Automated Medical Image Modality Recognition by Fusion of Visual and Text Information

Noel Codella, Jonathan Connell, Sharath Pankanti, Michele Merler, and John R. Smith

IBM T.J. Watson Research Center,
1101 Kitchawan Rd., Yorktown Heights, NY, USA

Abstract. In this work, we present a framework for medical image modality recognition based on a fusion of both visual and text classification methods. Experiments are performed on the public ImageCLEF 2013 medical image modality dataset, which provides figure images and associated fulltext articles from PubMed as components of the benchmark. The presented visual-based system creates ensemble models across a broad set of visual features using a multi-stage learning approach that best optimizes per-class feature selection while simultaneously utilizing all available data for training. The text subsystem uses a pseudo-probabilistic scoring method based on detection of suggestive patterns, analyzing both the figure captions and mentions of the figures in the main text. Our proposed system yields state-of-the-art performance in all 3 categories of visual-only (82.2%), text-only (69.6%), and fusion tasks (83.5%).

Keywords: medical image, modality, image recognition, image classification, text, visual, fusion.

1 Introduction

Medical image data has been growing by 20-40% every year [1], while the number of physicians per capita in the United States has remained relatively flat since the 1990s [2]. This trend makes automatic classification and categorization of medical images important for clinicians to handle increasing workload demands placed on them.

To facilitate research and development in the field of medical imaging, the Image Cross-Language Evaluation Forum (ImageCLEF) has organized a medical image modality benchmark [3,4,5], populated from Pubmed journal article figure images, text, and captions. The recognition tasks involved visual and textual information, or a fusion of both. In 2013, there were 10 performers with submitted runs, 2 of which consistently achieved top performance across all 3 run types: IBM [6], and the University of Ss Cyril and Methodius [7]. Visual methods employed by the University of Ss Cyril and Methodius involved extraction of Opponent SIFT features using a codebook size of 500 across a spatial pyramid (1x1, 2x2, 1x3). 1-vs-all classifiers were trained over these features, and multiclass decisions were rendered using a max operator. Text based approaches analyzed journal title and figure captions, using tokenization, stemming, and stop-word removal, along with TF-IDF weighting. Fusion between visual and text involved score averaging. IBM, which achieved state-of-art performance across all run types, used a fusion (score average) of visual approaches, including kernel fusion, linear kernel approximation, and optimized ensembles, over a set of low-level features.

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 487–495, 2014.

The text based system used an approach similar to TF-IDF weighting, while explicitly modeling uncertainty due to sample size, analyzing text from the figure captions only. Visual and text fusion involved score averaging.

In this work, we improve upon the state-of-the-art using a visual system that creates ensemble models across a variety of low and high level visual features, with a multi-stage learning approach that both optimizes per-class feature selection, as well as utilizing all data in the models. The text subsystem uses a pseudo-probabilistic scoring method based on detection of suggestive patterns, analyzing both the figure captions, as well as mentions of the figures in the main text. Fusion is performed using a learned weight that controls the relative contribution of the text subsystem versus the visual subsystem. 3 conditions are considered: 1) Visual data, where only image data is present, and 2) Text data, where English free-text descriptions are available, and 3) a fusion of both visual and text. We evaluate our system on the public ImageCLEF2013 benchmark, and achieve top results under all 3 conditions.

2 ImageCLEF2013 Dataset

For our experiments, we utilized the ImageCLEF 2013 Modality Classification dataset [4,5]. The dataset contains 31 categories, covering a variety of medical imaging modalities: Compound or multipane images (COMP), Ultrasound (DRUS), Magnetic Resonance (DRMR), Computerized Tomography (DRCT), X-Ray (DRXR), 2D Radiography, Angiography (DRAN), PET (DRPE), Combined modalities in one image (DRCO), Dermatology (DVDM), Endoscopy (DVEN), Other organs (DVOR), Electroencephalography (DSEE), Electrocardiography (DSEC), Electromyography (DSEM), Light microscopy (DMLI), Electron microscopy (DMEL), Transmission microscopy (DMTR), Fluorescence microscopy (DMFL), (D3DR) 3D reconstructions, Tables and forms (GTAB), Program listing (GPLI), Statistical figures (GFIG), Screenshots (GSCR), Flowcharts (GFLO), System overviews (GSYS), Gene sequence (GGEN), Chromatography (GGEL), Chemical structure (GCHE), Mathematics (GMAT), Non-clinical photos (GNCP), and Hand-drawn sketches (GHDR).

