Multi–scale visual words for hierarchical medical image categorisation

Dimitrios Markonis, Alba G. Seco de Herrera, Ivan Eggel, Henning Müller

University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland

ABSTRACT

The biomedical literature published regularly has increased strongly in past years and keeping updated even in narrow domains is difficult. Images represent essential information of their articles and can help to quicker browse through large volumes of articles in connection with keyword search. Content–based image retrieval is helping the retrieval of visual content. To facilitate retrieval of visual information, image categorisation can be an important first step. To represent scientific articles visually, medical images need to be separated from general images such as flowcharts or graphs to facilitate browsing, as graphs contain little information. Medical modality classification is a second step to focus search.

The techniques described in this article first classify images into broad categories. In a second step the images are further classified into the exact medical modalities. The system combines the Scale–Invariant Feature Transform (SIFT) and density–based clustering (DENCLUE). Visual words are first created globally to differentiate broad categories and then within each category a new visual vocabulary is created for modality classification. The results show the difficulties to differentiate between some modalities by visual means alone. On the other hand the improvement of the accuracy of the two–step approach shows the usefulness of the method. The system is currently being integrated into the Goldminer image search engine of the ARRS (American Roentgen Ray Society) as a web service, allowing concentrating image search onto clinically relevant images automatically.

Keywords: medical image categorisation, image classification, SIFT, DENCLUE

1. INTRODUCTION

Medical images carry an important part of the information in many biomedical articles. The amount of biomedical articles published in the past years has increased strongly. Many articles have now also become accessible online through open access publishing or when journals make all articles freely available after 12 or 24 months. PubMed Central* for example makes many articles including over a million images available publicly.

To allow for an efficient and effective retrieval of data from articles including images, several tools have been developed such as Springer ImageFinder† or Goldminer‡, which at the end of 2011 indexes over 230'000 radiological images. Most of the tools are text–based, only, but not all information on images can easily be obtained from the text alone. Content–based visual retrieval has been proposed several times[1,2] but quality in performance benchmarks has often shown to be much lower than text retrieval approaches.[3] On the other hand, filtering images based on an automatically extracted imaging modality can increase performance in an important way.[3] For search engines such as Goldminer that concentrate on radiology modalities, already a pre–filtering of all journal figures for radiology modalities based on visual information could be an important advantage. In some journals, particularly radiology journals, the figure captions are controlled and of high quality, so extracting the modality from the text alone is feasible in good quality. In other biomedical journals, the imaging modality can not be obtained from the caption text at all. Visual image classification techniques have other shortcomings as some modalities can easily be mixed up when categorising automatically such as CT (Computed Tomography) and MRI (Magnetic Resonance Imaging). In these cases text information of the captions can be used as additional cue to disambiguate the two.

Further author information: (Send correspondence to Dimitrios Markonis: Dimitrios.Markonis@hevs.ch)

*http://www.ncbi.nlm.nih.gov/pmc/

†http://www.springerimages.com/

‡http://goldminer.arrs.org/

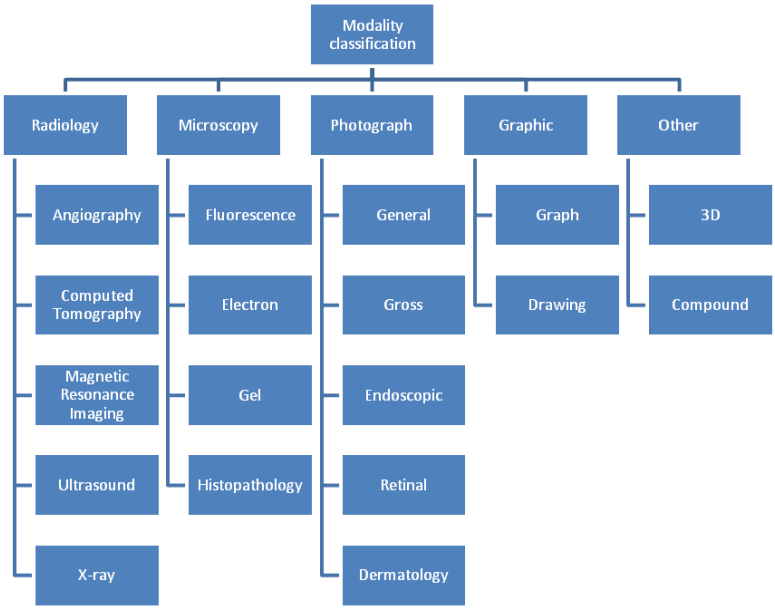


Figure 1: Modality categories of the ImageCLEF 2011 medical modality classification task.

