Beyond Bag of Words for Concept Detection and Search of Cultural Heritage Archives

Costantino Grana, Giuseppe Serra, Marco Manfredi, and Rita Cucchiara

Università degli Studi di Modena e Reggio Emilia, Modena MO 41125, Italy

Abstract. Several local features have become quite popular for concept detection and search, due to their ability to capture distinctive details. Typically a Bag of Words approach is followed, where a codebook is built by quantizing the local features. In this paper, we propose to represent SIFT local features extracted from an image as a multivariate Gaussian distribution, obtaining a mean vector and a covariance matrix. Differently from common techniques based on the Bag of Words model, our solution does not rely on the construction of a visual vocabulary, thus removing the dependence of the image descriptors on the specific dataset and allowing to immediately retargeting the features to different classification and search problems. Experimental results are conducted on two very different Cultural Heritage image archives, composed of illuminated manuscript miniatures, and architectural elements pictures collected from the web, on which the proposed approach outperforms the Bag of Words technique both in classification and retrieval.

Keywords: cultural heritage, bag of words, local descriptors, concept detection, image retrieval, similarity search.

1 Introduction

The creation of large digital archives of cultural heritage images, requires experts from different areas to explore the issues of digital collections [7], for the development of information systems and operating platforms able to support both the organization and the access to these repositories [12]. The success of text-based retrieval has raised user expectations about the possibilities of research on media collections, but search engines based only on textual keywords demonstrated their intrinsic limits: the entire content of an image cannot be easily summarized in few keywords. Cultural heritage repositories —incorporating by definition images, texts and often videos, 3D data etc.— should be considered as multimedia repositories so as to adopt all multimedia techniques for digging, understanding and handling such a heterogeneous amount of data.

To cope with multimedia collections, it is necessary to manage the content itself, which may require specific storage and presentation devices, and to manage the associated metadata that can be of different nature and be generated according to a variety of standards. In cultural heritage collections, in which objects are generally subject to some kind of expert analysis, an information

N. Brisaboa, O. Pedreira, and P. Zezula (Eds.): SISAP 2013, LNCS 8199, pp. 233–244, 2013.

system must meet two objectives: the first is to allow the association of elements with their descriptive metadata, and the second is to offer content based retrieval using state-of-the-art technologies. In the last five years most of the proposals for managing multimedia content assume the Bag of Words (BoW) paradigm as an effective approach to provide a compact representation, by clustering local features in a codebook and exploiting visual keyword data for search, concept detection and content understanding.

Recently a novel approach was proposed [9], which represents local features extracted form an image as a multivariate Gaussian distribution, obtaining a mean vector and a covariance matrix. In contrast with the BoW approaches this solution does not require the construction of a visual vocabulary, thus extracting the image descriptor independently from the specific dataset. Allowing the use of linear classifiers, the proposed representation exploits on-line solvers able to deal with large scale datasets that do not fit in memory.

In this paper, we propose to use this novel local features summarization technique for two different tasks, image retrieval and automatic annotation, on two archives of cultural heritage data: the first scenario targets image retrieval on pictures extracted from a Renaissance illuminated manuscript, the second scenario aims at automatically enrich a large dataset of images (automatically crawled from the web) with tags in a concept detection system. For the visual search task (content-based retrieval) we compare different metrics and show which distance is better suited for which descriptor configuration. For the semantic concept detection problem, Stochastic Gradient Descent on-line solver is used, which allows to deal with large scale datasets and high dimensional feature spaces. Differently from the previous use of the descriptor, we evaluate its applicability to coarsely annotated training datasets, showing its ability to deal with the semantic noise, that is the uncertainty of the automatically crawled labels.

