Semi–Supervised Learning for Image Modality Classification

Alba G. Seco de Herrera[1], Dimitrios Markonis[1], Ranveer Joyseeree[1,2], Roger Schaer[1], Antonio Foncubierta-Rodríguez[2], and Henning Müller[1]

[1] University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland,
alba.garcia@hevs.ch
http://medgift.hevs.ch/
[2] Swiss Federal Institute of Technology, Zurich, Switzerland

Abstract. Searching for medical image content is a regular task for many physicians, especially in radiology. Retrieval of medical images from the scientific literature can benefit from automatic modality classification to focus the search and filter out non–relevant items. Training datasets are often unevenly distributed regarding the classes resulting sometimes in a less than optimal classification performance. This article proposes a semi–supervised learning approach applied using a k–Nearest Neighbour (k–NN) classifier to exploit unlabelled data and to expand the training set. The algorithmic implementation is described and the method is evaluated on the ImageCLEFmed modality classification benchmark. Results show that this approach achieves an improved performance over supervised k–NN and Random Forest classifiers. Moreover, medical case–based retrieval benefits from the modality filter.

Keywords: semi–supervised learning, medical image classification, crowdsourcing, case–based retrieval

1 Introduction

A large amount of medical visual data is produced in hospitals daily. New imaging techniques continue to emerge, leading potentially to a further increase in data production. Many images are also made available continuously via publications in the scientific literature and are thus publicly available and not only in institutional archives. This adds up to an overwhelming amount of visual data for the physicians to analyse and take into consideration when taking clinical decisions.

Medical image analysis and retrieval have been active research fields over the past 15 years [29], trying to provide the tools to physicians to facilitate the access to and analysis of the medical visual data. Applications have been developed for Computer–Aided Diagnosis (CAD) [6, 12] and for general retrieval [29]. Recent medical information retrieval systems have followed a more user–oriented design [20], taking into consideration information seeking requirements of radiologists [26], which is a group of physicians with tasks focused strongly around images.

For the easier management of and access to the image datasets, searching and filtering by a specific image modality is often desired by radiologists [26]. Moreover, search systems that use content–based image retrieval can benefit both in speed and precision by reducing the search space with respect to the query image modalities [16]. Automatic modality classification is thus an important part of the performance and usability of modern medical retrieval systems, particularly when working with data from the literature, where a large number of image types are being used, not all of them clinically relevant. The ImageCLEFmed [21] evaluation campaign proposes a modality classification task to promote the research on this field.

In this paper, a method that uses semi–supervised learning (also referred to as training set expansion in this paper) to improve the classification accuracy based on the image modalities is proposed. Semi–supervised learning [7] uses a small number of labelled instances and a large amount of unlabelled data for training the classifier. It is used in cases where labelled data are rare and some classes are under–represented in the training set. This scenario is often met in medical image analysis, where accurate manual labelling of big datasets is difficult and expensive to obtain. Methods of semi–supervised learning have been applied to handwritten text recognition [4] and biological networks [34]. Related to this work, Csurka et al. [11] apply semi–supervised classification to medical image classification in order to expand the training set. The confidence scores for the unlabelled data are given by Support Vector Machine (SVM) classifiers using multi–modal (visual and textual) information and the expansion of the training set by visual retrieval is explored. In addition, our method uses multi–modal retrieval based on k–Nearest Neighbour (k–NN) and to expand the training set.

As a result of this semi–automatic learning a larger but noisy training set is obtained. This work proposes an iterative procedure to manually correct the expanded set by crowdsourcing. Crowdsourcing allows to divide the problem into microtasks that can be solved in a short amount of time by users familiar with medical images [18]. Crowdsourcing has recently emerged as a tool in bioinformatics because it can improve the quality, cost and speed of manual processing large amount of data [30]. This methodology allows dividing a data processing problem into simple manual micro–tasks [18]. Crowdsourcing has among other tasks been used for collecting and analysing health and medical research data or to create pre–clinical medical study material [2]. In particular several crowdsourcing challenges for image annotation have been proposed, such as generating models of proteins for successful molecular replacement and subsequent structure determination [22], classification of retinal fundus photography [27] or evaluating medical pictograms [35]. Furthermore, Vajda et al. [33] present a semi–automatic labelling strategy of images found in the literature by selecting for manual labelled only the images on a representative cluster.