Data is broken into training and test partitions, with 2845 and 2582 images, respectively. Benchmark performance is measured by multiclass accuracy. Each image in the dataset is a figure from a Pubmed journal article, with the fulltext and figure captions available for analysis. This textual data serves as the input for our text classification system. The images serve as input for the visual classification system.

Training data is partitioned into two sets of 80% and 20%, where the second set is used to compute parameters for fusion of visual and text classification systems.

3 Visual Classification System

Our visual classification system is comprised of 3 components: 1) A set of visual features, 2) An unit modeling algorithm that trains SVMs over individual features, 3) An ensemble modeling algorithm that optimally combines SVMs in late fusion. One-vs-All ensemble SVM models are trained for each of the 31 categories of the ImageCLEF2013 Benchmark. SVM scores are logistically normalized, and a MAX operator is used to

Semantic Model Vector Dataset Statistics		
Parent	**# Sub-cat.**	**# Training Img.**
CT	6	10071
DX	91	12995
ECG	3	904
MR	10	1049
PET	3	245
SM	4	335
US	3	402

Example Sub-Categories						
CT	**DX**	**ECG**	**MR**	**PET**	**SM**	**US**
Brain	Arm	Normal	Knee	B&W	Bloodsmear	Fetus
Chest	Elbow	Fibrillation	Spine	Color	Sicklecell	Cardiac
Lung Cancer	Chest		Brain		Normal	
Normal Lung	Chest AP		Brain Axial			
Pneumothorax	Chest Lat		Brain Sagital			
	Knee		Hip			
	Abdomen					

Fig. 1. Top level hierarchy statistics and sub-categories of dataset used for model vector

render a multiclass judgment between the categories. Each component is described in the following subsections.

3.1 Visual Features

Our system utilizes a spectrum of visual features for image classification. These include some standard features, such as Color Histogram, Color Correlogram, Color Wavelets, Edge Histogram, GIST, SIFT [9], LBP [10], a variant of Multiscale Color LBP [11], Thumbnails, and Fourier Polar Pyramid.

Multiscale color LBP is an extension of the common grayscale LBP, whereby LBP descriptors are extracted across 4 color channels (Red, Green, Blue, and Hue), with 1 histogram per color channel. In our implementation, for each color channel, LBP descriptors are extracted across multiple scales (1/1, 1/2, 1/4, and 1/8th image size), and aggregated into the same histogram, weighted by the inverse of the scale. For a 59-bin LBP histogram, this results in 59*4 = 236 total bins, including all scales.

The Fourier polar pyramid is similar to the curvelet feature [12], whereby each element of the feature vector represents the average of some region of Fourier-Mellin space. However, the regions are partitioned into a pyramid structure, introducing various degrees of scale and rotation invariance.

Each feature is extracted over a range of spatial granularities, or subpartitions, of the image. These divisions include global (entire image), pyramid (1x1, 2x2), pyramid-3 (1x1, 3x3), pyramid-23 (1x1, 2x2, 3x3), grid (5x5), and grid7 (7x7). For each granularity, the features for each subpartition are concatenated.

In addition to low-level features, we examine the performance improvements of learning over semantic model vectors, which have been introduced in previous work [6]. Our semantic model vector is trained from 26,000 images collected from multiple sources and organized into a hierarchical taxonomy structure of 120 categories of various radiological categories. The images used to train our semantic model vector were acquired from web crawled data, the Image Retrieval in Medical Applications (IRMA) 2009 dataset [13], The Cancer Imaging Archives (TCIA) [14], Cornell University SIMBA CT Dataset [15], and the Japanese Society of Radiological Technology (JSRT) [16]. Web crawl was performed using the Microsoft Bing search engine, with queries as the semantic concept label. The top 500 results were saved for each concept, and cleaned by a human labeller. Each dimension refers to a subcategory over 7 modality domains of CT, X-Ray, ECG, MRI, PET, Ultrasound, and Slide Microscopy. The rest of the dimensions represent subcategories covering body regions (such as chest, arm, leg, head, brain, neck, etc.), views (such as anterio-posterior, coronal, saggital), and

disease states (such as pneumothorax) under each modality. Fig. 1 shows additional information regarding the dataset. The detectors for the model vector were trained using a similar visual modeling pipeline described in this work.

3.2 Unit SVM Modeling

Our unit SVM models are trained using the following approaches: 1) grid search over a set of SVM parameters using 2-fold cross-validation, 2) Synthetic Minority Oversampling (SMOTE) [8] for imbalanced learning, and 3) Logistic SVM score normalization using fixed parameters (0.5 mapping to raw 0.0, 0.9 mapping to raw 1.0).