2 Related Work

Recently, several local features such as SIFT, SURF, ORB, HOG have become quite popular in representing images due to their ability to capture distinctive details of the images [14]. As introduced, a common approach to integrate the local features into a global representation is to use the the BoW approach, given its simplicity and effectiveness. It consists in three steps: extract local features, generate a codebook and then encode the local features into codes; pool all the codes together to generate the global image representation. The histogram is then fed to a classifier to predict the category [4]. In this approach a key step is the codebook generation, because it is the base to define a high-dimensional BoW histogram. Typically a codebook is built by quantizing local feature descriptors extracted from training images. In recent years, there have been numerous vector quantization approaches to build visual codebooks, such as k-means clustering, or vocabulary trees [15]. However, generated codebooks are not sufficiently flexible to model heterogeneous kinds of new datasets. This is an underlying problem of the BoW approach, because every time the dataset (or more generally the

context) changes, the feature vector of an image must be recomputed. Other elements that have attracted research efforts are the encoding and pooling. The simplest encoding in the literature assigns a local feature to the closest visual word and computes a histogram of visual word frequencies [5]. A recent approach replaces the hard quantization of features with soft-assignment in which each local feature is assigned to multiple visual words [6]. In spite of their simplicity, BoW approaches often introduce large quantization errors and limits in the classification performance. To alleviate these problems, several authors have proposed alternative encodings that retain more information about the original image features [21,16,10]. The Locality-constrained Linear Coding [21] applies locality constraint to select similar basis of local descriptors from a codebook, and learns a linear combination weight of these basis to reconstruct each descriptor. Fisher encoding [16], captures the average first and second order differences between the image descriptors and the centers of a Gaussian mixture model; while the Vector of Locally Aggregated descriptors [10] (VLAD) is a non probabilistic version of Fishers kernels. All of these techniques represent an image by exploiting different strategies to describe relationships between local descriptors and visual words of a codebook.

Instead, we propose to use a parametric distribution and compare its capabilities and proprieties to histogram based approaches. A reasonable first choice is to assume that our data follows a Gaussian distribution, because it has useful mathematical properties, it was extensively used and studied, and its representation requires few parameters [1]. In statistical learning a main aspect is to define function to measure similarity/dissimilarity between two distributions. Several measures in closed form expressions between two multivariate Gaussian densities have been proposed, such as the Bhattacharyya divergence and the symmetric Kullback-Leibler (KL) divergence [11]. Based on these dissimilarities, it is possible to build a non-linear kernel function, which can be used in the classification process. However, this would require an enormous computational effort and would soon become prohibitive when moving to large scale classification problem with high-dimensional feature vectors.

3 Multivariate Gaussian Descriptor

The proposed image signature represents local features extracted from an image (or a sub-region) by a multivariate Gaussian distribution. Let $F = \{\mathbf{f}_1 \dots \mathbf{f}_\mathbf{N}\}$ be a set of local features (e.g. SIFT descriptors, where $d = 128$) extracted through densely sampling in a regular grid on an image W (or a sub-region of W, when Spatial Pyramid Matching is used), we describe them with a multivariate Gaussian distribution supposing that they are normally distributed. The multivariate Gaussian distribution of a set of d-dimensional vectors F is given by

$$\mathcal{N}(\mathbf{f}; \mathbf{m}, \mathbf{C}) = \frac{1}{|2\pi\mathbf{C}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{f}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{f}-\mathbf{m})}, \tag{1}$$

where $|\cdot|$ is the determinant, $\mathbf{m}$ is the mean vector and $\mathbf{C}$ is the covariance matrix ($\mathbf{f}, \mathbf{m} \in \mathbb{R}^d$ and $\mathbf{C} \in \mathbb{S}_{++}^{d \times d}$, with $\mathbb{S}_{++}^{d \times d}$ the space of real symmetric positive semi-definite matrices).

Although the covariance matrix encodes information about the variance of the features and their correlation, it does not lie in a vector space. In fact, the covariance space is a Riemannian manifold and is not closed under multiplication with a negative scalar. Since most of the common machine learning algorithms assume that the data lies in a vector space, we need to define a suitable transformation. The covariance matrix is symmetric positive definite therefore we can adopt the Log-Euclidean metric. The basic idea of the Log-Euclidean metric is to construct an equivalent relationship between the Riemannian manifold and the vector space of the symmetric matrix.