The bag–of–visual–words representation, originating from using text retrieval techniques for image retrieval,[4,5] is a well known approach in content–based image retrieval and object recognition. In this approach, first a training set of local descriptors[6] is clustered and every cluster center is defined as a "visual word" depending on a size of the visual vocabulary. Each local image descriptor is then assigned to the closest cluster or the closest visual word. By doing so, the image is represented as a histogram of visual words[5] and similarity between images is calculated by comparing their histograms using for example a simple histogram intersection. Visual features can be extracted at points of high variance in the images (interest points), using random points in the image or using a dense regular grid of points. In a neighbourhood of these points the visual features are extracted. Using scale–invariant local descriptors as visual words has shown good results in the past.[3] Particularly, the Scale–Invariant Feature Transform[7] (SIFT) has been widely used in object recognition,[8,9] scene categorization,[10,11] concept detection,[12] and scalable image retrieval[10] and classification,[13] as it has proven to be robust and distinctive.[6] The SIFT descriptor is designed to be scale– and rotation–invariant and uses orientation histograms to describe the neighbourhood (4×4 pixels) of the interest point, while ignoring any color information.

A recent trend in content–based image retrieval is to use multiple channels for image representation.[12,14,15] A channel is defined as a combination of an interest point detector (or other sampling strategy) and a local descriptor, possibly accompanied by a schema for spatial information inclusion. However, caution needs to be taken when using such an approach in large image databases, as the dimensionality of the image representation directly affects the retrieval efficiency. At the same time, the size and the distinctiveness of the visual vocabulary is a key point for the performance of the method.

The categorisation system described in this paper is currently being integrated into the Goldminer retrieval system to extract potentially relevant images from biomedical journals and filter out images such as graphs and flow charts that are of little clinical relevance. To do so, a multi–step approach was chosen. First, a classification into broad categories is performed with a global vocabulary of visual words. Then, new, distinctive vocabularies are created based on finer sub–classes to separate the modalities. It is assumed that the vocabularies required to separate broad categories and fine categories have important differences.

This paper is organised as following: Section 2 describes the dataset used for the evaluation, the bag–of–visual–words approach and the density–based clustering method used for the creation of the vocabulary. Moreover, the multi–step approach is described. In Section 3, the results of the evaluation using the ImageCLEF 2011 medical image data set are presented and analysed. Finally, conclusions and future plans are discussed in Section 4.

Table 1: Distribution of training and test images in classes

Modality	# of training images	# of test images
Angiography	11	9
Computed tomography	70	83
Magnetic resonance imaging	17	20
Ultrasound	30	41
X–Ray	59	67
Fluorescence	44	28
Electron microscopy	16	18
Gel	50	50
Histopathology	208	195
General Photo	165	141
Gross pathology	43	32
Endoscopic imaging	10	11
Retinograph	5	3
Dermatology	7	15
Graphs	161	172
Drawing	43	74
3D reconstruction	32	45
Compound figure	17	20

2. METHODS

This section describes the data set and evaluation methodology used in this article. The main techniques and tools reused are also detailed.

2.1 The data set

The evaluation in this article uses a database created in the context of the ImageCLEF[§] 2011 benchmark[16] (Image retrieval task of CLEF, the Cross–Language Evaluation Forum). The database consists of over 230'000 images but for the modality classification 1'000 training and 1'000 test images were made available with modality class labels and only this subset is being used. The training images are used to train the system while the test images are used to validate the system quality. Labels are one of 18 categories including graphs and several radiology modalities (see the hierarchy in Figure 1). The sample images presented in Figure 2 demonstrate the visual diversity of the classes of the data set.