This paper discusses the details of the implementation and evaluates it using the ImageCLEFmed[3] 2013 dataset [14]. Results are compared over various

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

training sets, outperforming k–NN and Random Forest classifiers that use the original training set and unimodal semi–supervised classifiers.

The remainder of the paper is organized as follows. Section 2 describes in detail the implemented approach as well as the description of the crowdsourcing performance. Experimental results are reported and discussed in Section 3. Section 4 concludes the paper.

2 Methods

This section describes the visual features used for representing the images, the dataset used for the evaluation and the proposed semi–supervised learning. Finally the image modality classification approach is integrated on a case–based retrieval system.

2.1 Dataset

The ImageCLEFmed 2013 [14] modality classification dataset was used in this study. The dataset contains 5483 labelled images, which are split into a training set of 2582 images and a validation set of 2901 images. It is a subset of the medical case–based retrieval task dataset that contains 300,000 unlabelled images of 75,000 articles for retrieval evaluation. Both datasets are considered realistic datasets, containing images from the open access biomedical literature.

ImageCLEFmed presents 31 classes of image types in a hierarchy. The classes are images types occurring in the biomedical literature including medical modalities such as magnetic resonance imaging, angiography, but also general graphs or multipane images. The distribution of the labelled data among the classes is extremely uneven, making the dataset very suitable for semi–supervised learning (see Table 1).

The classification approach is implemented to filter images for the medical case–based retrieval task. The medical case–based retrieval task proposes 35 case description, with patient demographics, limited symptoms and test results including imaging studies (but not the final diagnosis). The goal of this task is to retrieve cases including images that are relevant for differential diagnosis of the given case, so have the same or a very closely related diagnosis.

2.2 Multi–modal Features

The proposed method uses multi–modal information for the representation of the images. The text representation of the images uses a simple vector space model with stopword removal, word stemming, tokenization and a tf/idf weighting, using the Lucene[4] search engine based on the captions of the images (in the modality classification task) or the full text of the articles (in the case–based retrieval task). For the visual content of the images, a set of low–level visual

[4] `http://lucene.apache.org/`

Table 1. Distribution in classes of the images among the varoius training sets from the ImageCLEFmed 2013 classification task.

	Modality	**RO**	**RE**
Compound or multipane images		1,105	8,913
Diagnostic images	**Radiology**		
	Ultrasound	60	379
	Magnetic Resonance	97	496
	Computerized Tomography	113	593
	X-Ray, 2D Radiography	70	440
	Angiography	54	233
	PET	16	45
	Combined modalities in one image	22	72
	Visible light photography		
	Dermatology, skin	79	493
	Endoscopy	64	268
	Other organs	70	488
	Printed signals, waves		
	Electroencephalography	21	66
	Electrocardiography	29	91
	Electromyography	18	79
	Microscopy		
	Light microscopy	91	710
	Electron microscopy	51	218
	Transmission microscopy	46	244
	Fluorescence microscopy	33	233
	3D reconstructions	46	237
Generic biomedical illustrations			
Tables and forms		65	522
Program listing		28	167
Statistical figures, graphs, charts		102	915
Screenshots		91	587
Flowcharts		94	487
System overviews		89	654
Gene sequence		68	501
Chromatography, Gel		55	428
Chemical structure		62	285
Mathematics, formulas		20	164
Non–clinical photos		96	585
Hand–drawn sketches		46	320

features was selected from the feature bank of the Parallel Distributed Image Search Engine (ParaDISE) [31] developed in the medGIFT[5] group. A set of appropriate features was selected for each of the tasks based on the training data. The features were selected with exhaustive search on a combination of features.

The same features are used as in [16] to measure only the effect of semi–supervised learning on the classifiers used: color and edge directivity descriptor (CEDD) [8]; bag of visual words using scale–invariant feature transform (BoVW–SIFT) [25]; fuzzy color and texture histogram (FCTH) [9]; bag of colors (BoC) [15]; and fuzzy color histogram (FCH) [19].

