Label the many with a few: Semi-automatic medical image modality discovery in a large image collection

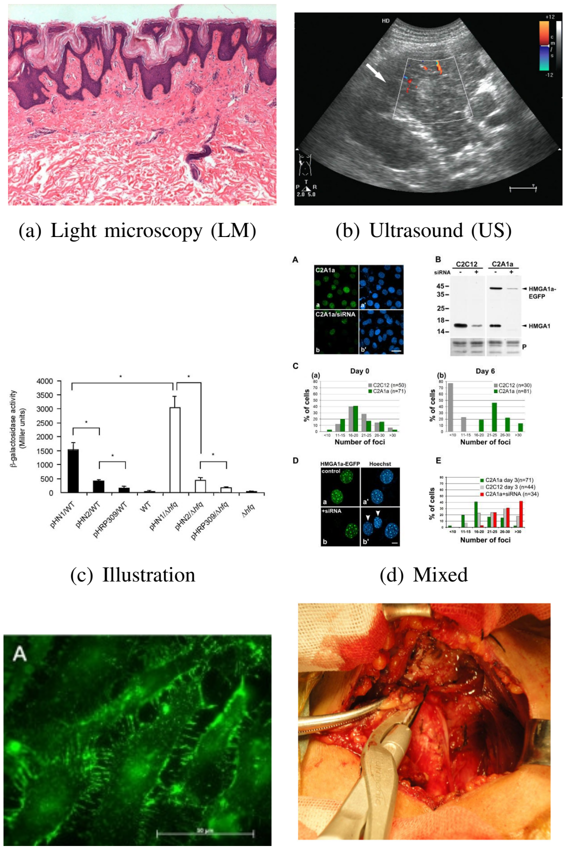
Szilárd Vajda, Daekeun You, Sameer K. Antani, George R. Thoma
Lister Hill National Center for Biomedical Communications
National Library of Medicine, National Institutes of Health
8600 Rockville Pike, Bethesda, MD, USA
{szilard.vajda,daekeun.you,sameer.antani,george.thoma}@nih.gov

***Abstract*—In this paper we present a fast and effective method for labeling images in a large image collection. Image modality detection has been of research interest for querying multimodal medical documents. To accurately predict the different image modalities using complex visual and textual features, we need advanced classification schemes with supervised learning mechanisms and accurate training labels. Our proposed method, on the other hand, uses a multiview-approach with minimal expert knowledge to semi-automatically label the images. All the images are projected in different feature spaces, which are then clustered in an unsupervised manner. Each cluster representative is mapped back to the image space, and labeled by an expert. The other images from the clusters "inherit" the labels from these cluster representatives. The final label is assigned to each image based on a voting mechanism, each vote providing an different opinion about the same image. The experimental setup showed that using only 0.3% of the labels was sufficient to annotate 300,000 medical images with 49.95% accuracy. Although, automatic labeling is not as precise as manual, it saves approximately 700 hours of manual expert labeling. We find that for this collection accuracy improvements are feasible with better disparate feature selection or different filtering mechanisms.**

I. Introduction

Medical image retrieval in the context of large collections is a challenging and demanding task, thus more attention has been recently focused on this type of systematic effort[1]. Knowing the modality of an image [1], i.e., either it is an X-ray, CT, MRI or a photograph, improves the performance of image or article retrieval because the search space may be greatly reduced. For more details on the modalities, please refer to [1], [2].

Besides the textual description of the images available in medical documents, visual categorization is also gaining significant interest. Textual search combined with visual -image based- search outperforms the different methods [2] due to the descriptive power of the visual appearance of the different modalities (see Fig. 1). While different image modalities can be characterized using textual representations (e.g. captions) and specialized vocabularies (e.g. UMLS®), visual classification relies on a large number of training samples.



(a) Light microscopy (LM)

(b) Ultrasound (US)

(c) Illustration

(d) Mixed

(e) Fluorescence microscopy (FM)

(f) Photo

Fig. 1. Different image modalities appearing in medical literature (images from [3].)

In order to outperform the limited capabilities of existing unsupervised classification methods like Lloyd clustering (variant of kmeans) [4], single linkage clustering [5], Self Organizing Map (SOM) [6], Growing Neural Gas (GNG) [7], more sophisticated supervised classification strategies such as Support Vector Machine (SVM) [8] or neural networks are needed to define complex and high dimensional decision

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

surfaces for the underlying vector representations.