For our experiments, we performed grid search over 10 SVM parameters (containing variants of histogram intersection, RBF, and Chi2 kernels), and SMOTE [8] with 3 nearest-neighbor interpolation to compensate for data imbalance.

3.3 Ensemble Modeling Algorithm

Our visual system is based on a multi-stage ensemble modeling approach. In the first stage, training data is partitioned into a Learning and a Validation set, with proportions of 50% and 50%, respectively. Unit SVM models are trained for subsamples across individual features, early fusions of features, and data on the Learning partition.

Unit SVM models are fused by evaluating their combined performance on the Validation partition. Combinations are chosen using a process of forward model selection. Forward Model Selection first selects the best performing unit model, determined by performance on the validation partition, and then continually combines unit models that boost ensemble model performance the most at each step. The combination consists of the sum of SVM scores, normalized by the number of unit SVM models. When performance ceases to increase, the algorithm terminates. Any evaluation criteria can be chosen as the measurement of performance, such as average precision, accuracy, precision, or recall. For our experiments, average precision was used to evaluate performance.

Once the fusion parameters for the ensembles have been determined, unit SVM models are retrained on second and third stages, using 80% and 100% of the training data. Unit SVM models trained on the 80% dataset are used for computation of visual and text fusion, whereas unit SVM models trained on the full 100% dataset are used for final external test-set scoring. In these two stages, the ensemble parameters learned in the first stage are held constant. In this manner, we manage to optimize ensemble models and text fusion, while at the same time maximally utilizing the precious limited data for our unit SVM models.

3.4 Visual System Results

The performance of the best of each type of feature in our modeling system is shown in Fig. 2. Features are named by their type and spatial granularity in parentheses. The top performing single feature was the 120 dimension semantic model vector.

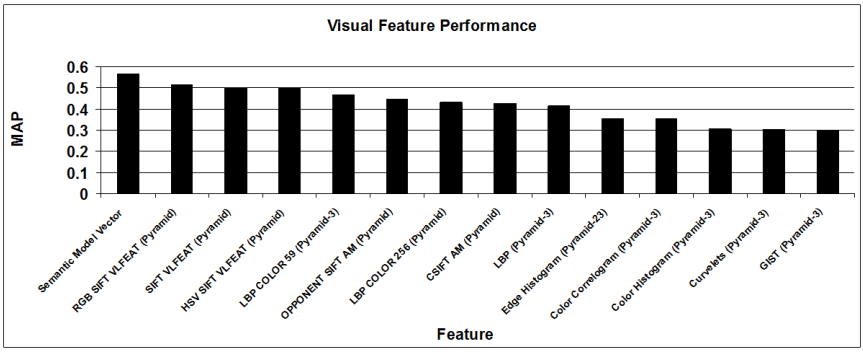


Fig. 2. Feature mean average precision results across all 31 categories

Following close behind were a variety of SIFT features, multiscale color LBP, and standard LBP variants.

Our visual system alone achieved a test accuracy of 82.2% using the semantic model vector trained from external datasets, and 81.17% without (ensemble fusion calculation is performed independently for both conditions). This represents an improvement over the current top ranked result of 80.79%.

4 Text Classification System

4.1 Text Features

The text-based modality classification system works by keyword spotting. A set of terms were manually selected from figure captions associated with example images of each modality in the training data. Generally, the heuristics used were to identify unusual words and phrases, ones that appeared in several examples for the selected modality, and those associated with the modality description (e.g. "gel electrophoresis"). These included name variants of the basic technique (e.g. "SEM", "sonogram") as well as specialization (e.g. "Doppler") and adjuncts (e.g. "gold"). Some were findings resulting from a technique (e.g. "papule") and adjectives applied to these findings (e.g. "erythematous"). Others were particular diseases (e.g. "vitiligo") or procedures (e.g. "dissection") linked to the modalites, or the portion of the anatomy they are typically applied to (e.g. "face", "brain", "heart"). These base terms were then consolidated into a smaller number of patterns by utilizing regular expressions to implement variable word endings. This cover plurals (e.g. "residue.*") and participles (e.g. "process.*"). In total, slightly over 400 terms were used.

