Medical Image Retrieval using a Bag of Meaningful Visual Words

Unsupervised visual vocabulary pruning with PLSA

Antonio Foncubierta-Rodríguez*
University of Applied Sciences
Western Switzerland
TechnoArk 3
3960 Sierre, Switzerland
antonio.foncubierta@hevs.ch

Alba García Seco de Herrera
University of Applied Sciences
Western Switzerland
TechnoArk 3
3960 Sierre, Switzerland
alba.garcia@hevs.ch

Henning Müller
University of Applied Sciences
Western Switzerland
TechnoArk 3
3960 Sierre, Switzerland
henning.mueller@hevs.ch

ABSTRACT

Content–based medical image retrieval has been proposed as a technique that allows not only for easy access to images from the relevant literature and electronic health records but also for training physicians, for research and clinical decision support. The bag–of–visual–words approach is a widely used technique that tries to shorten the semantic gap by learning meaningful features from the dataset and describing documents and images in terms of the histogram of these features. Visual vocabularies are often redundant, over–complete and noisy. Larger than required vocabularies lead to high–dimensional feature spaces, which present important disadvantages with the curse of dimensionality and computational cost being the most obvious ones. In this work a visual vocabulary pruning technique is presented. It enormously reduces the amount of required words to describe a medical image dataset with no significant effect on the accuracy. Results show that a reduction of up to 90% can be achieved without impact on the system performance. Obtaining a more compact representation of a document enables multimodal description as well as using classifiers requiring low–dimensional representations.

Categories and Subject Descriptors

I.4.8 [**Computing Methodologies**]: Image Processing and Computer Vision; H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*.

Keywords

Bag of visual words, language modelling, medical image retrieval

*Corresponding author.

ACM MM MIIRH 2013 Barcelona, SPAIN

http://dx.doi.org/10.1145/2505323.2505336.

1. INTRODUCTION

Image retrieval and image classification have been extremely active research domains with hundreds of publications in the past 20 years [1, 2, 3]. Content–based image retrieval has been proposed for diagnosis aid, decision support and enabling similarity–based easy access to medical information [4, 5].

One of the main domains of image retrieval has been the medical literature with millions of images being available [6, 7]. ImageCLEFmed[1], an annual evaluation campaign on retrieval of images from the biomedical open access literature [8]. In the ImageCLEF medical task, 12–17 teams compare their approaches each year on a variety of search tasks.

The Bag–of–Visual–Words (BoVW) is a visual description technique that aims at shortening the semantic gap by partitioning a low–level feature space into regions of the features space that potentially correspond to visual concepts. These regions are called visual words in an analogy to text–based retrieval and the bag of words approach. An image can be described by assigning a visual word to each of the feature vectors that describe local regions of the images (either via a dense grid sampling or interest points), and then representing the set of feature vectors by a histogram of the visual words. One of the most interesting characteristics of the BoVWs is that the set of visual words is created based on the actual data and therefore only concepts present in the data will be part of the visual vocabulary [9].

The creation of the vocabulary is normally based on a clustering method (e.g. k–means, DENCLUE) to identify local clusters in the feature space and then assigning a visual word to each of the cluster centers. This has been investigated previously, either by searching for the optimal number of visual words [10], by using various clustering algorithms [11] instead of the k–means or by selecting interest points to obtain the features [12].

Although the BoVW is widely used in the literature [13, 14] there is a strong performance variation within similar experiments when considering different vocabulary sizes [10]. In this paper, we hypothesize that this variance of the BoVW method is strongly related to the quality of the vocabulary used, understanding quality as the ability of the vocabulary

[1] `http://www.imageclef.org/`

to accurately describe useful concepts for the task. Therefore, we try to reduce the size of the vocabulary without reducing the performance of the method. The use of supervised clustering [15, 16] to force the clusters to a known number of classes was also considered as an option but it is against the notion of learning a variety of concepts present in the data. Instead, we compute the latent semantic concepts in the dataset in an unsupervised way by analyzing the probability of each word to occur. This allows to extract concepts from a combination of various visual word types, since the concepts are discovered based on the probability of co–occurrence of a set of visual words regardless of their origin. The resulting reduced vocabularies present two benefits over the full ones. First, a reduction of the descriptors leads to reduction of the computational cost of the online phase of retrieval but also in the offline indexing phase. This reduction becomes important in the context of large–scale databases or *Big Data*. The second benefit of the approach is that by removing non–meaningful visual words, the dataset description becomes more compact. A compact representation makes it easier to use neighbourhood–based classifiers, which tend to fail in high dimensional feature spaces due to the curse of dimensionality.