In [19] an approach to map from Riemannian manifolds to Euclidean spaces is described. The first step is the projection of the covariance matrices on an Euclidean space tangent to the Riemannian manifold, on a specific tangency matrix $\mathbf{P}$. The second step is the extraction of the orthonormal coordinates of the projected vector. In the following, matrices (points in the Riemannian manifold) will be denoted by bold uppercase letters, while vectors (points in the Euclidean space) by bold lowercase ones. The projection of $\mathbf{C}$ on the hyperplane tangent to $\mathbf{P}$ becomes:

$$\mathbf{c} = \mathrm{vec}_{\mathbf{I}} \left(\log \left(\mathbf{P}^{-\frac{1}{2}} \mathbf{C} \mathbf{P}^{-\frac{1}{2}} \right) \right), \tag{2}$$

where log is the matrix logarithm operator and $\mathbf{I}$ is the identity matrix, while the vector operator on the tanget space at identity of a symmetric matrix $\mathbf{Y}$ is defined as:

$$\mathrm{vec}_{\mathbf{I}}(\mathbf{Y}) = \left[y_{1,1} \ \sqrt{2} y_{1,2} \ \sqrt{2} y_{1,3} \ldots y_{2,2} \ \sqrt{2} y_{2,3} \ldots y_{d,d} \right]. \tag{3}$$

Thus, after selecting an appropriate projection origin, every covariance matrix is projected to an Euclidean space. Since $\mathbf{c}$ is a symmetric matrix of size $d \times d$ a $(d^2 + d)/2$-dimensional feature vector is obtained.

The projection point $\mathbf{P}$ is arbitrary and even if, as observed in [13], it could influence the performance (distortion) of the projection, from a computational point of view, the best choice is the identity matrix, which simply translates the mapping into a standard matrix logarithm.

The image descriptor is the concatenation of the mean vector and the projected covariance matrix on a Euclidean space obtaining, in the case of SIFT descriptor, a feature with 8384 dimensions. Finally, we empirically observe that most of the values in the concatenated descriptor are low, while few are high. In order to distribute the values more evenly, we adopt the power normalization method proposed by Perronnin et al. [16].

4 Image Similarity Search

The extracted features can be used for a visual search system. In particular we have considered three different metrics which allow to directly compare the

similarity of two images: Cosine Similarity, Euclidean Distance, Kullback-Leibler (KL) divergence.

The cosine distance treats both vectors as unit vectors by normalizing them, giving a measure of the angle between the two vectors. It does provide an accurate measure of similarity but with no regard to magnitude, in contrast to the Euclidean distance which gives the magnitude of difference between the two feature vectors.

The third metric, KL divergence, is a measure of the dissimilarity between two completely determined probability distributions. It is based on Kullback's measure of discriminatory information:

$$I(P_1, P_2) = -\int_{\epsilon} p_1 log(p_1/p_2)\, dx. \tag{4}$$

Kullback realizes the asymmetry of $I(P_1, P_2)$ and describes it as the directed divergence. To achieve symmetry, Kullback defines the divergence as $I(P1, P2) + I(P2, P1)$. The closed form expression for the symmetric KL (sKL) divergence between two multivariate Gaussian densities, $\mathcal{N}_1$ and $\mathcal{N}_2$, can be written as:

$$d_{KL}(\mathcal{N}_1, \mathcal{N}_2) = \frac{1}{2}\mathbf{u}^T(\mathbf{C}_1^{-1} + \mathbf{C}_2^{-1})\mathbf{u} + \frac{1}{2}tr(\mathbf{C}_1^{-1}\mathbf{C}_2 + \mathbf{C}_2^{-1}\mathbf{C}_1 - 2\mathbf{I}), \tag{5}$$

where tr is the matrix trace, $\mathbf{u} = (\mathbf{m}_1 - \mathbf{m}_2)$ and $\mathbf{I}$ is the identity matrix. Note that since the sKL divergence directly compares the Gaussian distributions is not necessary to project the covariance matrices in a vector space through matrix logarithm.

5 Large Scale Online Learning

Multivariate Gaussian Descriptors can be used to learn SVM for classication, in order to automatically enrich a large dataset with detection of semantic concept. Batch-type SVM solvers, such as LibSVM/LIBLINEAR are effective and well known solutions for train classifiers, however they are not feasible for training large digital archives of cultural heritage images. In fact, they are batch methods which require to go through all data to compute gradient in each iteration and most of them require to pre-load training data into memory, which is impossible when the size of the training data explodes.