For the case–based retrieval, the following features were chosen: BoVW–SIFT with a spatial pyramid matching [24]; BoC with $n \times n$ spatial grid (Grid BoC); CEDD; and Tamura texture [32].

2.3 Training Set Expansion

The training set is augmented because some classes contain only few annotated examples. only few annotated examples were available. The ImageCLEFmed 2013 modality classification training set is denoted as the set of labelled examples $x_1, \ldots, x_l \in X$ with $l = 2900$. Respectively, the corresponding labels are $y_1, \ldots, y_l \in Y$. The set of unlabelled examples $x_{l+1} \ldots x_{l+u} \in X$ refers to the ImageCLEFmed 2013 image retrieval dataset with $u + l = 300,000$. With X we denote the set that contains all the labelled and unlabelled examples. The proposed method labels $l \times k_r$ unlabelled examples, where k_r is a constant ($k_r = 10$ in this paper). Then, the expanded training set XX is used to train the classifier, where YY is the set of corresponding labels. The labelling of the unlabelled examples is described in the Algorithm 1.

In practice, because of the removal of double instances, the size of the expanded training set is slightly smaller than $l + l \times k_r$. Histogram intersection is used as similarity measure for the image retrieval on the set of visual features.

2.4 Crowdsourcing

This work uses the Crowdflower[6] platform to manually correct the automatic training set expansion described in Section 2.3. The internal crowdsourcing interface allows to carry out tasks by a known set of experts to guarantee the precision of the results. In this work, eight experts in the medical imaging domain participated in the crowdsourcing job. More details on the described crowdsourcing task can be found in [13].

The correction task was divided into several steps that were executed in an iterative way:

[5] `http://medgift.hevs.ch/`
[6] `http://www.crowdflower.com/`

Data: $X; Y$
Result: $XX \subseteq X; YY$
$XX = \{x_i, i \in [1 \dots l]\}$ `/* initialize to the original training set */`
for $i = 1 \dots l$ **do**
  query x_i against X
  retrieve the top k_r results r
  for $j = 1 \dots k_r$ **do**
    `/* do not re--include original examples */`
    if $r_j \neq x_m \forall m \in [1 \dots l]$ **then**
      $y_j = y_i$ `/* assign label to result */`
      $XX = XX \cup r_j$ `/* expand training set */`
    end
  end
end
remove examples with multiple labels

Algorithm 1: Training expansion algorithm.

Verification The crowdsourcing verification task was set up to verify the automatically assigned labels. Since about 40% of the figures in the biomedical open access literature [10] are compound or multipane figures (with several subfigures in a single image) an additional option was added to facilitate the following steps. Apart from the options 'Yes, perfect classification' and 'No, wrong classification' it was possible to choose 'No, compound image' for incorrectly classified compound figures and it was possible to mark 'Not sure'.

Relabelling Images that were incorrectly classified automatically or tagged as 'not sure' were manually relabelled in a second crowdsourcing iteration. The images were relabelled into the 31 classes of the hierarchy presented in Section 2.1. As these are the images difficult to classify in the first iteration, each of the images was classified by two participants. In case of disagreement between the answers a third expert labelled the image and the majority decision was taken.

Algorithm 2 describes the iterative crowdsourcing process.

In this algorithm, m is the size of the expanded training set.

2.5 Classification

Two methods were employed for the classification of the biomedical images based on the modalities. The use of two classifiers allows a further study of the effects of the proposed semi–supervised learning method. The first method follows previous work [16] applying a k–NN classifier using weighted voting.

This was followed using Random Forests [3] over the visual features. The randomForest package in the R programming language[7]was used. All default settings were kept, meaning that there are 500 trees in the forest. Bootstrap

[7] `http://cran.r-project.org/doc/manuals/r-release/R-lang.html`

```
Data: XX, YY
Result: YY
XX = {x_i, i ∈ [1 ... l ... m]}
for i = l + 1 ... m do                                  /* manual verification */

    verify label y_i of image x_i;
    if y_i ≠ 'Yes, perfect classification' then
        if y_i is compound image then
            y_i ='COMP';
            else                                        /* manual relabelling */

             | relabelled x_i
            end
        end
    end
    if i = (m − l)/2 & i ≤ l then                       /* automatic reclassification */

     | reclassify x_i, i ∈ [i ... l ... m]} using updated subset
    end
end
```

Algorithm 2: The iterative crowdsourcing algorithm.

aggregation (bagging) was employed whereby training is carried out with only a subset of the total number of variables per feature. The number of chosen variables was set to the floored value of the square root of the total number of variables per feature. Sampling of the training set was done with replacement and the sample size was the total number of training points.