The rest of the paper is organized as follows: Section 2 explains in details the materials and methods used with focus on the data set, the probabilistic latent semantic analysis and how it is used to remove meaningless visual words from the vocabulary. Section 3 contains factual details of results of the experiments run on the dataset, while Section 4 discusses them. Conclusions and future work are explained in Section 5.

2. MATERIALS AND METHODS

In this section, further details on the data set and the techniques employed are given.

2.1 Data set

Image modality is one of the characteristics of medical image retrieval that practitioners would like to see included in existing systems [17]. Medical image search engines such as GoldMiner[2] and Yottalook[3] contain modality filters to improve retrieval results. Whereas DICOM headers often contain metadata that can be used to filter modalities, this information is lost when exporting images for publication in journals or conferences where images are stored as JPG, GIF or PNG files. In this case visual appearance is key to identify modalities or the caption text can be analyzed for respective keywords to identify modalities. The ImageCLEFmed evaluation campaign contains a modality classification task that is regarded as an essential part for image retrieval systems. In 2012, the modality classification data set contained 2,000 images from the medical literature organized in a hierarchy of 31 categories [18]. Figure 1 shows the hierarchical structure of modalities. All images in the dataset belong to a single leaf node in the hierarchy.

The modality classification dataset is divided into two subsets of 1,000 images each, one for training and one for testing. The training set and its corresponding ground truth are made public for the groups to train and optimize their methods but the comparison is performed on a test set of

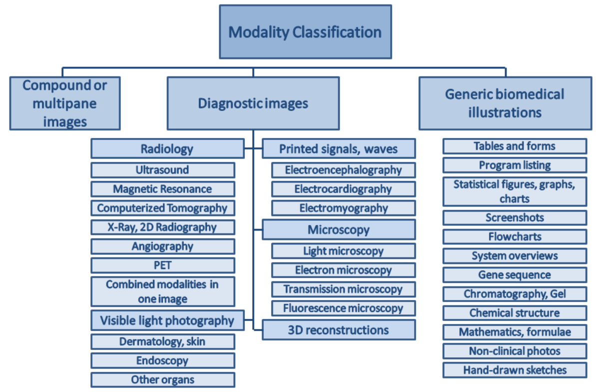


Figure 1: Hierarchy of modalities or image types considered in the modality classification task.

which the ground truth is not known by the groups. Figure 2 shows the distribution of images across modalities in the training and test sets.

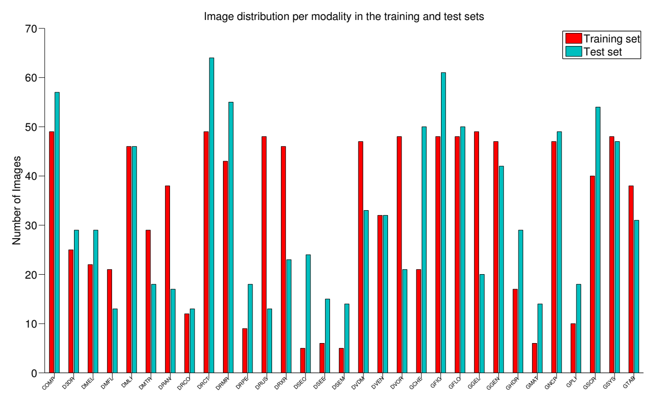


Figure 2: Distribution of images across modalities for the modality classification training and test sets.

Besides modality classification, an image retrieval task is also performed during the benchmarking event where independent assessors judge the relevance of each document in the pool of results submitted by the groups. The retrieval task is performed on a dataset containing the full ImageCLEFmed data set, which in 2012 consisted of more than 306,000 images.

Both data sets were used in the experiments described in this article. Methods were first tested on the modality classification data set (training and testing) to investigate the effect of parameters on the system. Then, fewer parameter combinations were tested on the retrieval task with a larger data base.

2.2 Descriptors

In this section, the descriptors used in our experimental evaluation are presented. Scale Invariant Feature Transform (SIFT) and Bag–of–Colors (BoC) were chosen as images descriptors.

[2] http://goldminer.arrs.org/
[3] http://www.yottalook.com/

2.2.1 SIFT

In this work, images are described with a BoVW based on their SIFT [19] descriptors. This representation has been commonly used for image retrieval because it can be computed efficiently [14, 20, 21]. The SIFT descriptor is invariant to translations, rotations and scaling transformations and robust to moderate perspective transformations and illumination variations. SIFT encodes the salient aspects of the greylevel–images gradient in a local neighbourhood around each interest point.