To deal with large datasets, we propose to use the stochastic gradient descent (SGD) algorithm, recently introduced for SVM classifiers training, because it is an online method and can be easily parallelized to simultaneously train several classifiers. In fact it updates the learning system on the basis of the loss function measured for a single example.

We have training data that consists of N feature-label pairs, denoted as $\{\mathbf{x}_t, y_t\}_{t=1}^{N}$, where $\mathbf{x}_t$ is a $s \times 1$ feature vector representing an image and $y_t \in \{-1, +1\}$ is the label of the image. The selected cost function for binary SVM classification is the hinge loss, that can be written as:

Fig. 1. Example of pictures grouped by class

$$L = \sum_{t=1}^{T} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \max\left[0, 1 - y_t(\mathbf{w}^T \mathbf{x}_t + b)\right], \tag{6}$$

where $\mathbf{w}$ is $s \times 1$ SVM weight vector, λ (nonnegative scalar) is a regularization parameter, and b (scalar) is a bias term. In the SGD algorithm, training data are fed to the system one by one, and the update rule for $\mathbf{w}$ and b respectively are:

$$\begin{aligned} \mathbf{w}_t &= (1 - \lambda\eta)\mathbf{w}_{t-1} + \eta y_t \mathbf{x}_t \\ b_t &= b_{t-1} + \eta y_t \end{aligned} \tag{7}$$

if margin $\Delta_t = y_t(\mathbf{w}^T \mathbf{x}_t + b)$ is less than 1; otherwise, $\mathbf{w}_t = (1 - \lambda\eta)\mathbf{w}_{t-1}$ and $b_t = b_{t-1}$. The parameter η is the step size. We set $\eta = (1 + \lambda t)^{-1}$, following the `vl_pegasos` implementation [20].

In order to parallelize the computation for training SVM classifiers, we randomize the data on disk and we load the data in chunks which fit in memory. We then train the classifiers on further randomizations of the chunks, so that different epochs (one training epoch is defined as providing all training samples to the classifier once) will get the chunks data with different orderings. This last step of randomization turns out to be essential to make the SGD algorithm work properly.

6 Experimental Results

6.1 Datasets Description

We perform the experiments on two different datasets: the first one is a set of pictures from an illuminated manuscript used for image retrieval purposes, the second one was created by querying GoogleImages and used for concept detection.

Fig. 2. Sample images extracted from the GoogleCH dataset, with the corresponding ground truth annotations

The first dataset ("Bible dataset") was created from digitalized pages of the Holy Bible of Borso d'Este, duke of Ferrara (Italy) from 1450 A.D. to 1471 A.D. It is one of the best Renaissance illuminated manuscripts in the world, whose original is held in the Biblioteca Estense Universitaria in Modena (Italy). It is composed by 640 pages, with two-column layered text in Gothic font, spaced out with some decorated drop caps, enclosing thousands of painted masterpieces surrounded by rich decorations. These pages have been digitized at 10 Mpixels. Then an automatic procedure [8] has been adopted to segment the miniature illustrations. The set of images obtained from the segmentation process has been manually refined to define the final dataset of 2281 pictures, publicly available for scientific purposes [2] [1]. In collaboration with a group of art experts, the authors performed a manual classification obtaining a subset of 13 classes, characterized by a clear semantic meaning and a significant search relevance (see Fig. 1). As a result, 41% of the original dataset (903 images) has been uniquely annotated into those classes, while the remaining pictures are considered as distractors, often with similar color, shape and texture distribution.