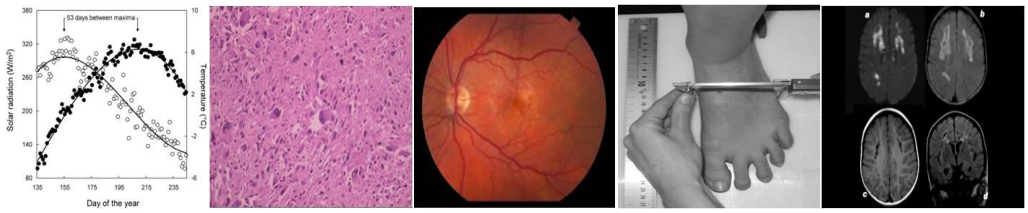
The number of images per class in the training and test sets varies from fewer than ten to several hundred (see Table 1 for the exact numbers). This uneven distribution can affect the training of the classifiers and the resulting performance. Participating groups in the ImageCLEF challenge[17] addressed this challenge and took actions to automatically expand the training set. This inclusion of noise on images to increase the training set may improve the retrieval performance but can also worsen its efficiency. In our study, it was adopted to select a subset of 100 images uniformly distributed across the classes for the creation of the visual vocabulary.

Besides the categories proposed in the benchmark, the system was also evaluated based on broad categories such as radiology vs. all other images and graphs vs. all other images. Radiology images are typically of interest for systems such as Goldminer that concentrate on this and these systems also often try to filter out all graphs, flow charts etc., as they are not clinically relevant images to index.

2.2 Bag–of–visual–words

The pipeline of the bag–of–visual–words approach is shown in Figure 3. A training set of images is chosen and local descriptors (in the case of SIFT, 128–dimensional vectors) are extracted from interest points of each image of this set. The descriptors are then clustered using a clustering method and the centroids of the clusters are

[§]`http://www.imageclef.org/`



(a) Graph (b) Histopathology (c) Endoscopic (d) General Photo (e) Radiology

Figure 2: Sample images from ImageCLEF2011 medical data set

used as visual words. The visual vocabulary represents all cluster centers. Then, the local visual features are also extracted from all other images in the database and mapped to the cluster centers to create for each image a histogram of visual words. Images are thus indexed as histograms of the visual words (bag–of–visual–words) by assigning the nearest visual word to each feature vector.

When a new image is classified, a similarity measure is used to compare the histogram of the new image with the histograms of the training images, providing a similarity score for images in the training set. In order to include spatial information into this representation, several approaches exist. The most used, called spatial pyramid matching[11] partitions the image into increasingly fine sub–regions and creates a histogram of visual word occurences for each sub–region. Then, finer histograms are weighted more to favour matches found in close locations. Recently, an extension of this method was proposed using Fisher kernels instead of histograms of occurences.[18] Spatial reranking has also been used in image retrieval[19] by estimating transformations of query regions to top–ranked retrieved images.

In the study described in this paper, an $n \times n$ partition of the image was used for the extraction of visual features at fixed localizations, mainly for simplicity and efficiency reasons. The image was divided into a grid of $n \times n$ sub–images and a different histogram was computed for each sub–image. As preliminary tests for $n = 2, 3, 4$ showed no improvement of the classification results by this step it was subsequently abandoned. The SIFT implementation existing in the Fiji image processing package[¶] was used for the local feature description of the images, while histogram intersection was applied for similarity calculation.

2.3 Density–based Clustering — DENCLUE

For the creation of the visual vocabulary a common practice is to use K–means as clustering technique. This algorithm is an attractive solution as it uses a single parameter (K — the number of the clusters). However, it also suffers from several drawbacks. First, the divergence of the algorithm is not guaranteed as it can get trapped in local minima. It works well if the data set is a mixture of Gaussian distributions, which is difficult to decide in high dimensional spaces. Moreover, it is not robust to noise and outliers, nor can it detect arbitrarily shaped clusters. K–means has a time complexity of $O(kNd)$ per iteration where N is the data set size and d its dimensionality. While this is acceptable for small vocabularies, the optimal vocabulary size can vary from $O(10^3)$ to $O(10^6)$[20] depending on the image data set. After taking these points into account another approach was followed for the clustering step in this study. A density–based clustering method called DENCLUE[21] was chosen, which is designed to deal with high dimensional data and large data sets. The main idea behind this method is to cluster the data using the local minima of a function that represents their density in the d–dimensional space. Before presenting the algorithm, some definitions of basic concepts are provided.