The importance of each variable in a feature was calculated at training time using out–of–bag (oob) data. A variable was deemed to have high importance if it lead to a high decrease of accuracy in the trees formed. The values obtained were used to rank the variables in order of importance. This was carried out for all the features and a new feature was formed by combining the most important variables in the original features. The maximum number of variables used for the new feature was set to the maximum variable count for the set of features used, being 238. Limiting the number of variables used helps to minimize the effects of overfitting.

2.6 Medical Case–based Retrieval

This section details the retrieval tools that are used to create the multi–modal retrieval baseline. The approach is based on the approach presented for the ImageCLEFmed 2013 case–based tasks [16].

Previous studies have shown that it is successful for image retrieval and it is robust to many transformations [5, 1].

Optimal fusion techniques were also selected following previous work [17]. Fusion is performed in several cases in the retrieval pipelines: to handle multiple

query images in content–based image retrieval (combMAX) and to combine the various visual (Borda) and textual features (linear).

Modality Filter The effect of modality filtering on the retrieval quality was investigated. All the images of the ImageCLEFmed 2013 dataset and the query images of each topic were classified. A set of query modalities was produced for each topic extracted from the query images. For each query image a rank list is retrieved and filtered. Then, the fusion of the various ranked lists obtained from each of the query images are combined. Four approaches of modality filtering were tested:

- *Exact* – uses a single modality of the query image for filtering the list of this image.
- *Close* – uses a set of all modalities occurring in the query images of each topic to filter the list of each image query.
- *Prefix* – is similar to the first but the broadest category (diagnostic, general, compound) was used instead of the exact modality.
- *Diagnostic* – only diagnostic images from the database are retrieved. This approach does not depend on the query.

3 Experimental Results

This section presents the evaluation results for a subset of the experiments over the ImageCLEFmed 2013 database. First of all, results achieved for the ImageCLEFmed modality classification task are presented. Then, image classification is applied as a filter for ImageCLEFmed the case–based retrieval task and results of this are shown.

3.1 Classification

First, results using the k–NN classifier are investigated as well as the results of the crowdsourcing procedure.

In [11] an arbitrary k is used for the k–NN classifier, explaining that the choice of k may not be optimal. We compute the classification accuracy for a range of k to investigate the robustness of the method. Table 2 shows the results for the experiments using the k–NN classifier over the following training sets:

- **RO**: original training set.
- **RE**: automatically expanded training set.
- **REN**: automatically expanded training set without expanding the compound images.
- **REH**: automatically expanded training set with half of the expanded images manually relabelled.
- **RENH**: automatically expanded training set without expanding the compound images. Half of the expanded images are manually relabelled.

- **RET**: automatically expanded training set all of the expanded images manually relabelled.
- **RENT**: automatically expanded training set without expanding the compound images. All of the expanded images are manually relabelled.
- **REW**: automatically expanded training set without the images labelled as 'compound' or 'correct'. All of the expanded images are manually relabelled.

The experiments carried out over each of the training set are discussed in detail below.

Table 2. Accuracy (%) obtained applying the k–NN classifier with various $k's$ and the Random Forest (RF) classifier for the ImageCLEFmed 2013 modality classification using various training sets. Average (Avg) and standard deviation (SD) over the $k's$ are also shown.