To perform such supervised training, a large amount of labeled data is necessary. Labeling medical images (i.e. CT, X-ray, Ultrasound, Microscopy, etc.) involves trained human annotator(s), and a significant amount of time and effort to review and label each image. Our goal is to propose a method which can overcome these drawbacks - requiring expert knowledge and the invested time and effort. We seek to robustly infer labels for large medical image collections using only a minimal amount of expert involvement.

The proposed semi-supervised labeling strategy exploits several visual descriptors such as CEDD (Color and Edge Directivity Descriptor), CLD (Color Layout Descriptor)) extracted from the raw images, and a textual descriptor computed from the image captions. Each feature representation is clustered in an unsupervised manner (i.e. using kmeans, SOM or GNG). The only constraint is the control over the number of clusters to minimize of the human annotator's involvement. Each cluster center is annotated and the labels are propagated through all the samples belonging to the same cluster. The different image representations are likely to cluster the different items in different arrangements, thus a majority vote will provide the necessary robustness to infer a reliable label for each item. However, the ultimate goal is not to label the images, but to extend the ground truth data with the newly discovered, admittedly noisy data, and retrain the image modality classifier [9].

The remainder of the paper is organized as follows: Section II gives a brief overview about the different modality recognition attempts along with the semi-automatic labeling strategies found in the literature. Section III presents the proposed labeling strategy, while Section IV is dedicated to the experiments. Finally, Section V provides the conclusions.

II. Related Work

The following section gives a brief overview of modality detection in general, describing the most prominent visual and textual features, and at the same time presents some semi-automatic labeling systems based on active learning.

A. Modality detection

Image modality classification has been one of the main tasks in the medical image classification and retrieval track of ImageCLEF. ImageCLEF started in 2003 as part of the Cross Language Evaluation Forum (CLEF). ImageCLEFmed medical information retrieval track was added to the evaluation in 2004. Modality classification was introduced in the ImageCLEFmed track in 2010. The goal of the task is to detect the modality of the images in the collection using visual, textual, or mixed methods.

In our ImageCLEFmed2012 participation [2], we extracted 15 different visual features from the images and textual features from the article citations, figure captions, mentions, and MeSH® terms. Also, class-specific contents such as text strings and polygons (e.g., rectangles in flowcharts, hexagons in chemical diagrams, etc.) are extracted from illustration figures to assist our SVM-based main modality classifiers [2]. Flat (a single multiclass classifier) and hierarchical classification approaches were developed using the features separately or in combination. Our best classification result ranked within the top three groups in the competition. Details of modality classification techniques and results from various participating groups are presented in the proceedings [3]. To our knowledge, besides the experiments conducted in ImageCLEF, no other work can be found in the literature regarding medical image modality detection.

B. Active learning

Active learning systems [10] try to reduce the manual labeling work by asking an "oracle" (e.g., a human annotator) to label some unknown (unlabeled) data instances, and based on this knowledge learn to classify the rest of the samples. Usually in these setups there is a huge amount of unlabeled data, and only a limited amount of data available with correct labels or no labeled data at all. The goal is to robustly infer labels for the unknown samples exploiting the limited information available. The labeling process is performed in such a way to minimize the cost of labeling (involvement of a human expert). The known labels must be reliable. To get reliable labels the involvement of the human expert is mandatory; moreover, the annotated samples should be representative for the unlabeled set as typically the new labels are inferred through similarity measures.

To robustly propagate a concept in multiview-learning, a strategy of using an ensemble of learners is proposed [11], [12]. Each of these learners has a different opinion (label) on the data, e.g., by using different feature representation. Decisions are made by combining the outputs of different learners. A well-known strategy is using a majority vote [12]. The advantages of incorporating ensembles in semi-supervised learning approaches for robust propagation are, for example, discussed in [13].

