou, H., Douze, M., and Schmid, C. (2011). Product quantization for nearest neighbor search. *Pattern Analysis and Machine Intelligence*, 33(1):117–128.

Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., and Schmid, C. (2012). Aggregating local image descriptors into compact codes. *Pattern Analysis and Machine Intelligence*, pages 1704–1716.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York.

Nister, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Proc. CVPR*, pages 2161–2168.

Perronnin, F. and Dance, C. R. (2007). Fisher kernels on visual vocabularies for image categorization. In *Proc. CVPR*, pages 1–8.

Perronnin, F., Liu, Y., Snchez, J., and Poirier, H. (2010a). Large-scale image retrieval with compressed Fisher vectors. In *Proc. CVPR*, pages 3384–3391.

Perronnin, F., Sánchez, J., and Mensink, T. (2010b). Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, pages 143–156.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, pages 1–8.

Sánchez, J., Perronnin, F., Mensink, T., and Verbeek, J. (2013). Image classification with the Fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245.

Shahbaz Khan, F., Anwer, R., van de Weijer, J., Bagdanov, A., Vanrell, M., and Lopez, A. (2012). Color attributes for object detection. In *Proc. CVPR*, pages 3306–3313.

Tuytelaars, T. and Mikolajczyk, K. (2008). Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280.

Wengert, C., Douze, M., and Jégou, H. (2011). Bag-of-colors for improved image search. In *ACM Multimedia*, pages 1437–1440.

Zhang, S., Yang, M., Cour, T., Yu, K., and Metaxas, D. (2012). Query specific fusion for image retrieval. In *Proc. ECCV*, pages 660–673.

Zheng, L., Wang, S., Zhou, W., and Tian, Q. (2014). Bayes merging of multiple vocabularies for scalable image retrieval. In *Proc. CVPR*, pages 1963–1970.