k	RO	RE	REN	REH	RENH	RET	RENT	REW
2	65.05	62.19	43.67	65.88	53.51	68.19	70.67	66.37
3	68.19	67.11	49.13	69.82	60.44	**69.16**	73.5	69.86
4	68.62	68.38	50.14	69.78	61.76	68.66	73.27	70.13
5	68.70	69.55	51.41	70.40	64.20	67.34	73.30	70.44
6	68.81	69.70	52.54	70.4	65.44	67.53	**73.89**	70.98
7	68.85	70.63	53.20	70.36	67.07	67.45	73.42	70.71
8	68.50	71.10	54.20	70.86	67.96	66.83	72.68	**71.83**
9	68.26	71.68	54.39	70.55	68.19	66.80	73.27	71.25
10	**68.93**	71.52	55.60	70.75	69.35	66.52	73.23	71.33
11	68.62	71.76	55.52	70.63	70.44	66.49	72.84	71.41
12	68.00	72.26	56.49	**71.02**	70.71	66.83	72.80	71.68
13	67.76	71.87	56.88	70.40	70.90	65.83	72.80	71.33
14	67.49	72.49	56.92	70.59	71.76	65.40	72.57	71.17
15	67.34	72.26	57.07	70.40	72.57	65.13	72.10	71.02
16	66.99	**72.61**	57.50	70.28	72.34	64.94	72.14	70.36
17	66.99	72.26	**57.96**	70.44	72.65	64.20	71.83	70.59
18	66.68	72.41	58.04	70.28	72.65	64.20	71.91	70.55
19	66.37	72.38	57.57	70.48	72.84	64.24	71.37	70.24
20	66.56	71.83	57.81	70.21	73.03	64.01	71.52	70.52
21	66.21	72.49	57.61	69.90	**73.34**	63.62	71.41	70.52
22	66.18	**72.61**	57.88	69.82	73.30	63.27	71.10	70.24
23	65.79	72.57	57.69	69.74	73.19	62.92	71.02	70.24
24	65.71	72.14	57.77	65.25	73.23	62.81	71.29	69.97
25	65.71	72.22	57.77	64.97	72.76	62.46	71.33	69.55
Avg	67.35	71.08	55.20	69.72	69.32	65.62	**72.30**	70.51
SD	1.21	2.37	3.61	1.72	5.06	1.96	**0.94**	1.06
RF	68.13	69.13	60.88	–	–	67.62	65.03	**69.25**

Using the original training set (RO) results achieved 68.93% accuracy when $k = 10$. The accuracy is increased to 72.61% when using $k_r = 10$ for the ex-

pansion of the training set (RE) and $k = 16$. Because a majority of the figures in the dataset are compound figures it was proposed to not include the compound figure class to the training set expansion [16]. An additional experiment which does not expand this class was run (REN) but with worse results than the previous experiments.

Half of the expanded training data were then manually corrected using crowdsourcing as explained in Section 2.4. New labels of this first iteration of the correction procedure (see Algorithm 2) were then updated in the expanded data. Therefore, two more experiments were run using the new labels (REH and RENH) increasing the accuracy to 73.34% with $k = 21$. This result was obtained without including the images tagged as 'No, compound images' during the crowdsourcing iteration (RENH) in the training data.

To crowdsource the label correction of the images of the second half of the expanded training set, images were previously reclassified using the RENH training set and a k–NN classifier. $k = 21$ was chosen because it generates the highest accuracy in the experiments. Hence, the number of correct images classified increased and fewer images were relabelled in the crowdsourcing. Figure 1 shows the distribution of each of the answers in the two verification tasks carried out during the iterative process described in Section 2.4.

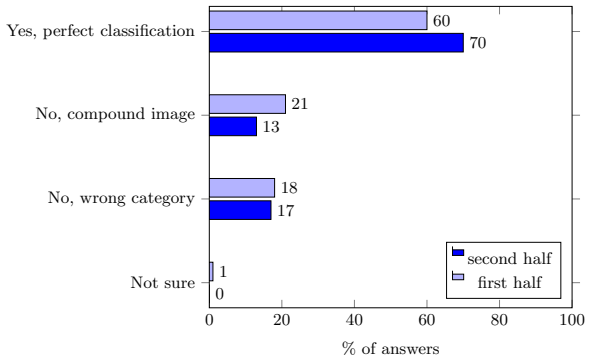

Fig. 1. Each bar represents the distribution of each of the answers in the verification crowdsourcing task. The second half of the set was reclassified using the updated training set after the first half was relabelled.