Definition 1. *A kernel function* $K : \mathbb{R}^d \to \mathbb{R}, K(x) \geq 0$, *which has*

$$\int_{\mathbb{R}^d} K(x)dx = 1$$

[¶] http://fiji.sc/

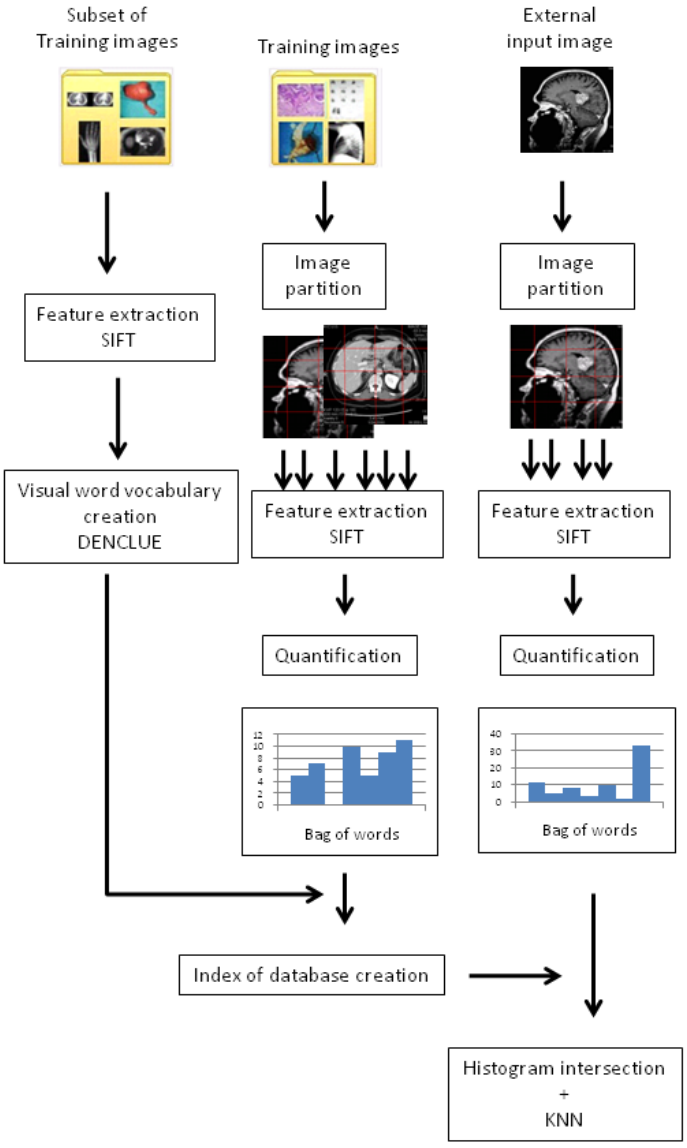


Figure 3: Overview of the one–step approach for the extraction of visual words.

The density function is defined as the sum of the kernels of all data points. Given N data objects described by a set of d–dimensional feature vectors $D = x_1, \ldots, x_N \subset F^d$ the density is defined as

$$f^D(x) = \frac{1}{Nh^d} \sum_{i=1}^{N} K(\frac{1}{h}(x - x_i))$$

DEFINITION 2. *A point $x^* \in F^d$ is called a density attractor of a density function f^D, iff x^* is a local maximum of f^D. A point $x \in F^d$ is density attracted to a density attractor x^* iff a hill–climbing procedure started at x converges to x^*.*

DEFINITION 3. *The local density $\hat{f}^D(x)$ is*

$$\hat{f}^D(x) = \frac{1}{Nh^d} \sum_{x' \in near(x)} K(\frac{1}{h}(x' - x))$$

where $near(x) = x' : dist(x_1, x) \leq \delta_{near}$

DENCLUE consists of two steps. The first step approximates the density function. This can be done by using a cubeMap data structure.[21] After the local density has been computed for each point, the points with $\hat{f}^D < \xi$ where ξ is one of the parameters of the algorithm, are considered noise. For the remaining points, density attractors are determined using a hill–climbing procedure, guided by the local density gradient $\nabla \hat{f}^D(x^i)$.

$$x = x^0, x^{i+1} = x^i + \delta \frac{\nabla \hat{f}^D(x^i)}{\|\nabla \hat{f}^D(x^i)\|}$$