The second dataset ("GoogleCH dataset", see sample images in Fig. 2) was automatically crawled from GoogleImages, by searching for 20 semantic concepts related to cultural heritage (altar, archaeological sites, bell tower, bridge, building, capital, ceremony, church, city square, concerts, crown, Gothic glass wall, inscription, manuscript, mosaic, musical instruments, orchestra, rose window, statue, Tuscany food). For each concept, about 500 images were downloaded and, excluding some which were wrong links, resized to a fixed width of 640 pixels, with a proportional height scaling. The final dataset contains 9594 images, each annotated with the single concept used on the query. Another 1000 images was downloaded, and manually annotated selecting all the concepts present in the image. In this way the training set can be considered a noisy source of information, but definitely containing some useful information. Of course some of the images thus obtained suffer from the ambiguity of the concept terms or their different meaning in different languages (e.g. "inscription" is the French word for "subscription").

[1] Download the Bible dataset at
`http://imagelab.ing.unimo.it/files/bible_dataset.zip`

6.2 Content-Based Visual Similarity Retrieval

In order to propose a valuable comparison, a large variety of visual descriptors based on BoW has been tested in addition to the Multivariate Gaussian Model on the Bible dataset. In particular, we relied on the code and the implementation proposed by [17], employing the following descriptors: RGB Color Histogram (localCH), a combination of three 1D histograms based on the R, G, and B channels of the RGB color space; Transformed Color Histogram (localTCH), RGB histogram obtained by normalizing the pixel value distributions, achieving scale-invariance and shift-invariance with respect to light intensity; Color Moments (localCM), generalized color moments up to the second order, giving a 27-dimensional shift-invariant descriptor; SIFT descriptor (128-dimensional feature vector); RGB-SIFT descriptor (rgbSIFT), for a total 3×128-dimensional feature vector; RG-SIFT descriptor (rgSIFT), computed for R and G channels independently, for a total 2×128-dimensional feature vector; HSV-SIFT descriptor (hsvSIFT), computed converting the original image into the HSV color space, and considering each channel independently, for a total 3×128-dimensional feature vector; Opponent-SIFT descriptor (oppSIFT), describing all of the channels in the opponent color space [18] using SIFT descriptors; C-SIFT descriptor (cSIFT), as proposed by [3], using the C-invariant color space which eliminates the remaining intensity information from the opponent channels; SURF, a scale- and rotation-invariant interest point detector and descriptor which uses integral images and other optimization and approximations to reduce the computational time.

All these descriptors were extracted using the Harris-Laplace region detector. A codebook has been created for every descriptor through a k-means clustering over 10% of the annotated dataset, randomly selected among all the classes in order to ensure an equal amount of visual information for each of them. The employed distance function is the histogram intersection. The sizes k of the codebooks have been determined empirically. In fact, since the clustering is a process of data compression, too small k's (large compression ratio) will force diverse keypoints into the same visual word reducing the quality of the representation; instead too large k's (small compression ratio) might lead to a sparse representation with similar keypoints mapped into different visual words, increasing the computational requirements without any real benefit. Therefore in our experiments we tested values of k between 2^9 and 2^{14}.

Table 1 reports the detailed results obtained using the different features in terms of Mean Average Precision (MAP). For the BoW approaches every column reports the performance using several codebook sizes. The bottom part of the table reports the results obtained with the proposed descriptor, changing the similarity measure. When using the Cosine similarity or the Euclidean distance, the covariance matrix is projected on the tangent space with the matrix logarithm, while the symmetric KL divergence works directly on the covariance matrix. It is possible to observe that the best results are achieved with the Multivariate Gaussian Model of the rgbSIFT descriptors using the dot product, and the result is significantly better then the best result obtained with the BoW

Table 1. Detailed MAP results obtained using the different features. For the BoW approaches (top of the table) every column reports the performance using several codebook sizes. The distance used is always Histogram Intersection. The bottom part of the table reports the results obtained with the proposed descriptor, changing the similarity measure.