2.2.2 Bag of Colors

BoC is used to extract a color signature from the images [22]. The method is based on BoVW image representation, which facilitates the fusion with the SIFT–BoVW descriptor. The CIELab[4] color space was used since it is a perceptually uniform color space [23]. A color vocabulary $\mathcal{C} = \{c_1, \dots, c_{100}\}$, with $c_i = (L_i, a_i, b_i) \in CIELab$, is defined by automatically clustering the most frequently occurring colors in the images of a subset of the collection containing an equal number of images from the various classes.

The BoC of an image I is defined as a vector $BoC = \{\bar{c}_1, \dots, \bar{c}_{100}\}$ such that, for each pixel $p_k \in I$:

$$\bar{c}_i = \sum_{k=1}^{P} \sum_{j=1}^{P} g_j(p_k)$$

with P the number of pixels in the image I, where

$$g_j(p) = \begin{cases} 1 \; if \; d(p, c_j) \leq d(p, c_l) \\ 0 \; otherwise \end{cases} \quad (1)$$

and $d(x, y)$ is the Euclidean distance between x and y.

2.3 Vocabulary pruning using probabilistic latent semantic analysis

2.3.1 Probabilistic latent semantic analysis

Visual words are often referred to as an extension of the bag of words technique used in information retrieval from textual to visual data. Similarly, language modelling techniques have also been extended from text to visual words–based techniques [24, 25].

Latent Semantic Analysis (LSA) [26] is a language modelling technique that maps documents to a vector space of reduced dimensionality, called *latent semantic space*, based on a Singular Value Decomposition (SVD) of the terms–documents co–ocurrence matrix. This technique was later extended to statistical models, called *Probabilistic* Latent Semantic Analysis (PLSA), by Hofmann [27]. PLSA removes restrictions of the purely algebraic former approach (namely, the linearity of the mapping).

Hofmann defines a generative model that states that the observed probability of a word or term $w_j, j \in 1, \dots, M$ occurring in a given document $d_i, i \in 1, \dots, N$, is linked to a latent or unobserved set of concepts $\mathcal{Z} = \{z_1, \dots, z_K\}$ that happen in the text:

$$P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k) P(z_k|d_i). \quad (2)$$

[4]CIELab is a color space defined by the International Commission on Illumination (Commission Internationale de l'Éclairage) describing all colors visible for humans while trying to mimic the nonlinear response of the eye.

The model is fit via the EM (Expectation–Maximization) algorithm. For the expectation step:

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|d_i)}. \quad (3)$$

and for the maximization step:

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, w_m) P(z_k|d_i, w_m)}, \quad (4)$$

$$P(z_k, d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z_k|d_i, w_j)}{n(d_i)}. \quad (5)$$

where $n(d_i, w_j)$ denotes the number of times the term w_j occurred in document d_i; and $n(d_i) = \sum_j (d_i, w_j)$ refers to the document length.

These steps are repeated until convergence or until a termination condition is met. As a result, two probability matrices are obtained: the word–concept probability matrix $W_{M \times K} = (P(w_j|z_k))_{j,k}$ and the concept–document probability matrix $D_{K \times N} = (P(z_k|d_i))_{k,i}$.

2.3.2 PLSA for visual words

The PLSA technique only requires a word–document co–occurrence matrix and therefore the technique can be referred to as feature–agnostic. Since it does not set any requirements on the nature of the low level features that yield these co–occurence matrices (other than being discrete), the extension to visual words is simple. PLSA in combination with visual words for classification purposes was also applied in [28, 29].

In our approach, images are described in terms of a BoC in the CIELab color space and a BoVW based on SIFT descriptors. Therefore, the dataset can be described using the following co–occurrence matrices:

$$C_{N \times N_C} = (n(d_i, c_j))_{i,j}, \quad (6)$$

$$S_{N \times N_S} = (n(d_i, s_l))_{i,l}, \quad (7)$$

where N is the number of images in the dataset, N_C the length of the color vocabulary, N_S the length of the SIFT–based vocabulary and $n(d_i, c_j)$ or $n(d_i, s_l)$ is the number of occurrences of the color word c_j or SIFT word s_l occurring in the image d_i.

2.3.3 Vocabulary pruning

The key idea in our approach is that not only the color and SIFT vocabularies are over–complete and redundant individually for the dataset, but they may as well contain visual words that model the same latent concepts. Therefore, a full color–SIFT representation of the dataset is obtained by concatenating the two matrices C and S into a single $N \times (N_C + N_S)$ visual features matrix V.

The matrix V is then analysed using the PLSA technique with a varying number of concepts K and the resulting visual word–concept conditional probability matrices $W_{(N_C+N_S) \times K}$ are used to find the meaningless visual words that need to be removed from the vocabulary.