For handwritten graphical multi-stroke symbols an annotation assistance is proposed by Li et al. [14], where the annotation of the symbols is reduced to finding sub-graphs in a relation graph built from different segments. In the graph the nodes are the segments, and the arcs represent the spatial relationships between them. The authors show that only 58.2% of the strokes need to be labeled. With respect to the goal of reducing the manual effort in the transcription of historical documents, the work introduced by Toselli et al. in [15] has a similar goal as ours. However, the principle differs from our approach. Recently, the work proposed by Vajda et al. [16], [17] deal with the same problem in a character recognition scenario, where unlabeled character data collections were labeled using a multi-view system, and a majority voting decides about the labels of handwritten characters. In this study, authors proved that only a few hundred labeled characters are necessary to accurately label several thousands.

III. Method

This section describes the semi-supervised labeling method and the underlying feature representations used in our experiments, as well the choice of the selected clustering strategy.

A. Feature representations

You et al. [2] considered 15 different low-level visual features for modality detection including color features, edge features, texture features, and their different combinations [18]. After ranking the different features by their discriminative power, it was found that the CEDD and the CLD are the most prominent features, therefore, these were selected to provide visual descriptions for the images in our experiments.

In order to extract textual features, approximately 283 text terms from figure captions were identified based on their ability to provide the most relevant information about image modalities. The terms are selected from the captions of 1,000 biomedical images and include, for example, "computed tomography", "CT", "confocal microscopy", "T1-weighted", "flowchart", "photograph", etc. In the construction of a textual representation, the captions of the images were examined, and a 283-dimensional feature vector was built, indicating the frequency of each term in the caption. This kind of representation of the image is powerful, and in many cases outperforms the visual features [2].

In order to minimize the amount of work to be accomplished by the expert performing the labeling, we selected only these three features, although a larger variability in the feature spaces would likely have improved the final recognition (labeling) performance. The odd number of features selected is motivated by our preference for a simple majority in the voting scheme.

B. The choice of the clustering method

Clustering data, in general terms, refers to partitioning the input into meaningful groups based on some proximity measure such as Euclidean distance, Hamming distance, Mahalanobis distance, etc. A large number of methods have been proposed, but for any given set of data points, the different clustering strategies provide different outcomes. This is mainly due to the type of the data, their representation, and last but not least the distance metric applied in the methods.

To objectively evaluate different clustering methods, various measures were proposed based on quality and quantity [19], [20]. However, for our purpose, instead of using the Davis-Bouldin index [20], we used a measure based on the cluster distribution, namely the cluster compactness, which is computed using the generalized definition of variance of a vector:

$$v(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d^2(x_i, \bar{x})}. \tag{1}$$

Here $d(x_i, x_j)$ is a metric defined between the vectors x_i and x_j, respectively. N stands for the number of items in X (set of vectors), while $\bar{x} = \frac{1}{N}\sum_i x_i$ is the mean of X. Using equation 1, the cluster compactness for $C_1, C_2, \ldots, C_K$ (K stands for the number of clusters) is defined such as

$$Compactness = \frac{1}{K}\sum_{i}^{K} \frac{v(C_i)}{v(X)}. \tag{2}$$

In our experimental setup, we considered single linkage clustering, kmeans, SOM and GNG. These methods have the advantage of closely controlling the number of clusters represented in the dendogram, the number of clusters and the number of neurons, respectively. The well-known MeanShift algorithm [21] was also considered, but due to the inconvenience of selecting a proper bandwidth value (h) in the kernel density estimator function, - a parameter highly dependent on the data, this approach was abandoned.

For our data (visual and textual features), kmeans clustering was determined to be the most efficient in terms of compactness as depicted by equation 2. Hence, we retained only the kmeans method for further experiments. The kmeans has the advantage of easy implementation, and setting the number of clusters in advance. This property is advantageous as one can directly control the amount of work for the expert. The goal is to keep the burden for the annotator as low as possible.

However, the method allows the usage of different clustering methods, and different distance metrics for partitioning each feature space. Even the number of clusters can differ from one feature space to another. The only important aspect is to control the number of clusters, and thereby control the amount of work to be invested by the human expert.

C. Manual labeling and voting

Once the clusters are created in different feature spaces, the different cluster centers, so-called "centroids" represent the underlying clusters. However, these centroids do not represent real data points as they are created over the iterative processes during the different clustering strategies. Hence the closest real data items (using the same distance metric as for the clustering) are selected to represent the clusters.