Finally, a training set with 19,905 images containing correct labels is obtained. Using this training set (RET) and $k = 3$ the accuracy obtained is 69.16%. Two more experiments were then carried out. One excluded the images tagged as 'No, compound images' during the crowdsourcing iteration (RENT) into the training data. Considering the hypothesis that the correctly classified and the compound images of the expanded dataset do not add information because the algorithm was already able to classified them well, another experiment was done adding only the images tagged as 'No, wrong category' and 'Not sure' during

the crowdsourcing iteration (REW) to the original training set. Best results were obtained using the RENT training set and $k = 6$ achieving 73.89% of accuracy. The used of training set also achieved best accuracy on the average over the k values. Indeed the standard deviation is also the lowest (0.94) showing that the results obtained with most of the k values are very close to the average.

Classification was also carried out using Random Forests over only the visual descriptors (see Section 2.5). Experimental results are shown in Table 2. The REH and RENH training sets were not used because they were an intermediate step for crowdsourcing the label correction.

Results also show an improvement when using the semi–supervised learning variant REW (69.25%), which validates the hypothesis that initially wrongly labelled images add information to the classifier.

The proposed semi–supervised approach achieves its best performance using k–NN algorithm when $k = 6$, while for the baseline approach, this happens when $k = 10$. Relative lower standard deviations of the algorithm suggest that the k–NN algorithm is stable across k choices (see Table 2). Using k–NN , the average accuracy shows that the variant without the added compound images, so the majority class, (RENT) of the proposed multi–modal approach achieves the best results. The results obtained by the full expanded dataset (RET) were poorer than the ones obtained without the extra compound images (RENT). This can be explained by the fact that the compound figures belong to the most frequent class in the test set and do not require expansion. It also demonstrates that the application of semi–supervised learning is not trivial and the algorithms need to take into account the data distribution across classes.

Using Random Forest classifiers best results were achieved by the training set with the images labelled as 'wrong' or 'not sure' (RETW). This shows that these images provide more relevant information than the expanded compound or the correctly classified by the k–NN algorithm. Again not all of the expanded training sets improve the result.

Both classifiers perform better using the expanded and corrected training sets, demonstrating the effectiveness of the presented method. Better accuracy scores than the ones reported in this paper were obtained by two participants in ImageCLEFmed 2013 [14](81.68% and 78.04%). However, these multi–modal approaches used much more sophisticated classifiers such as support vector machines or multiple classifiers [23, 28] compared to k–NNs on more complex visual features. They also used additional external training data, making the results difficult to compare. The proposed approach using a more sophisticated classifier (the Random Forest classifier) using unimodal information (only visual features) achieves an accuracy of 69.04%. In addition, default settings were used and no optimisation was attempted for this method, meaning that there is an improvement potential. Thus, we are confident that using the technique in a multi–modal classification using Random Forests can achieve even better performance. The purpose of this study, however, was to demonstrate the effect of semi–supervised learning on the classification performance. More importantly, other supervised

learning techniques can benefit from the proposed method, as demonstrated by the Random Forest tests.

3.2 Modality Filter for Medical Case–Based Retrieval

Table 3 shows the performance of the four types of modality filters applied on the medical case–based retrieval task. To carry out these experiments all the images were classified using the best classification approach presented above. Results are also compared with the baseline presented in Section 2.6 and with the best result achieved in ImageCLEFmed 2013. Best results were achieved when using the "prefix" approach, outperforming the best ImageCLEFmed results as well as the baseline presented in this work. More important is the reduction of the image search space obtained in the content–based image retrieval step.

Table 3. Results of the approaches when using various modality filter strategies on the ImageCLEFmed 2013 case–based retrieval task. Results are compared with the baseline and the best mix results submitted to ImageCLEFmed 2013.

Run ID	MAP	GMAP	Bpref	P10	P30
Best mix ImageCLEFmed	0.1608	0.0779	0.1426	0.1800	0.1257
baseline	0.1889	0.1190	0.1720	0.2257	0.1629
exact	0.1887	0.1193	0.1728	**0.2286**	0.161
close	0.1892	0.1191	0.1720	**0.2286**	0.1629
prefix	**0.1904**	**0.1208**	0.1732	0.2257	**0.1638**
diagnostic	0.1874	0.1177	**0.1735**	0.2257	0.1581

4 Conclusions

Modality classification is important in medical image retrieval systems both for overall retrieval quality and usability. This paper describes a method for improving medical image classification and in a subsequent step medical case–based retrieval. The method uses multi–modal retrieval to exploit unlabelled data for semi–supervised learning of a k–NN classifier. A crowdsourcing platform was used to manually correct the assigned labels.