	512	**1024**	**2048**	**4096**	**8192**	**16384**
localCH	0.142	0.145	0.147	0.147	0.149	0.147
localTCH	0.129	0.135	0.139	0.141	0.145	0.147
localCM	0.135	0.141	0.146	0.150	0.152	**0.155**
SIFT	0.134	0.136	0.138	0.139	0.142	0.144
rgbSIFT	0.136	0.137	0.138	0.139	0.139	0.142
rgSIFT	0.144	0.147	0.149	0.152	0.152	0.150
hsvSIFT	0.137	0.137	0.138	0.140	0.141	0.139
oppSIFT	0.138	0.141	0.141	0.142	0.143	0.145
cSIFT	0.139	0.139	0.140	0.143	0.143	0.143
SURF	0.119	0.127	0.130	0.129	0.129	0.128

	Cosine Similarity	**Euclidean**	**sKL divergence**
mgm-SIFT	0.146	0.134	0.151
mgm-rgbSIFT	**0.191**	0.119	0.130

approaches (i.e. the local color moments). Fig. 3 shows the performance comparison of the best configuration for each feature summarization method.

6.3 Concept Detection for Image Enrichment

For the concept detection task, we employ the proposed descriptor and compare it with the state-of-the-art BoW approach setting the histogram size, i.e. the number of cluster centers for the k-means algorithm, to 4000. For this task, the rgbSIFT descriptors are extracted at four scales, defined by setting the width of the spatial bins to 4, 6, 8, and 10 pixels respectively, over a dense regular grid with a spacing of 3 pixels. We use the function `vl_phow` provided by the `vl_feat` library [20] and, apart from the spacing step, the defaults options are used. Since the rgbSIFT descriptor is a 384-dimensional feature, the multivariate Gaussian descriptor of an image (or a sub-region) would become an extremely large vector. For this reason, we obtain the image feature by concatenating the multivariate Gaussian descriptors computed for each color channel separately. Images are hierarchically partitioned into 1×1, 2×2 and 1×3 blocks on 3 levels respectively. The resulting descriptors are then concatenated for both methods. The Mean Average Precision (MAP) is used to evaluate the performance, because commonly adopted in concept annotation scenarios.

With this dataset, we apply SGD, which allows us to deal with the large number of images available. Loading the entire training set on memory (9594 samples) occupies about 8.0GB, requiring to split the data in chunks, each loaded