The human annotator labels all these cluster representatives by mapping back the feature vectors to their original data source - the images. This mapping from the different feature spaces back to the image domain allows the annotator to retrieve the original image data, and therefore make it readable. Thus, each cluster representative gets assigned a label representing a certain modality such as Photo, Illustration, Microscopy, Ultrasound, X-ray, CT, etc. That same label is then propagated through all the data points (other images) residing in the same cluster. In that case each data sample gets as many labels as many feature representations are involved in the process. In our case each image has three labels according to the CEDD, CLD and modality term frequency, considered in our experimental setup (see for details Section III-A).

For each image sample unanimity (complete majority) or simple majority vote [12] is performed to determine the label. A formal description of the decision can be described as follows:

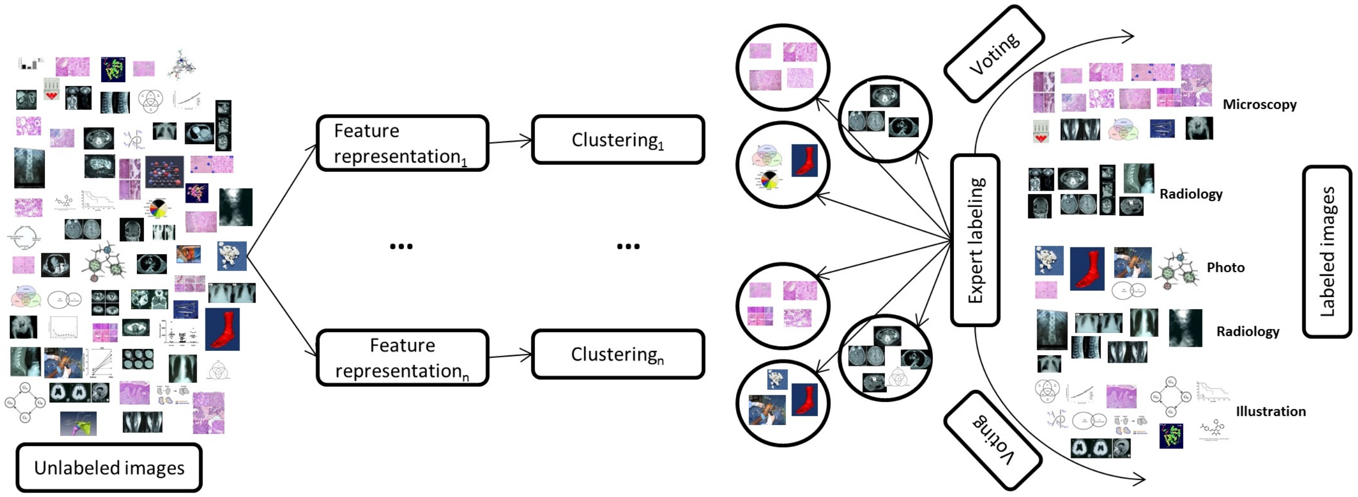


Fig. 2. The semi-automatic labeling process: a system overview.

Given a set of patterns $p_i \in P$ belonging to C different classes represented by the labels $[L_{i1}, \ldots, L_{iC}] \in \{0,1\}^C$, where $i = 1 \cdots N$, N is the number of samples and n the number of classifiers available, the pattern p_i is labeled as belonging to class $L_c \in \{L_1, \ldots, L_C\}$.

$$\sum_{\substack{k=1 \\ c \in \{1,\ldots,C\}}}^{n} L_{ic} = n. \tag{3}$$

$$\sum_{\substack{k=1 \\ c \in \{1,\ldots,C\}}}^{n} L_{ic} \geq \left[\frac{n}{2}\right] + 1. \tag{4}$$

If unanimity (see condition of equation 3) or the simple majority (see condition of equation 4) of the votes go for a particular label, the data point is labeled accordingly. Otherwise, the data is rejected as being uncertain. The choice of the vote can be used as a quality measure. If the vote is unanimous the label is more likely to be correct with greater certainty than in the case of just a simple majority vote. The method is similar to multi-view learning, where each feature space can be considered a separate view of the same object (data point). The more views agree upon a certain label, the more likely the assigned label is correct. However, unanimity does not guarantee that the label will be correct. In some cases all views (classifiers) can vote wrongly for the same class, thus implanting error in the labeling system. The complete labeling procedure is depicted in figure 2.