The presence or absence of patterns (terms) in each caption was then fed into a pseudo-probabilistic combination scheme, based on simple conditional probability:

$$\begin{aligned} p(mod|term) &= p(term\&mod)/p(term) \\ &= [p(term|mod) * p_0(mod)]/p(term) \\ &= [p(term|mod)/p(term)] * p_0(mod) \end{aligned} \tag{1}$$

where $p_0(mod)$ is the prior expectation for that particular modality. This can be viewed as a one step update rule for the previous probability estimate of the modality given the presence of that term. The various terms can thus be chained together to yield a final estimate:

$$\begin{aligned} p'(mod) &= p_0(mod) * \textstyle\prod_j (p(term_j|mod)/p(term_j)) \\ &= exp(ln(p_0(mod)) + \textstyle\sum_j ln(p(term_j|mod)/p(term_j))) \end{aligned} \tag{2}$$

As shown, it is more convenient to perform this as a summation in the logarithmic domain. Note that we only increment the sum with the specified value only if that particular term is found. Skipping the increment means incrementing by 0, which implies $p(term_j|mod) = p(term_j)$ so there is no information gain. Yet the explicit absence of a certain pattern can also be significant. Thus the final scoring function is extended as shown below.

$$\begin{aligned} score(mod) = ln(p'(mod)) &= ln(p_0(mod)) \\ &+ \textstyle\sum_j ln(p(term_j|mod)/p(term_j)) \\ &+ \textstyle\sum_j ln(p(\sim term_j|mod)/p(\sim term_j)) \end{aligned} \tag{3}$$

This scoring scheme is only pseudo-probabilistic because we increment the sum for every term found, even though the terms themselves are not independent (among other things there are many terms coupled through the same modality). We also increment the sum multiple times if the term occurs multiple times in a caption, although these occurrences are obviously not independent. Still, the scoring function retains the feel of TF-IDF (term frequency, inverse document frequency) as used in classical information retrieval. In the end, for each image we separately compute the score for each modality and then select the modality with the highest value as the answer.

However this formulation has problems with term probabilities near 0 or 1 (which implies p(˜term) near 0) since division by 0 is undefined. This is particularly problematic for small training samples. If a term was never seen for a certain modality in training, then its presence in a test example implies that modality is not likely. Similarly, if a term was always seen for some class (e.g. “Xray” for DRXR) then its (possibly inadvertent) absence in a test case implies the modality may not be likely. This can happen for captions such as “See text for description of figure”.

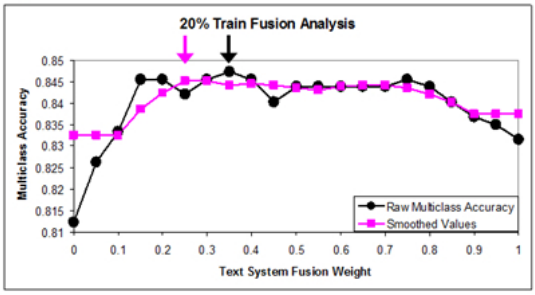
Therefore we “soften” the probability ratios since only a finite sample has been seen. We do this by considering the addition of u new examples which buck the observed trend.

$$\begin{aligned} p(x) &= n_x/n_{total} \\ p^+(x) &= (n_x + u)/(n_{total} + u) \quad highestimate \\ p^-(x) &= n_x/(n_{total} + u) \quad lowestimate \end{aligned} \tag{4}$$

This pushes the values closer to 0.5 (but is not symmetric around p(x)). As a candidate for u we use $k * sqrt(n_{total})$, which is loosely based on the Central Limit Theorem:

$$(\sigma^2(\sum_n x) = n * \sigma^2). \tag{5}$$

a)



b)

Run Type	Acc.
Visual Only	82.2%
Text Only	69.6%
Fusion (T=1.0)	81%
Fusion (T=0.35)	83.2%
Fusion (T=0.25)	83.5%

Fig. 3. a) Fusion weight grid search results. Arrows show weight selections for each method. **b)** Multiclass accuracies on test set. Text weight shown as T value.

These estimates are then used to set the score increment for a term. The presence of a term is only deemed significant when $p\text{-}(term_j \mid mod) > p+(term_j)$ and, similarly, for the absence of a term. Otherwise the term is considered irrelevant for that modality (any impact is presumed due to sampling noise) and thus its associated score increment is set to 0.

So far we have been treating the presence and the absence of some term equivalent. Yet a term may be present many times in a particular caption (multiple hits), but it can absent only once (single miss). Moreover the general probability of observing a particular term is fairly close to 0, whereas the probability of the absence of that term is nearly one. For these reasons we use different ks for the uncertainty in hits and misses. Optimal values of $k_h = 0.02$ and $k_m = 1.2$ were found using grid search with 5 fold cross-validation on the training dataset. This resulted in 76.6% accuracy in cross-validation. Note that setting either or bothk_h and k_m to 0 always yielded lower performance. Hence explicitly taking account of uncertainty seems to help.