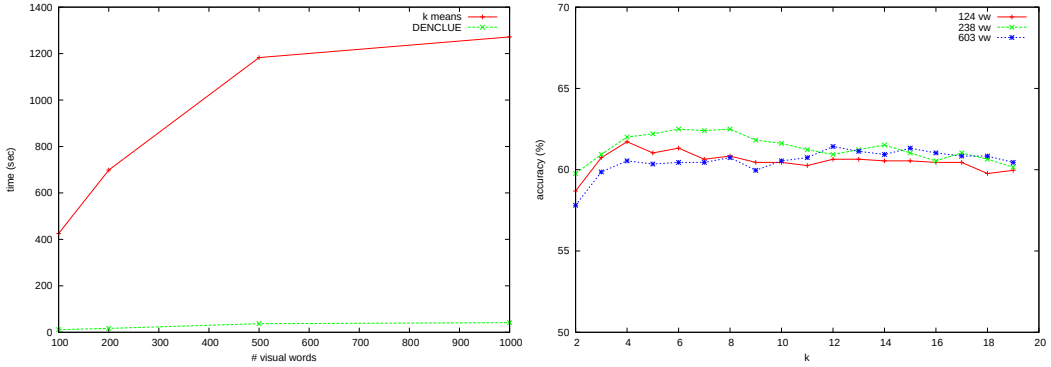
where δ is a parameter that controls the speed of the hill–climbing procedure. This procedure stops at $k \in N$ if $\hat{f}^D(x^{k+1}) < \hat{f}^D(x^k)$ and takes $x^* = x^k$ as a density attractor. When using averaged shifted histograms[22] for the approximation of the local density,[23] the time complexity is $O(d^2 N \log(N))$ for the worst case scenario that every point is in a different hypercube. Note that its complexity is independent of the number of clusters, something that for large vocabularies is an important speed–up. The parameters h, ξ, δ of the method were empirically chosen. However, the speedup that DENCLUE provides allows even a grid search for the optimal parameters. Algorithms to detect the optimal parameters are proposed in.[24,25]

2.4 Hierarchical approach

For the hierarchical visual words approach proposed, six visual vocabularies were created. A global vocabulary, using descriptors from all the classes and the five class–specific vocabularies, each created by clustering descriptors of the training images from one of the broader classes (Radiology, Microscopy, Photograph, Graphic, Other). Every image in the training and test sets is represented in two different indexes (corresponding to the two types of vocabularies), a global index and a category index based on the first step of broad classification. The classification consists of two steps. First, the image to be classified is transformed into a bag–of–visual–words using the global vocabulary. The image is then classified into one broad class by weighted k–NN voting of the global index. Then, a new histogram is created for the image using the assigned category vocabulary and a new classification of the exact classed takes place.

2.5 RESTful web service

A RESTful web service[26] using Java was implemented in order to provide easy access to our image categorization API (Application Programming Interface). Independently of the programming language used by the client, this service can simply be called via an HTTP (HyperText Transfer Protocol) get request that takes an image URL (Uniform Resource Locator) as a parameter. Having the public URL of the image, the server is able to download the image and pass it locally to the modality classification core component for further processing. As soon as the image is classified the server sends back an HTTP response that contains the modality of the image embedded in XML (eXtensible Markup Language). At the same time the XML also specifies if the image is considered a graph and/or a radiology image or not. Below a short example XML message returned by the service is shown:



(a) (b)

Figure 4: (a) Execution times of K–means and DENCLUE clustering for different vocabulary sizes (b) Accuracy using different sizes of vocabularies

```
<Result imageURL="http://api.ning.com/files/6YQud7l0Rdx8LVH2Gn-yzHqubxeTMF-7gnGsWLDE9N213SZBz
rtnDHfVbxSncRqPJDQR-Xsou2pQOdfjrTMX6EJB9XyiUQqV/SplenicInfarctCTAxial.jpg">
  <Graph>false</Graph>
  <Radiology>true</Radiology>
  <Modality>CT</Modality>
</Result>
```

A RESTful client is able to acccess the service via the following URL[‖] where {0} represents a base 64–encoded string of the public image URL. The web service is based on a Glassfish v2.1 application server hosted on an Intel Xeon 2.93GHz machine with 12 cores and 96 GB of RAM that is hosting several other services.

3. RESULTS

To evaluate the efficiency of the DENCLUE clustering the times required for the creation of the visual vocabulary between the DENCLUE and the state–of–the–art method K–means were measured for different vocabulary sizes. The results in terms of efficiency and effectiveness that are shown in Figure 4a for DENCLUE were obtained using a cubeMap structure[21] to compute the local densities. The results show that DENCLUE has a significant speed advantage compared to K–means. In the following experiments for the creation of a medium sized vocabulary of 10'000 visual words out of $\tilde{1}$00'000 descriptors, K–means required more than 22 hours to complete the clusterisation while DENCLUE less than 30 minutes.