IV. Evaluation

To completely analyze the proposed method, different experiments were conducted. First the data will be presented, followed by the description of the experimental protocol, and the results achieved.

A. Data description

We use the ImageCLEF2012[2] dataset that is made available through our participation in the forum. The dataset consists of over 300,000 biomedical figures that originate from the open access subset of biomedical articles available through the PubMed Central® (PMC) repository[3], hosted by the U.S. National Library of Medicine. Each article contains the full text and all figures in the article. Our analysis of the data set found that illustrations (i.e., graphs, charts) comprise nearly 80% of all figures in collection. For test purposes 7,245 figures were separated, and labeled manually - involving an expert annotator.

In the experiments 11 different modalities were considered, such as AN (angiography), EM (electron microscopy), FM (fluorescence microscopy), Illustration, Mixed (containing mixture of modalities in the same image), Photo, LM (light microscopy), CT (computer tomography), US (ultrasound), MRI (magnetic resonance imaging), and X-ray. An Unknown class was also designated to label all the images not fitting the modalities listed above. However, while the test set contains images belonging to "Unknown", the training material does not contains such class samples.

B. Experiments

The semi-automatic labeling procedure was performed on 300,000 unlabeled images, and the accuracy of the labeling was measured involving those 7,245 labeled images considered as test samples.

Concerning the different feature representations (CEDD, CLD, and modality term frequency), we considered the best performing ones based on some preliminary experiments. A similar trend can be discovered in the experiments described in [2]. The number of only three feature spaces (CEDD, CLD, and word frequency) is motivated by the fact that the lower

[2] http://www.imageclef.org/2012
[3] http://www.ncbi.nlm.nih.gov/pmc/

kmeans	Unanim.	Simple maj.	Disagreement
K = 100	59.20%	31.87%	8.91%
K = 200	59.04%	31.76%	9.20%
K = 300	61.01%	30.34%	8.65%

TABLE I
THE PERCENTAGE OF DATA WHERE THERE WAS UNANIMITY, SIMPLE MAJORITY OR DISAGREEMENT AMONG THE LABELS ASSIGNED AUTOMATICALLY BY THE LABELING PROCESS.

kmeans	kNN	k=1	k=5	k=11
K = 100	3 votes	40.00%	39.36%	38.12%
	2 votes	45.53%	46.32%	46.69%
	2 or 3 votes	46.10%	47.10%	**47.29%**
K = 200	3 votes	**48.14%**	47.55%	46.73%
	2 votes	43.02%	43.67%	44.46%
	2 or 3 votes	44.07%	45.45%	45.73%
K = 300	3 votes	**49.95%**	39.36%	38.12%
	2 votes	47.32%	48.14%	47.96%
	2 or 3 votes	48.26%	49.20%	49.06%

TABLE II
LABELING ACCURACY OF THE 300,000 SAMPLES MEASURED ON 7,245 SAMPLES USING kNN (K=1, 5 AND 11) USING THOSE LABELS WHERE UNANIMITY (3 VOTES), JUST SIMPLE MAJORITY (2 VOTES) OR AT LEAST A SIMPLE MAJORITY (2 OR 3 VOTES) WAS OBSERVED.

the number of feature representation, the less work is to be performed by the human expert. The choice of the clustering technique is motivated by the cluster compactness measure. The choice of the kmeans clustering and the experiments conducted using rather low K values (100, 200 and 300) also allows the annotator to assign a small number of labels, thus reducing the annotation workload.

Table I. shows the distribution of the labels selected based on unanimity (3 votes), simple majority (2 votes), and disagreement, respectively. For our subsequent experiment, only those samples were considered where at least simple majority (2 votes) was observed.

To measure the performance of the labeling, a k-nearest neighbor classifier (kNN) was considered. The use of more sophisticated methods like SVM or neural network would be a boost for the system. However, our goal is not only to discover the labels, but rather use these newly discovered data to train more sophisticated classifiers such as mentioned previously. Table II. presents the results using different voting schemes and different neighborhoods. For the kNN classification the CEDD feature was considered, while as for reference set, the 7,245 labeled images were used.