A visual word is considered meaningless if its conditional probability is below the *significance threshold* T_k for every latent concept. Since each concept can be linked to a different number of visual words, the significance threshold is not an absolute value, but relative to each concept. In our approach, T_k takes the value of the p_T–th percentile of each concept. This allows to keep only the $(100 - p_T)\%$ most significative visual words for each concept while removing the remaining visual words. A visual word can be significative for several concepts (*polysemic words*) and several visual words can be equally significative for a given concept (*synonyms*). These factors, which are common in language modelling, have as a result that the vocabulary reduction cannot be estimated directly using the value of p_T, since it depends on the distribution of synonyms and polysemic words in the experimental data model.

The number of latent concepts as well as the value of the significant percentile are parameters of the technique presented in this paper. Section 3 explains the results of the experimental evaluation of the technique for various values of K and p_T.

2.4 Experiments

Several experiments were run to evaluate the performance of the vocabulary pruning technique. In this section, the experiments are described.

2.4.1 *Classification with a truncated descriptor*

Preliminary experiments on the vocabulary pruning technique over the training set were based on removing meaningless visual words from the descriptors but not from the vocabulary (i.e. the histogram values for meaningful visual words remain the same and therefore histograms are no longer normalized).

By running a 2–fold cross validation on the modality classification training set, the effect of the parameters K (number of latent concepts) and p_T (significant percentile threshold) was investigated. All descriptors were computed using the full vocabulary and visual words below the significance threshold were later removed from the descriptors. No fusion rules were applied to the SIFT–BoVW and BoC descriptors.

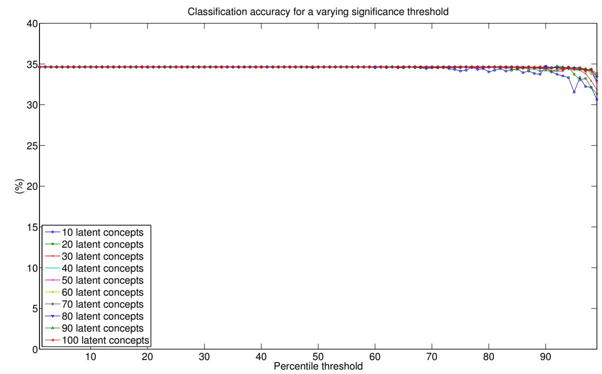
2.4.2 *Classification with a reduced vocabulary*

In this experiment, meaningless visual words were removed from the vocabulary, histograms were recomputed and therefore stayed normalized. Due to the presence of very unbalanced classes in the dataset, experiments included 2–fold cross–validation on the training set and cross–validation based on separate training and test set. The same experiments were run with the full vocabularies.

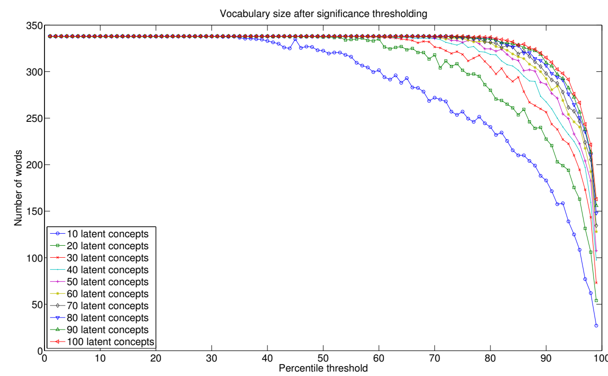
Classification using the SIFT–BoVW and BoC can benefit from a fusion technique to include color and texture information. The similarity scores were calculated using both descriptors separately and the CombMNZ fusion rule [30] was used to obtain final scores. Images were classified using a weighted k–NN (k–Nearest Neighbors) voting [31]. Experiments were run with various k values for the voting.

2.4.3 *Retrieval with a reduced vocabulary over the complete data set*

In this experiment, the complete ImageCLEF dataset for medical images was indexed for retrieval. The number of images in the dataset (306,000) is sufficiently large to allow measures on speed gain when reducing the vocabulary. Retrieval was performed using the fusion rule described in Section 2.4.2. The retrieval experiment consisted of 22 topics (each consisting of 1 to 7 query images), corresponding to the ImageCLEF 2012 medical track.



(a) Effect on classification accuracy.



(b) Effect on effective vocabulary size.

Figure 3: Evaluation of descriptor truncation over the modality classification training set using cross–validation. 1–NN classification was performed for a varying number of latent concepts K and significant percentile p_T

3. RESULTS

In this section a summary of the results for each experiment is given.