The classification accuracy for the same size of visual vocabularies is approximately the same for the two clustering approaches, with DENCLUE having slightly better results for fewer visual words and K–means having better results for larger vocabularies. A possible reason for this is that in the 128–dimensional space of the clustering, highly–dense areas are scarcely found. A way to address this is to either use the averaged shifted histogram[22] for the approximation of the local densities or to simply use a much larger data set. For the described approach only 100 training images were used. While, the whole set of 1'000 training images was too large in size for experiments using K–means for 10'000 visual words, only DENCLUE was tested giving similar performance results to K–means in a much shorter computation time.

Regarding the k–NN voting, k plays a significant role as the images in the ground truth are not distributed uniformly in the classes. Figure 4b shows the classification accuracy for different choices of the number of nearest neighbors. It can be seen that there is no overall optimal choice for k and that the number of nearest neighbors to use also depends on the size of the vocabulary. As expected from the small number of training images that some classes contained, the accuracy declined after approximately k = 12 nearest neighbors. The size of the

[‖] `http://fast.hevs.ch:8080/MedgiftService/resources/modality/imageurl/{0}`

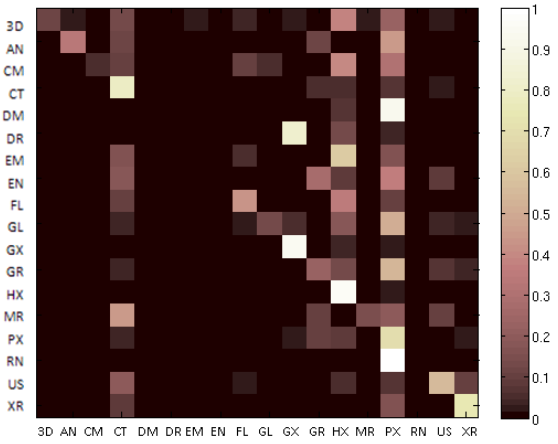


Figure 5: Confusion matrix obtained for the modality classification results.

vocabulary is also an important factor for the classification performance and the system efficiency. The accuracy of the one–step approach for the 18 classes is 62.5% for 238 visual words. The hierarchical approach currently improves this result by 1–2%. Higher accuracies ($\tilde{6}5\%$) were observed for vocabularies of sizes $> 5'000$. This improvement comes, however, with a price as the speed of the online part of the representation of the image to be classified and the retrieval of training images decreases. In the hierarchical approach the same accuracy can be obtained using only one–tenth of the visual words for the image representation (global histogram and intra–class histogram).

Two preliminary classifications that correspond to realistic needs are the graph vs. non–graph and the radiology vs. non–radiology image categorisation. The bag–of–visual–words approach achieves an accuracy of 96.97% and 90.80%, respectively for these two problems. Choosing a uniform (across the classes) training set in order to create the global vocabulary, instead of using the full training set, showed improvement of 0.5% of the overall accuracy. For the two–step approach, for the first step classification a larger value of k is found to perform better, while a smaller value is obtaining better results for the second step. As seen in the confusion matrix in Figure 5, there are certain classes that are frequently misclassified. Computed tomography (CT), histopathology (HX) or general photos (PX) show a lower accuracy. It should be noted, that the images of classes with few training images (e.g. MRIs, Dermatology images, Drawings) tend to be classified to the classes of the same broad category that contain the most training images (e.g. CTs, General Photos, Graphics). This highlights the need for a balanced and particularly a larger training set.

The web service that was developed to implement the classifier described, is currently being integrated into the Goldminer search system. A diagram presenting the steps for the classification of an image is shown in Figure 6.

4. CONCLUSIONS

This paper describes an approach for modality classification using two classification steps with separate visual vocabularies. First, a broad vocabulary is taken to separate the main image types and then a new vocabulary is calculated for each of the broad categories to have a fine–grained classification.