One might observe that the type of the vote applied in the process can serve also as a measurement, a sort of confidence value for the attached labels. In Table II. there is a clear view about the quality of the results. If a unanimity vote (3 votes) is considered only 40.00% of the labels are guessed correctly (for k = 1, K =100). If we allow simple majority or a mixture of the two, scores up to 47.29% can be achieved. When the number of clusters increases (K=200, 300), the trend is changing. It is more appropriate to trust the unanimity vote (49.49%) than mixed votes with lower confidence, which introduce some errors due to their lower confidence in the voting.

Modality	Accuracy [Recognized/Total]
AN	0% [0/268]
EM	32.18% [28/87]
FM	47.06% [32/68]
Illustration	**87.97% [1353/1538]**
LM	**68.54% [1684/2457]**
Mix	0% [0/0]
Photo	13.00% [39/300]
CT	46.30% [332/717]
US	0% [0/380]
MRI	**54.55% [126/231]**
X-ray	0% [0/1122]
Unknown	0% [0/77]

TABLE III
MODALITY RECOGNITION ACCURACY INCLUDING THE NUMBER OF RECOGNIZED AND THE NUMBER OF TOTAL SAMPLES FOR K=300 USING 1NN CLASSIFIER ON CEDD FEATURES.

In the modality-wise recognition, there are four classes (AN, US, X-ray and Unknown), which do not get recognized (see Table III.). The so-called Unknown class serves to gather the images not fitting in any of the other classes. After a thorough analysis of the different cluster labels during the manual labeling, we realized that cluster representatives labeled as X-ray appeared for the CEDD and modality term frequency features only, therefor none of the X-ray classes got vote. A similar situation can be noticed for the US and AN classes too. While during the manual labeling some clusters were labeled as Mixed, none of the test samples had been annotated as such, therefore no data was labeled by the process.

As described in Section IV-A, the image collection is not balanced. Approximately 80% of the data belongs to the class Illustration, which influences the clustering results, and therefore, the overall recognition scores of the other classes. A more balanced data set would produce more representatives clusters, and clusters such as AN, US, X-ray would also be present in some clusters, and the chance to correctly label images belonging to these classes would increase. However, the precondition for a proper clustering also depends on the discriminative power of the feature space considered during the partitioning.

In the Unknown class, the variability of images is high, which accounts for the low performance achieved during the recognition. The relatively high performing classes include modalities such as EM, FM, Illustration, LM, Photos and CT. The MRI (magnetic resonance imaging) class also performs rather well. However, this success is due to the discriminative power of the features used to represent the different views. The more compact the clustering, the better the chances to label them correctly; hence higher the chance to get a unanimous vote for each image in particular.

The confusion matrix analysis also supports the previous theory. Large confusions can be observed for classes like AN, X-ray, LM, and Photo as they are confused with the Illustration class. The LM class is often confused with the Photo class as well. The X-ray images are often recognized as being CT or MRI images. While in the first case the artifact is due to the unbalanced nature of the data, as the Illustration class is over-represented in the collection, in the latter case the X-ray,

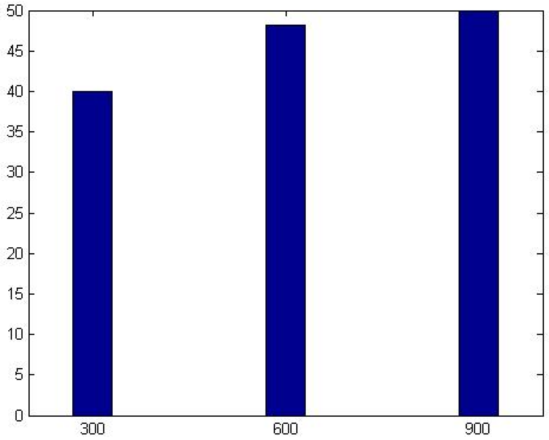


Fig. 3. Labeling accuracy as a function of the number of annotations for 300,000 images (unbalanced).

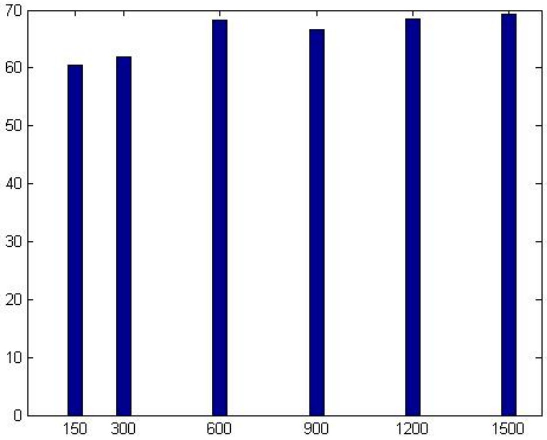


Fig. 4. Labeling accuracy as a function of the number of annotations for 7,245 images (balanced).