3.1 Truncated descriptor

This section explains the results of the experiment described in Section 2.4.1. Since the descriptor requires the full vocabulary before performing the truncation of meaningless words no speed gain in the offline phase was obtained.

Figure 3(a) shows the results of the accuracy obtained using a 1–NN classifier compared to the effect of truncating descriptors on vocabulary size in Figure 3(b). The number of latent concepts K varies from 10 to 100 in steps of 10 and the significant percentile threshold for each concept p_T from 1 to 99.

The effect of increasing the significant percentile is much stronger on the number of visual words used than on the classification accuracy. Similarly, the number of latent concepts has a limited impact on accuracy while having a strong impact on the vocabulary size. Rather unsurprisingly, the fewer latent concepts considered, the easier it becomes to find meaningless visual words. Also, vocabulary sizes tend to be more similar for various K values when p_T is high.

Statistical significance tests were run to compare the results distributions using the truncated descriptors. These tests failed to show a statistically significant difference between classification using the full descriptor or any of the reduced descriptors over the training set.

3.2 Reduced vocabulary over modality classification training and test sets

This section contains a summary of the results of the experiments described in Section 2.4.2.

Table 1 contains a summary of the best results for a significant percentile $p_T = 80$ and a varying number of concepts. It also includes the results obtained with the full vocabulary using the same classifier. Although it is not shown in the table, all of the removed words for $p_T = 80$ belonged to the SIFT–BoVW vocabulary.

Latent Concepts	Removed words	Accuracy (reduced vocabulary)	Accuracy (complete vocabulary)
10	**27.22%**	**44.20%**	43.79%
20	**17.16%**	**44.20%**	43.79%
30	**6.8%**	**43.99%**	43.79%
40	**3.25%**	**43.79%**	43.79%
50	**2.96%**	**43.99%**	43.79%
60	**2.07%**	**43.99%**	43.79%
70	**1.18%**	**43.79%**	43.79%
80	**0.59%**	**43.79%**	43.79%
90	**0.59%**	**43.79%**	43.79%
100	**0.3%**	**43.79%**	43.79%

Table 1: Best classification results (varying the k–NN voting) over the training set for varying number of latent concepts and a fixed significant percentile $p_T = 80$. The last column contains the accuracy when using the complete vocabulary with the same classifier. Results are shown in bold when a reduced vocabulary produces better or equal classification than the complete vocabulary.

Table 2 contains the corresponding results for a 99–percentile as significance threshold. In this experiment meaningless words were found in both the BoC and the SIFT–BoVW vocabularies.

Latent Concepts	Removed words	Accuracy (reduced vocabulary)	Accuracy (complete vocabulary)
10	**91.72%**	**41.55%**	41.34%
20	**84.32%**	**44.20%**	43.18%
30	**78.99%**	**43.79%**	42.16%
40	**72.78%**	**45.01%**	41.34%
50	**67.75%**	**44.81%**	42.16%
60	**61.83%**	**44.60%**	42.97%
70	**59.47%**	**43.81%**	42.97%
80	**54.73%**	**45.62%**	42.97%
90	**53.85%**	**43.99%**	42.97%
100	**50%**	**43.79%**	42.97%

Table 2: Best classification results (varying the k–NN voting) over the training set for varying number of latent concepts and a fixed significant percentile $p_T = 99$. The last column contains the accuracy when using the complete vocabulary with the same classifier. Results are shown in bold when a reduced vocabulary produces better or equal classification than the complete vocabulary.

Tables 3 and 4 contain the corresponding results over the test set when performing cross–validation with separate test and training sets. The vocabularies used are the same as those from Tables 1 and 2.

Latent Concepts	Accuracy (reduced vocabulary)	Accuracy (complete vocabulary)
10	**40.14%**	38.94%
20	**39.24%**	38.94%
30	**39.54%**	38.64%
40	**39.24%**	38.24%
50	**39.34%**	38.94%
60	**39.24%**	38.94%
70	**39.24%**	38.94%
80	**39.24%**	38.94%
90	**39.24%**	38.94%
100	**39.24%**	38.94%

Table 3: Best classification results (varying the k–NN voting) over the test set for varying number of latent concepts and a fixed significant percentile $p_T = 80$. The last column contains the accuracy when using the complete vocabulary with the same classifier. Results are shown in bold when a reduced vocabulary produces better or equal classification than the complete vocabulary.