The choice of the DENCLUE algorithm for the creation of the vocabularies is justified by the speed and clustering quality of the algorithm. DENCLUE is highly efficient with large data sets, it handles out–liners and noise well, detects arbitrary shaped clusters and performs well with a data set of high dimensionality as opposed to other density–based clustering methods. Moreover, its input parameters can be tuned to better cluster a specific dataset. This was particularly useful for the creation of the different types of visual vocabularies in this

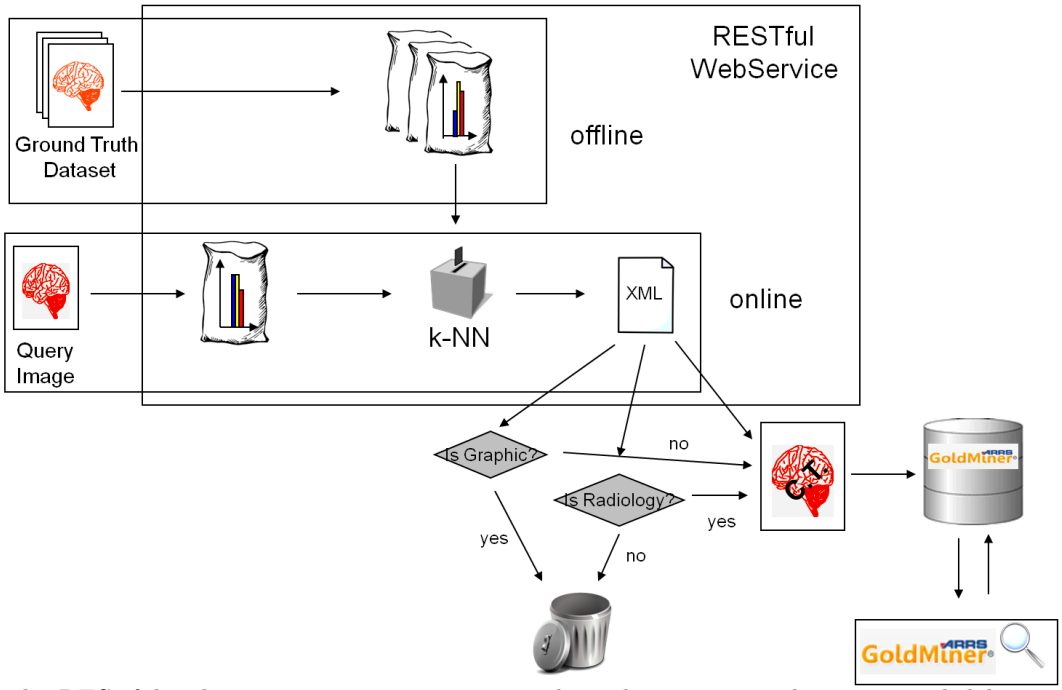


Figure 6: The RESTful web service — given an image, the web service get the image and delivers as a result the modality and also the decision of whether or not a radiology image or a graph was the image type.

paper. The structure that it uses[21] provides useful insight into the characteristics of the data set and can also be used for a more efficient assignment of the visual image descriptors to visual words.

As the system was foreseen to work on large image data sets a relatively small vocabulary was chosen, because it provides good results on the important graphics/non–graphics and radiology/non–radiology classifications. However, larger vocabularies could be used by taking advantage of the already existing DENCLUE structure or a parallel implementation in order to improve classification performance without efficiency costs.

The presented system makes a simple web service available to an image classification system that categorises submitted images into several modality classes. The performance of the system seems sufficient for routine use of analysing very large numbers of images and filter all images potentially relevant for retrieval. Based on the confusion matrix shown in Figure 5, several difficulties were encountered. CT and MRI images are frequently mixed up, which can be caused mainly by their visual proximity. General photos are very difficult to categorise, as a broad variety of images is observed in this class with high intra–class variety. In addition, dermatology, endoscopic and gross pathology images are often confused with general photos. In future work we plan to include color features to identify histopathology images in an easier way and, therefore, to easier distinguish electron microscopy images of them.

(a) MRI (b) CT

Figure 7: MRI and CT images can be difficult to visually discriminate

A larger number of training images than the currently 1000 might also lead to a better performance and such a manual work is currently under way. The fusion of text and content–based retrieval using the captions of the images is another promising field of research. The quality of the captions strongly depends on the type of journal with radiology journals allowing to classify about 80% of the images correctly based on solely the caption information. Fusion techniques could take into account the assumed quality of text and visual information and then combine the two. Taking into account the confusion matrices for visual and caption–based classification can also lead to better quality as visual retrieval is good for some classes whereas other classes such as CT and MRI (Figure 7) are frequently mixed up.