MRI and CT images share many visual similarities, thereby limiting the ability of features to separate such classes.

Labeling accuracy in function of the number of samples annotated by the expert, is shown in figure 3. An increase in performance can be observed for setups involving more labels provided by the human expert. However, it is seen that accuracy levels off rather quickly, so more labels will not improve the labeling performance. The balance between the number of annotations and the accuracy should be established in such a way as to achieve good labeling performance, while keeping the annotator-provided labels low.

To compare our results, we considered the recent work proposed by Rahman et al. [9]. Although the number of classes is slightly different, they use 2, 3, 4, 8 and 14 modality classes, respectively. Our results are comparable or outperform the accuracy figures reported by the authors. Their result (63.2%) outperforms ours only for a multimodal setup using hierarchical classification. While in our case only a global, high-level and unsupervised clustering was performed, in the modality recognition work proposed by Rahman and his colleagues, 15 different feature representations were utilized in an SVM-based supervised training scenario using a multitude of training samples.

To directly measure the impact of our method, we conducted a side experiment, using as input the 7,245 images considered for test purpose. We clustered the data using the different feature representations exactly as we presented previously. However, the huge advantage was that the labels of the cluster representatives were known, thereby allowing the manual annotator labeling to be replaced was by an automatic process. The rest of the process followed the method described in Section III. The images were partitioned using kmeans clustering into 50, 100, 200, 300, 400 and 500 clusters, respectively. Using only 150 annotations, an accuracy of 60.44% was achieved, while using 1500 labels the results achieved up to 69.26% accuracy, a score which outperforms even the most sophisticated system proposed recently [9]. It is also in the same range as the result reported in [2]. One fact should be noted, that this dataset is more balanced, hence the different clusters are more compact and representative, thereby the labeling makes more sense. The results based on the growing number of labels can be observed in figure 4.

During the experiments, we also conducted an evaluation of the time spent for the annotations. In total 1,700 images were annotated (for K=100, 200 and 300), and on average labeling 100 images took some 14 minutes. Using this analogy about 700 hours would be necessary to manually annotate the complete data collection. Using our proposed labeling method, the process can be reduced to less than three hours, involving the unsupervised clustering and the expert's labeling of the clusters.

V. Conclusion

In this paper we propose a fast and efficient method to label large medical image collections involving a small amount of human effort. Recently image modality detection has been of research interest for querying medical documents. Hence, the importance of such labeling - being able to train afterwards sophisticated classifiers which can determine the different image modalities with high accuracy.

Our labeling method borrows a concept from multiview learning. Each image is represented in different feature spaces, -in our case two image feature spaces and a textual feature space derived from the image caption, and clustered in an unsupervised manner by controlling the number of clusters to be produced. Each cluster representative (the closest sample to its centroid) is labeled by a human expert, and the label is propagated over the other images belonging to the same cluster. The final decision for a label is based on unanimity or simple majority vote. The choice of the voting strategy serves as a quality measure for the final label accuracy.

The experiments conducted on a large medical image collection showed promising results. Out of 300,000 images 149,850 images (49.95%) were labeled correctly involving only 900 labels provided by the annotator. While labeling the complete data collection by an expert would take some 700 hours of work, our method allows the task to be completed in less than 3 hours, though less accurately. The experiments conducted on a balanced collection show the huge potential of the method reaching high scores up to 69.26% accuracy.

Our strategy is not an end in itself. The labels discovered are considered only as future training material (admittedly including noise) for more sophisticated classifiers involving textual and visual features alike. Those classifiers, capable of learning high dimensional complex decision surfaces, would produce higher recognition scores than the voting.

VI. Future work

In order to further improve the quality of the labels assigned by out method, we would like to invest into some post-processing stage meant to filter out the incorrect labels [22], [23] or apply training strategies able to handle noisy data [24].