4.2 Text Sources

In our system we additionally harvest all the sentences in the body of the article that refer to a particular figure and concatenate them into a pseudo-caption for the figure. We utilized an early fusion method in which we simply appended the mentions and the caption to form a super-caption. Using values $k_h = 0.05$ and $k_m = 0.4$ (derived as previously described) this yields a training cross-validation accuracy of 77.1%.

4.3 Text Results

The text based system alone, using the learned $k_h = 0.05$ and $k_m = 0.4$ on the early fusion of figure text mentions and captions yielded a test set performance of 69.6%, compared to 64.17% as the previously best performing system [6], representing a significant improvement in performance.

5 Fusion System

5.1 Algorithm

For fusion between Visual and Text, we summed the normalized score confidences from each system (visual and text), and performed grid search between 0.0 to 1.0 at step sizes of 0.05 for a single weight for text (visual confidences were fixed with a weight of 1.0). The internal 20% datasplit was used to measure performance of the fusion, and ultimately determine the weight by that which maximized performance. In addition, grid search performance results were smoothed by taking the average of 5 datapoints, in order to reduce the risk of overfitting.

5.2 Results

Fusion grid search is displayed in Fig. 3a. Both unsmoothed and smoothed grid search results are displayed. Fig. 3b shows the performance of each fusion, in addition to equal weighted fusion, on the external held-out test set. Our best performance of 83.5%, utilizing the semantic model vector in the visual system, and 83.11% without, is an improvement over the currently reported top performance of 81.68% [6].

6 Conclusion

In summary, we have presented a system for medical image modality recognition that is composed of two independent parts, visual and text, which are fused using grid search and smoothing. Our system achieves state-of-the-art performance in the ImageCLEF 2013 modality classification tasks of visual, text, and the fusion of both.

References

1. Frost & Sullivan: 2004 Healthcare Storage Report (2004), http://www.frost.com/c/10024/sublib/display-report.do
2. National Center for Health Statistics, http://www.cdc.gov/nchs/data/databriefs/db105.pdf
3. Muller, H., Seco de Herrera, A.G., Kalpathy-Cramer, J., Fushman, D.D., Antani, S., Eggel, I.: Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: Third International Conference of the CLEF Initiative Workshop (2012)
4. Caputo, B., et al.: ImageCLEF 2013: The Vision, the Data and the Open Challenges. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 250–268. Springer, Heidelberg (2013)
5. Seco de Herrera, A.G., Kalpathy-Cramer, J., Fushman, D.D., Antani, S., Muller, H.: Overview of the ImageCLEF 2013 medical tasks. In: American Medical Informatics Association (AMIA) ImageCLEF Medical Image Retrieval Workshop (2013)
6. Cao, L., Codella, N., Connell, J., Merler, M., Nguyen, Q., Pankanti, S., Smith, J.: IBM Multimedia Analytics @ ImageCLEF2013. In: American Medical Informatics Association (AMIA) ImageCLEF Medical Image Retrieval Workshop (2013)

7. Kitanovski, I., Dimitrovski, I., Loskovska, S.: FCSE at Medical Tasks of ImageCLEF 2013. In: American Medical Informatics Association (AMIA) ImageCLEF Medical Image Retrieval Workshop (2013)
8. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, P.: SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
9. van de Sande, K., Gevers, T., Snoek, C.: Evaluating Color Descriptors for Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1582–1596 (2010)
10. Ahonen, T., Hadid, A., Pietikäinen, M.: Face Recognition with Local Binary Patterns. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
11. Zhu, C., Bichot, C., Chen, L.: Multi-scale Color Local Binary Patterns for Visual Object Classes Recognition. In: 20th IAPR International Conference on Pattern Recognition (ICPR), pp. 3065–3068. IEEE Press, New York (2010)
12. Candès, E., Demanet, L., Donoho, D., Ying, L.: Fast Discrete Curvelet Transforms (2005), http://www.curvelet.org/
13. Image Retrieval in Medical Applications, http://ganymed.imib.rwth-aachen.de/irma/index_en.php
14. The Cancer Imaging Archive (TCIA), http://cancerimagingarchive.net/
15. Cornell University Vision and Analysis Group Public Image Databases, http://www.via.cornell.edu/visionx/simba/
16. Japanese Society of Radiological Technology, http://www.jsrt.or.jp/jsrt-db/eng.php