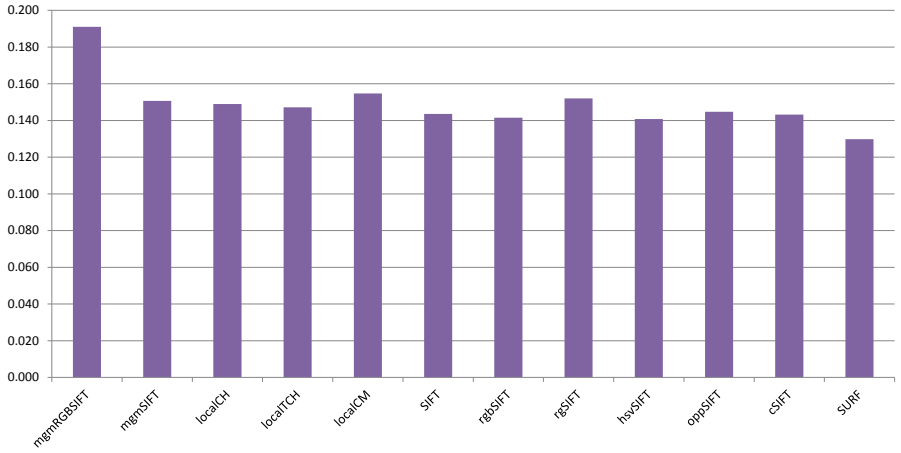

Fig. 3. Comparison of the best MAP values obtained using the different features

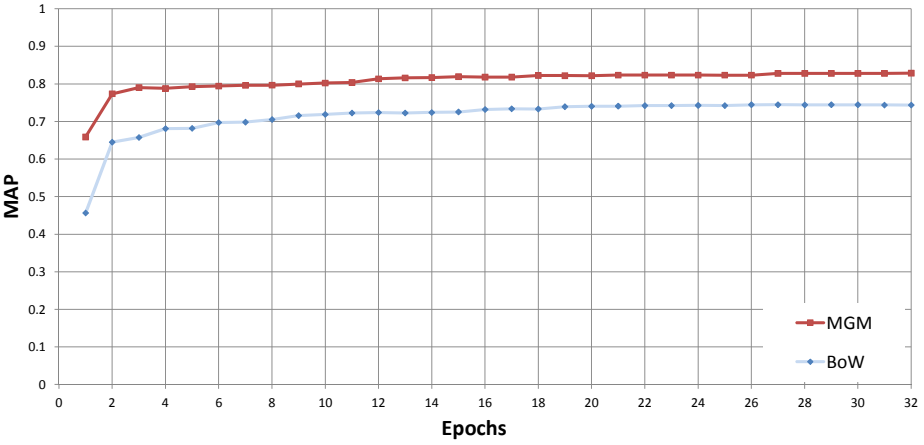

Fig. 4. Mean Average Precision values obtained on the GoogleCH dataset, using the proposed approach and BoW

in turns. To select an appropriate regularization parameter λ for the SGD solver, we randomly split the training set in two and run the SGD varying λ from 10^{-3} to 10^{-7} in power of 10 steps. Based on this preliminary experiments we fix $\lambda = 10^{-5}$. Fig. 4 reports the results of both the proposed approach and BoW in term of MAP at different number of training epochs. Note that the performance rapidly increases in the first 10 epochs, and later tends to remain quite constant. In addition, our method obtains a MAP of 0.83 compared to 0.74 of the BoW approach (at the 30th epoch) and presents better performance at all epochs.

7 Conclusions

In this paper we propose a novel approach for image retrieval and automatic annotation of cultural heritage images. For image retrieval scenario we analyzed three different metrics, while for the automatic annotation we explored the possibility to use noisy data in the training set. The experimental results, on the Bible and GoogleCH datasets, show interesting results both in classification and similarity search with respect to a large variety of visual signatures based on BoW.

References

1. Ali, S., Silvey, S.: A general class of coefficients of divergence of one distribution from another. J. of the Royal Stat. Soc (B) 28(1), 131–142 (1966)
2. Borghesani, D., Grana, C., Cucchiara, R.: Miniature illustrations retrieval and innovative interaction for digital illuminated manuscripts. In: Multimedia Systems (2013)
3. Burghouts, G.J., Geusebroek, J.M.: Performance evaluation of local colour invariants. Computer Vision and Image Understanding 113, 48–62 (2009)
4. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC (2011)
5. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV Workshop Stat. Learn. Comput. Vision (2004)
6. van Gemert, J.C., Geusebroek, J.-M., Veenman, C.J., Smeulders, A.W.M.: Kernel codebooks for scene categorization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 696–709. Springer, Heidelberg (2008)
7. Gonçalves, M.A., Fox, E.A., Watson, L.T., Kipp, N.A.: Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. ACM Trans. Inf. Syst. 22(2), 270–312 (2004)
8. Grana, C., Borghesani, D., Cucchiara, R.: Automatic segmentation of digitalized historical manuscripts. In: Multimedia Tools and Applications, pp. 1–24 (2010)
9. Grana, C., Serra, G., Manfredi, M., Cucchiara, R.: Image classification with multivariate gaussian descriptors. In: ICIAP (2013)
10. Jegou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3304–3311 (2010)
11. Kailath, T.: The divergence and Bhattacharyya distance measures in signal selection. IEEE T. Commun. Techn. 15(1), 52–60 (1967)
12. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: an architecture for complex objects and their relationships. Int. J. Digit. Libr. 6(2), 124–138 (2006)
13. Martelli, S., Tosato, D., Farenzena, M., Cristani, M., Murino, V.: An FPGA-based Classification Architecture on Riemannian Manifolds. In: DEXA Workshops (2010)
14. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE T. Pattern Anal. 27(10), 1615–1630 (2005)
15. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: IEEE International Conference on Computer Vision and Pattern Recognition (2006)
16. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)

17. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE T. Pattern Anal. 32(9), 1582–1596 (2010)
18. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision 3(3), 177–280 (2007)
19. Tuzel, O., Porikli, F., Meer, P.: Pedestrian Detection via Classification on Riemannian Manifolds. IEEE T. Pattern Anal. 30(10), 1713–1727 (2008)
20. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008), http://www.vlfeat.org/
21. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR (2010)