This method is applied and evaluated on k–NN and Random Forest classifiers. The results show that this method provides higher accuracy than a similar k–NN classifier trained in a supervised way using a smaller k. Removing the additional compound figures in this pipeline (as overrepresented) improves performance when using the k–NN classifier. Removing as well the automatically corrected images achieves even better results when using the Random Forest classifier. Applying both classifiers demonstrates that semi–supervised learning improves classification accuracy.

Three approaches are investigated for using an image modality filter in the case–based retrieval step. Results show that filtering by a broad modality in the hierarchy improves the retrieval performance and reduces the search space.

In the future, we plan to evaluate the training set expansion using other classifiers, such as support vector machines and also using more complex visual features. Evaluation with multi–kernel classifiers could also better demonstrate the added value of multi–modal semi–supervised learning. In addition, Random Forests will be optimized on the training sets to improve accuracy.

Acknowledgements. This work was partly supported by the EU 7^{th} Framework Program in the context of the Khresmoi project (FP7–257528).

References

1. Sabri Boughorbel, Jean-Philippe Tarel, and Nozha Boujemaa. Generalized histogram intersection kernel for image recognition. In *IEEE International Conference on Image Processing*, volume 3, pages III–161. IEEE, 2005.
2. Hansen C. Bow, Jonathan R. Dattilo, Andrea M. Jonas, and Christoph U. Lehmann. A crowdsourcing model for creating preclinical medical education study tools. *Academic Medicine*, 88(6):766–770, 2013.
3. Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
4. Javier Cano, Juan-Carlos Pérez-Cortes, Joaquim Arlandis, and Rafael Llobet. Training set expansion in handwritten character recognition. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 548–556. Springer, 2002.
5. Rishav Chakravarti and Xiannong Meng. A study of color histogram based image retrieval. In *Sixth International Conference on Information Technology:New Generations ITNG*, pages 1323–1328, 2009.
6. Heang-Ping Chan, Jun Wei, Yiheng Zhang, Mark A. Helvie, Richard H. Moore, Berkman Sahiner, Lubomir Hadjiiski, and Daniel B. Kopans. Computer–aided detection of masses in digital tomosynthesis mammography: Comparison of three approaches. *Medical Physics*, 35(9):4087–4095, 2008.
7. Olivier Chapelle, Bernhard Schölkopf, Alexander Zien, et al. *Semi–Supervised Learning*, volume 2. MIT press Cambridge, 2006.
8. Savvas A. Chatzichristofis and Yiannis S. Boutalis. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Lecture notes in Computer Sciences*, volume 5008, pages 312–322, 2008.
9. Savvas A. Chatzichristofis and Yiannis S. Boutalis. FCTH: Fuzzy color and texture histogram: A low level feature for accurate image retrieval. In *Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Service*, pages 191–196, 2008.
10. Ajad Chhatkuli, Dimitrios Markonis, Antonio Foncubierta-Rodríguez, Fabrice Meriaudeau, and Henning Müller. Separating compound figures in journal articles to allow for subfigure classification. In *SPIE Medical Imaging*, 2013.
11. Gabriela Csurka, Stephane Clinchant, and Guillaume Jacquet. XRCE's participation at medical image modality classification and ad–hoc retrieval task of ImageCLEFmed 2011. In *Working Notes of CLEF 2011*, 2011.