Latent Concepts	Accuracy (reduced vocabulary)	Accuracy (complete vocabulary)
10	36.44%	37.94%
20	36.24%	37.94%
30	36.84%	38.64%
40	38.44%	38.94%
50	37.24%	38.64%
60	37.34%	38.94%
70	**38.94%**	38.94%
80	37.94%	38.94%
90	**38.94%**	38.94%
100	**39.44%**	38.94%

Table 4: Best classification results (varying the k–NN voting) over the test set for a varying number of latent concepts and a fixed significant percentile $p_T = 99$. The last column contains the accuracy when using the complete vocabulary with the same classifier. Results are shown in bold when a reduced vocabulary produces better or equal classification than the complete vocabulary.

3.3 Reduced vocabulary for the retrieval task

Based on the results in Section 3.2, two vocabularies were selected for obtaining results in the ImageCLEFmed retrieval task. The smallest vocabulary corresponds to the $p_T = 99$ and 10 latent concepts vocabulary, whereas the most accurate vocabulary was the $p_T = 80$ and 10 latent concepts.

Table 5 contains a summary of the results in terms of time required for indexing the complete dataset for the most accurate configuration ($p_T = 80$ and 10 latent concepts) , the smallest vocabulary ($p_T = 99$ and 10 latent concepts) and the complete vocabulary.

Table 6 shows the results when performing the retrieval task on the complete ImageCLEFmed 2012 dataset with the selected vocabularies for each of the 22 topics or queries.

(a) Average time per image for the reduced vocabulary with parameters $p_T = 99$ and $K = 10$.

Feature type	Index time	Size
BoC	2.14 s	19 words
SIFT–BoVW	0.74 s	9 words

(b) Average time per image for the reduced vocabulary with parameters $p_T = 80$ and $K = 10$.

Feature type	Index time	Size
BoC	4.86 s	100 words
SIFT–BoVW	1.15 s	146 words

(c) Average time per image for the complete vocabulary.

Feature type	Index time	Size
BoC	4.86 s	100 words
SIFT–BoVW	1.67 s	238 words

Table 5: Average indexing time per image for the smallest vocabulary, the most accurate and the complete vocabulary.

4. DISCUSSION

As shown in Figure 3 the impact of PLSA–based pruning has a stronger effect on the size of the vocabulary than on the performance of the classifiers. Table 2 shows that a vocabulary reduction of up to 91.72% can be obtained with a comparable accuracy for the same classifier. For the 99–percentile value, the best classification method with the reduced vocabulary always obtains higher accuracy than the same classification method on the full vocabulary.

However, significance tests have failed to show a statistically significant difference between the various accuracy results obtained. Therefore, the main contribution of this work is a method that can enormously reduce visual word vocabularies while obtaining a comparable (and often slightly higher) accuracy.

Another important aspect of the results is that the PLSA–based pruning finds a more meaningful vocabulary than the SIFT–BoVW one. Whereas in the complete vocabulary the SIFT–based words outnumbered the color words by a factor of 2.38, this relationship is inverted in the smallest vocabulary where there are more than two color words for each SIFT–based word.

Results in Table 5 show that the reduction of the indexing time is smaller than the reduction in the number of words. However, the smallest vocabulary presents an indexing time 55.9% lower than the complete vocabulary. Studies have shown that the reduction of the number of features used as a descriptor can increase the speed of online retrieval [32]. This is confirmed in Table 5(c), with retrieval times up to 64% lower when using the smallest vocabulary.

Results in Tables 1 to 4 show that the performance is much better for modality classification tasks than for retrieval in the complete ImageCLEFmed dataset (see Table 6), probably due to the size of the training set used (1000 images) in comparison with the 306000 images in the complete dataset. For the retrieval task, the vocabularies present a comparable performance in terms of recall, being the $p_T = 80$, $K = 10$ vocabulary slightly better than the others. However, mean

(a) Retrieval results for each vocabulary and various queries. Results with higher recall are shown in bold.

	Relevant items	Items retrieved (complete vocabulary)	Items retrieved ($p_T = 80$, $K = 10$)	Items retrieved ($p_T = 99$, $K = 10$)
Topic 1	21	7	**8**	**8**
Topic 2	33	**21**	20	16
Topic 3	47	**35**	**35**	29
Topic 4	22	15	**16**	15
Topic 5	58	**7**	**7**	4
Topic 6	13	7	7	**8**
Topic 7	11	2	2	**3**
Topic 8	6	**3**	**3**	2
Topic 9	2	0	0	0
Topic 10	17	**6**	**6**	**6**
Topic 11	72	17	**19**	8
Topic 12	27	5	6	**9**
Topic 13	147	**50**	48	38
Topic 14	521	**57**	56	48
Topic 15	0	0	0	0
Topic 16	3	**1**	**1**	**1**
Topic 17	7	0	0	**2**
Topic 18	4	0	0	0
Topic 19	6	**3**	**3**	2
Topic 20	5	0	0	0
Topic 21	49	5	5	**7**
Topic 22	19	**7**	**7**	5
Total	1090	248	**249**	211

(b) Mean Average Precision (MAP) across all topics.