VII. Acknowledgment

This research is supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine, and Lister Hill National Center for Biomedical Communications (LHNCBC).

.

References

[1] H. Müller, J. Kalpathy-Cramer, D. Demner-Fushman, and S. Antani, "Creating a classification of image types in the medical literature for visual categorization," in *SPIE Medical Imaging*, 2012.

[2] D. You, M. M. Rahman, S. Antani, D. Demner-Fushman, and G. R. Thoma, "Text- and content-based biomedical image modality classification," in *Proc. SPIE Medical Imaging*, 2013, pp. 86 740L–86 740L–8.

[3] H. Müller, A. G. S. de Herrera, J. Kalpathy-Cramer, D. Demner-Fushman, S. Antani, and I. Eggel, "Overview of the ImageCLEF 2012 medical image retrieval and classification tasks," 2012.

[4] S. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Inf. Theor.*, vol. 28, no. 2, pp. 129–137, Sep. 2006.

[5] R. Sibson, "SLINK: an optimally efficient algorithm for the single-link cluster method," *Comput. J.*, vol. 16, no. 1, pp. 30–34, 1973.

[6] T. Kohonen, M. R. Schroeder, and T. S. Huang, Eds., *Self-Organizing Maps*, 3rd ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2001.

[7] B. Fritzke, "A growing neural gas network learns topologies," in *NIPS*, 1994, pp. 625–632.

[8] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[9] M. Rahman, D. You, M. Simpson, S. K. Antani, D. Demner-Fushman, and G. R. Thoma, "Multimodal biomedical image retrieval using hierarchical classification and modality fusion," *International Journal of Multimedia Information Retrieval*, vol. 2, no. 3, pp. 159–173, 2013.

[10] B. Settles, "Active Learning Literature Survey," University of Wisconsin–Madison, Tech. Rep. 1648, 2009.

[11] L. Rokach, *Pattern Classification Using Ensemble Methods*, ser. Series in Machine Perception and Artificial Intelligence. World Scientific Publishing Company, 2009.

[12] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[13] Z.-H. Zhou, "When semi-supervised learning meets ensemble learning," in *MCS*, 2009, pp. 529–538.

[14] J. Li, H. Mouchère, and C. Viard-Gaudin, "An annotation assistance system using an unsupervised codebook composed of handwritten graphical multi-stroke symbols," *Pattern Recognition Letters*, pp. on–line, Dec. 2012.

[15] A. H. Toselli, V. Romero, M. Pastor, and E. Vidal, "Multimodal interactive transcription of text images." *Pattern Recognition*, vol. 43, no. 5, pp. 1814–1825, 2010.

[16] S. Vajda, A. Junaidi, and G. A. Fink, "A semi-supervised ensemble learning approach for character labeling with minimal human effort," in *ICDAR*, 2011, pp. 259–263.

[17] J. Richarz, S. Vajda, R. Grzeszick, and G. A. Fink, "Semi-supervised learning for character recognition in historical archive documents," *Pattern Recognition*, vol. 47, no. 3, pp. 1011–1020, 2014.

[18] M. S. Simpson, M. M. Rahman, S. Phadnis, E. Apostolova, D. Demner-Fushman, S. Antani, and G. R. Thoma, "Text and content-based approaches to image modality classification and retrieval for the imageclef 2011 medical retrieval track," in *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[19] J. He, A.-H. Tan, C.-L. Tan, and S.-Y. Sung, *On Quantitative Evaluation of Clustering Systems*. Kluwer Academic Publishers, 2003.

[20] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–227, Feb. 1979.

[21] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[22] D. Guan, W. Yuan, Y.-K. Lee, and S. Lee, "Identifying mislabeled training data with the aid of unlabeled data," *Applied Intelligence*, vol. 35, no. 3, pp. 345–358, 2011.

[23] M. Guo, Y. Liu, J. Li, H. Li, and B. Xu, "A knowledge based approach for tackling mislabeled multi-class big social data," in *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, 2014, pp. 349–363.

[24] T. Leung, Y. Song, and J. Zhang, "Handling label noise in video classification via multiple instance learning." in *ICCV*, D. N. Metaxas, L. Quan, A. Sanfeliu, and L. J. V. Gool, Eds. IEEE, 2011, pp. 2056–2063.