Another important challenge are also compound figures showing in a single image several sub images. These can be of the same modality (such as several CT slices) but can also include a mix of modalities such as PET and CT images next to each other. The cleanest approach would be to separate them and then treat the sub images separately.

Modality classification as a basic technology for image retrieval seems important and can allow to narrow potential results sets or allow for a diversity in the results if text queries are performed. For the given small number of classes a high accuracy could be obtained.

5. ACKNOWLEDGEMENTS

This work was partially funded by the European Union in the context of the KHRESMOI project (grant number 257528), the Swiss National Science Foundation (205321–130046) and the HES–SO.

REFERENCES

1. H. Müller, A. Rosset, A. Garcia, J.-P. Vallée, and A. Geissbuhler, "Benefits from content–based visual data access in radiology," *RadioGraphics* **25**, pp. 849–858, May 2005.
2. H. D. Tagare, C. Jaffe, and J. Duncan, "Medical image databases: A content–based retrieval approach," *Journal of the American Medical Informatics Association* **4**(3), pp. 184–198, 1997.
3. H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, R. Said, B. Bakke, C. E. K. Jr., and W. Hersh, "Overview of the CLEF 2010 medical image retrieval track," in *Working Notes of CLEF 2010 (Cross Language Evaluation Forum)*, September 2010.
4. D. M. Squire, W. Müller, H. Müller, and T. Pun, "Content–based query of image databases: inspirations from text retrieval," *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99)* **21**(13–14), pp. 1193–1198, 2000. B.K. Ersboll, P. Johansen, Eds.
5. J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, *ICCV '03*, pp. 1470–1477, IEEE Computer Society, (Washington, DC, USA), 2003.
6. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis & Machine Intelligence* **27**(10), pp. 1615–1630, 2005.
7. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision* **60**(2), pp. 91–110, 2004.
8. J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," in *Proceedings of Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '06).*, p. 13, June 2006.
9. F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
10. H. Jegou, M. Douze, and C. Schmid, "Aggregating local descriptors into a compact image representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3304 – 3311, June 2010.
11. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, *CVPR '06*, pp. 2169–2178, IEEE Computer Society, (Washington, DC, USA), 2006.

12. K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "The university of amsterdams concept detection system at imageclef 2009," *Lecture Notes in Computer Science* **6242/2010**, pp. 261–268, 2010.
13. D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006* **2**, pp. 2161–2168, June 2006.
14. K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**, pp. 1582 – 1596, September 2010.
15. D. Wang, X. Liu, and L. Luo, "Video diver: generic video indexing with diverse features," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, 2007.
16. J. Kalpathy-Cramer, H. Müller, S. Bedrick, I. Eggel, A. S. de Herrera, and T. Tsikrika, "The CLEF 2011 medical image retrieval and classification tasks," in *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)*, September 2011.
17. G. Csurka, S. Clinchant, and G. Jacquet, "Xrce's participation at medical image modality classification and ad-hoc retrieval tasks of imageclef 2011," in *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)*, September 2011.
18. J. Krapac, J. Verbeek, and F. Jurie, "Modeling spatial layout with fisher vectors for image categorization," in *International Conference on Computer Vision*, November 2011.
19. J. Philbin, O. Chum, and M. Isard, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, (Minneapolis, MN, USA), June 2007.
20. G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1 – 7, June 2007.
21. A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in *Conference on Knowledge Discovery and Data Mining (KDD)*, **5865**, pp. 58–65, AAAI Press, 1998.
22. D. W. Scott, "Averaged shifted histograms: Effective nonparametric density estimators in several dimensions," *Annals of Statistics* **13**(3), pp. 1024 – 1040, 1985.
23. A. Hinneburg and D. A. Keim, "A general approach to clustering in large databases with noise," *Knowledge and Information Systems* **5**(4), pp. 387–415, 2003.
24. A. Hinneburg and H.-H. Gabriel, "Denclue 2.0: Fast clustering based on kernel density estimation," *Advances in Intelligent Data Analysis VII* **4723/2007**, pp. 70–80, 2007.
25. W. Gan and D. Li, "Optimal choice of parameters for a density-based clustering algorithm," *Lecture Notes in Computer Science, Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing* **Volume 2639/2003**, p. 577, 2003.
26. R. Thomas, *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, Doctoral dissertation, University of California, 2000.