Vocabulary used	MAP
Complete vocabulary	6.51%
$p_T = 80$, $K = 10$	6.52%
$p_T = 99$, $K = 10$	1.51%

(c) Average execution times of the online phase for a single query image.

Vocabulary used	Online retrieval time
Complete vocabulary	125 s
$p_T = 80$, $K = 10$	107 s
$p_T = 99$, $K = 10$	45 s

Table 6: Results of retrieval experiments for each vocabulary.

average precision strongly varies between large vocabularies and the smallest vocabulary ($p_T = 99$, $K = 10$).

It can be discussed that de benefits of the PLSA–based pruning presented in this paper are not the ability to discover new and meaningful visual words for retrieval but the ability to recognize those visual words that convey most of the meaning among those present in the vocabulary.

5. CONCLUSIONS AND FUTURE WORK

In this paper a vocabulary pruning method based on probabilistic latent semantic analysis of visual words for medical image retrieval and classification is presented. The selection of optimal visual words is performed by removing visual words with a conditional probability over all learnt latent concepts that is below a given threshold. The vocabulary pruning process is completely unsupervised, since the learning of the concepts is performed without taking into consideration the number of classes or what is the actual class assigned to each image. Therefore, it can be used to reduce massive fine–grained vocabularies to smaller vocabularies that contain only the most meaningful visual words even before training the classifier. To obtain these fine–grained vocabularies, simple clustering algorithms can be used to produce a large number of small clusters that later will be pruned using the methods explained in this paper. Smaller clusters are supposed to encode subtle visual differences among images, which will be preserved by the PLSA–based pruning if they are meaningful for some latent concept. Future applications of the technique also include the use of multiple vocabularies that can be merged and pruned as a single set of discrete features.

We are currently extending the techniques to images obtained for clinical use, where the use of low–dimensional descriptors can achieve fast and accurate characterization of large–scale datasets of high–dimensional (3D, 4D, multimodal) images. This is expected to lead to different results as for the modality classification tasks and retrieval tasks from the literature color plays a more important roles than for most clinical images. Still, the possibility to reduce visual vocabularies strongly can lead to larger base vocabularies that can potentially capture the image content much better but can then be reduced for efficient retrieval.

6. ACKNOWLEDGMENTS

This work was partially supported by the Swiss National Science Foundation (FNS) in the MANY2 project (205320–141300), the EU 7^{th} Framework Program under grant agreements 257528 (KHRESMOI) and 258191 (PROMISE).

7. REFERENCES

[1] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content–based image retrieval systems in medicine–clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23, 2004.

[2] Ceyhun Akgül, Daniel Rubin, Sandy Napel, Christopher Beaulieu, Hayit Greenspan, and Burak Acar. Content–based image retrieval in radiology: Current status and future directions. *Journal of Digital Imaging*, 24(2):208–222, 2011.

[3] Lilian H. Y. Tang, R. Hanka, and H. H. S. Ip. A review of intelligent content–based indexing and browsing of medical images. *HIJ*, 5:40–49, 1999.

[4] Dina Demner-Fushman, Sameer Antani, Mohammad-Reza Siadat, Hamid Soltanian-Zadeh, Farshad Fotouhi, and Kost Elisevich. Automatically finding images for clinical decision support. In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, ICDMW '07, pages 139–144, Washington, DC, USA, 2007. IEEE Computer Society.

[5] Barbara Caputo, Henning Müller, Tanveer Syeda Mahmood, Jayashree Kalpathy-Cramer, Fei Wang, and James Duncan. Editorial of miccai workshop proceedings on medical content–based retrieval for clinical decision support. In *Proceedings on MICCAI Workshop on Medical Content–based Retrieval for Clinical Decision Support*, volume 5853 of *Lecture Notes in Computer Science (LNCS)*. Springer, 2009.

[6] Henning Müller, Jayashree Kalpathy-Cramer, Charles E. Kahn Jr., and William Hersh. Comparing the quality of accessing the medical literature using content–based visual and textual information retrieval. In *SPIE Medical Imaging*, volume 7264, pages 1–11, Orlando, Florida, USA, February 2009.

[7] Thomas M. Deserno, Sameer Antani, and L. Rodney Long. Content–based image retrieval for scientific literature access. *Methods of Information In Medicine*, 48(4):371–380, July 2009.

[8] Henning Müller, Alba García Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner Fushman, Sameer Antani, and Ivan Eggel. Overview of the imageclef 2012 medical image retrieval and classiïňĄcation tasks. In *Working Notes of CLEF 2012 (Cross Language Evaluation Forum)*, September 2012.