12. Adrien Depeursinge, Jimison Iavindrasana, Gilles Cohen, Alexandra Platon, Pierre-Alexandre Poletti, and Henning Müller. Computer–aided diagnostic for interstitial lung diseases in HRCT: the talisman project. In *Swiss Conference on Medical Informatics*, Sierre, Switzerland, June 2008.
13. Alba García Seco de Herrera, Antonio Foncubierta-Rodríguez, Dimitrios Markonis, Roger Schaer, and Henning Müller. Crowdsourcing for medical image classification. In *Annual Congress SGMI 2014*, 2014.
14. Alba García Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner Fushman, Sameer Antani, and Henning Müller. Overview of the ImageCLEF 2013 medical tasks. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.
15. Alba García Seco de Herrera, Dimitrios Markonis, and Henning Müller. Bag of colors for biomedical document image classification. In Hayit Greenspan and Henning Müller, editors, *Medical Content–based Retrieval for Clinical Decision Support*, MCBR–CDS 2012, pages 110–121. Lecture Notes in Computer Sciences (LNCS), October 2013.
16. Alba García Seco de Herrera, Dimitrios Markonis, Roger Schaer, Ivan Eggel, and Henning Müller. The medGIFT group in ImageCLEFmed 2013. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.
17. Alba García Seco de Herrera, Roger Schaer, Dimitrios Markonis, and Henning Müller. Comparing fusion techniques for the ImageCLEF 2013 medical case retrieval task. *Computerized Medical Imaging and Graphics*, 39:46–54, 2015.
18. Benjamin M. Good and Andrew I. Su. Crowdsourcing for bioinformatics. *Bioinformatics*, 16(29):1925–1933, 2013.
19. Ju Han and Kai-Kuang Ma. Fuzzy color histogram and its use in color image retrieval. *IEEE Transactions on Image Processing*, 11(8):944–952, 2002.
20. Allan Hanbury, Célia Boyer, Manfred Gschwandtner, and Henning Müller. KHRESMOI: Towards a multi–lingual search and access system for biomedical information. In *Med–e–Tel, Luxembourg, 2011*, pages 412–416, 2011.
21. Jayashree Kalpathy-Cramer, Alba García Seco de Herrera, Dina Demner-Fushman, Sameer Antani, Steven Bedrick, and Henning Müller. Evaluating performance of biomedical image retrieval systems an overview of the medical image retrieval task at ImageCLEF 2004–2014. *Computerized Medical Imaging and Graphics*, 39(0):55 – 61, 2015.
22. Firas Khatib, Frank DiMaio, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177, 2011.
23. Ivan Kitanovski, Ivica Dimitrovski, and Suzana Loskovska. FCSE at medical tasks of ImageCLEF 2013. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.
24. Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.
25. David G. Lowe. Distinctive image features from scale–invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
26. Dimitrios Markonis, Markus Holzer, Sebastian Dungs, 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.
27. Danny Mitry, Tunde Peto, Shabina Hayat, James E Morgan, Kay-Tee Khaw, and Paul J. Foster. Crowdsourcing as a novel technique for retinal fundus photography classification: Analysis of images in the epic norfolk cohort on behalf of the UK biobank eye and vision consortium. *PLOS ONE*, 8(8), 2013.
28. André Mourão, Flávio Martins, and João Magalhães. NovaSearch on medical ImageCLEF 2013. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.
29. 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.
30. Benjamin L. Ranard, Yoonhee P. Ha, Zachary F. Meisel, David A. Asch, Shawndra S. Hill, Lance B. Becker, Anne K. Seymour, and Raina M. Merchant. Crowdsourcing–harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine*, 29(1):187–203, 2014.
31. Roger Schaer, Dimitrios Markonis, and Henning Müller. Architecture and applications of the Parallel Distributed Image Search Engine (ParaDISE). In *FoRESEE 2014, 1st International Workshop on Future Search Engines at INFORMATIK 2014*, 2014.
32. Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):460–473, June 1978.
33. Szilárd Vajda, Daekeun You, Sameer K Antani, and George R. Thoma. Label the many with a few: Semi–automatic medical image modality discovery in a large image collection. In *Computational Intelligence in Healthcare and e-health (CICARE), 2014 IEEE Symposium on*, pages 167–173. IEEE, 2014.
34. Kevin Y. Yip and Mark Gerstein. Training set expansion: An approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. *Bioinformatics*, 25(2):243–250, 2009.
35. Bei Yu, Matt Willis, Peiyuan Sun, and Jun Wang. Crowdsourcing participatory evaluation of medical pictograms using Amazon mechanical turk. *Journal of Medical Internet Research*, 15(6), 2013.