[9] Leibe Bastian Grauman, Kristen and. *Visual Object Recognition*. 2011.

[10] Antonio Foncubierta-Rodríguez, Adrien Depeursinge, and Henning Müller. Using multiscale visual words for lung texture classification and retrieval. In Hayit Greenspan, Henning Müller, and Tanveer Syeda Mahmood, editors, *Medical Content–based Retrieval for Clinical Decision Support*, volume 7075 of *MCBR–CDS 2011*, pages 69–79. Lecture Notes in Computer Sciences (LNCS), September 2012.

[11] Alexander Hinneburg and Hans-Henning Gabriel. DENCLUE 2.0: Fast clustering based on kernel density estimation. *Advances in Intelligent Data Analysis VII*, 4723/2007:70–80, 2007.

[12] Sebastian Haas, Rene Donner, Andreas Burner, Markus Holzer, and Georg Langs. Superpixel–based interest points for effective bags of visual words medical image retrieval. In Hayit Greenspan, Henning Müller, and Tanveer Syeda-Mahmood, editors, *Medical Content-based Retrieval for Clinical Decision Support*, volume 7075 of *MCBR-CDS 2011*. Lecture Notes in Computer Sciences (LNCS), September 2011.

[13] Uri Avni, Hayit Greenspan, Eli Konen, Michal Sharon, and Jacob Goldberger. X–ray categorization and retrieval on the organ and pathology level, using patch–based visual words. *IEEE Transactions on Medical Imaging*, 30(3):733–746, 2011.

[14] Dimitrios Markonis, Alba García Seco de Herrera, Ivan Eggel, and Henning Müller. Multi–scale visual words for hierarchical medical image categorization. In *SPIE Medical Imaging 2012: Advanced PACS–based Imaging Informatics and Therapeutic Applications*, volume 8319, pages 83190F–11, February 2012.

[15] Sugato Basu, Arindan Banerjee, and Raymond Mooney. Semi–supervised clustering by seeding. In *19th Internaional Conference on Machin Learning (ICML-2002)*, pages 19–26, July 2002.

[16] Mikhail Bilenko, Sugato Basu, and Raymond Mooney. Integrating constraints and metric larning in semi–supervised clustering. In *21t Internaional Conference on Machine Learning (ICML-2004)*, July 2004.

[17] Dimitrios Markonis, Markus Holzer, Sebastian Dung, Alejandro Vargas, Georg Langs, Sascha Kriewel, and Henning Müller. A survey on visual information search behavior and requirements of radiologists. *Methods of Information in Medicine*, 51(6):539–548, 2012.

[18] Henning Müller, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, and Sameer Antani. Creating a

classification of image types in the medical literature for visual categorization. In *SPIE medical imaging*, 2012.

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

[20] Yi Yang and Shawn Newsam. Bag–of–visual–words and spatial extensions for land–use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 270–279, New York, NY, USA, 2010. ACM.

[21] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, volume 2, pages 506–513, Washington, DC, USA, June 2004.

[22] Christian Wengert, Matthijs Douze, and Hervé Jégou. Bag–of–colors for improved image search. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 1437–1440, New York, NY, USA, 2011. ACM.

[23] M.Sheerin Banu and Krishnan Nallaperumal. Analysis of color feature extraction techniques for pathology image retrieval system. IEEE, 2010.

[24] Pierre Tirilly, Vincent Claveau, and Patrick Gros. Language modeling for bag-of-visual words image categorization. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 249–258. ACM, 2008.

[25] Qi Tian, Shiliang Zhang, Wengang Zhou, Rongrong Ji, Bingbing Ni, and Nicu Sebe. Building descriptive and discriminative visual codebook for large-scale image applications. *Multimedia Tools and Applications*, 51(2):441–477, 2011.

[26] Scott Deerwester, Susan T. Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[27] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.

[28] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Scene classification via plsa. In *Computer Vision–ECCV 2006*, pages 517–530. Springer, 2006.

[29] Ismail El sayad, Jean Martinet, Thierry Urruty, and Chabane Djeraba. Toward a higher–level visual representation for content–based image retrieval. *Multimedia Tools and Applications*, 60(2):455–482, 2012.

[30] Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. In *Text REtrieval Conference*, pages 243–252, 1993.

[31] David J. Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining (Adaptive Computation and Machine Learning)*. The MIT Press, 2001.

[32] David McG. Squire, Henning Müller, and Wolfgang Müller. Improving response time by search pruning in a content–based image retrieval system, using inverted file techniques. In *IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL '99)*, pages 45–49, 1999.