Text- and Content-based Approaches to Image Modality Detection and Retrieval for the ImageCLEF 2010 Medical Retrieval Track

Matthew Simpson, Md Mahmudur Rahman, Sachin Singhal, Dina Demner-Fushman, Sameer Antani, and George Thoma

Lister Hill National Center for Biomedical Communications
National Library of Medicine, NIH, Bethesda, MD, USA

Abstract. This article describes the participation of the Image and Text Integration (ITI) group from the U.S. National Library of Medicine (NLM) in the ImageCLEF 2010 medical retrieval track. Our methods encompass a variety of techniques relating to document summarization and text- and content-based image retrieval. Our text-based approaches utilize the Unified Medical Language System (UMLS) synonymy to identify concepts in information requests and image-related text in order to retrieve semantically relevant images. Our image content-based approaches utilize similarity metrics based on computed "visual concepts" and low-level image features to identify visually similar images. In this article we present an overview of the application of our methods to the modality detection, ad-hoc image retrieval, and case-based retrieval tasks and describe our submitted runs and results.

Keywords: Image Retrieval, Case-based Retreival, Image Modality

1 Introduction

This article describes the participation of the Image and Text Integration (ITI) group from the U.S. National Library of Medicine (NLM) in the ImageCLEF 2010 medical retrieval track.

ImgeCLEFmed'10 [12] consists of an image modality detection task and two medical retrieval tasks. For the modality detection task, the goal is to automatically classify given medical images according to eight modalities (e.g., CT or MRI). In the first retrieval task, a set of ad-hoc information requests is given, and the goal is to retrieve the most relevant images for each topic. Finally, in the second retrieval task, a set of case-based information requests is given, and the goal is to retrieve the most relevant articles describing similar cases.

In the following sections, we describe the text- and content-based features that comprise our image and case representation (Sections 2–3) and our methods for the modality detection (Section 4) and medical retrieval tasks (Sections 5–6). Our text-based retrieval approach relies on mapping information requests and image-related text to concepts in the Unified Medical Language System (UMLS) [8] Metathesaurus, and our modality detection and content-based retrieval approaches analogously rely on mapping the content of medical images to "visual concepts" using supervised machine learning techniques.

In Section 7, we describe our submitted runs, and in Section 8 we present our results. For the modality detection task, our best submission achieved a classification accuracy of 92% which was the 2nd ranked submission overall. For the retrieval tasks, our results were lower than expected yet reveal new insights which we anticipate will improve future work.

2 Image Representation

Images contained in biomedical articles can be represented using both text- and content-based features. Text-based features include text that pertains to an image, such as in captions and "mentions" (snippets of text within the body of an article that discuss an image), and content-based features include information derived from the image itself, such as shapes, colors and textures. We describe our text- and content-based image representations below.

2.1 Text-Based Features

We represent each image in the ImageCLEFmed'10 collection [12] as a structured document of image-related text. Our representation includes the title, abstract, and MeSH terms[1] of the article in which the image appears as well as the image's caption and mention.

We organize the content of an image's caption into the well-formed clinical question framework following the method described by Demner-Fushman and Lin [3]. Extractors identify UMLS concepts related to problems, interventions, age, anatomy, drugs, and image modality. We assign one of the eight modality classes to an image according to the extracted modality terms. Additionally, we extract textual Regions of Interest (ROIs) from image captions. A textual ROI is a noun phrase describing the content of an interesting region of an image which is identified within a caption by a pointer. For example, in the caption "MR image reveals hypointense indeterminate nodule (arrow)," the word *arrow* points to the ROI containing a *hypointense indeterminate nodule.*

The above structured documents can be indexed and searched with a traditional search engine or the extracted concepts may be combined with additional features (discussed below) for use in a multimodal representation. For the latter approach, "keywords" in a structured document D_j can be represented as an N-dimensional feature vector

$$\mathbf{f}_j^{\text{keyword}} = [w_{j1}, w_{j2}, \cdots, w_{jN}]^{\mathrm{T}} \quad (1)$$

where w_{jk} denotes the weight (typically *tf-idf*) of keyword t_k in document D_j.

2.2 Image Content-Based Features

In addition to the above textual features, we also represent the visual content of images using various low-level global image features and several derived features intended to capture high-level semantic content.

[1] MeSH is a controlled vocabulary created by NLM to index biomedical articles.

Low-level Global Features We represent the spatial structure and global shape/edge features of images with the Color Layout Descriptor (CLD) and Edge Histogram Descriptor (EHD) of MPEG-7 [2]. CLD is extracted to form the feature vector $\mathbf{f}^{\text{cld}}$ and EHD is extracted to form $\mathbf{f}^{\text{ehd}}$. Additionally, we extract the Color and Edge Directivity Descriptor (CEDD) and Fuzzy Color and Texture Histogram (FCTH) using the Lucene image retrieval (LIRE) library[2]. CEDD incorporates color and texture information into $\mathbf{f}^{\text{cedd}}$, and FCTH uses the high frequency bands of the Haar wavelet transform to form $\mathbf{f}^{\text{fcth}}$.

"Bag of Concepts" Feature In a heterogeneous medical image collection, it is possible to identify specific local patches in images that are perceptually and/or semantically distinguishable, such as homogeneous texture patterns in gray-level radiological images, differential color and texture structures in microscopic pathology and dermoscopic images. The variation in the local patches can be effectively modeled as "visual concepts" [13] by using supervised learning-based classification techniques, such as Support Vector Machines (SVMs).

For concept model generation, we utilize a multi-class SVM composed of binary SVM classifiers combined using the *one-against-one* strategy [5]. To train the SVM, a set of L labels are assigned as $C = \{c_1, \cdots, c_i, \cdots, c_L\}$, where each $c_i \in C$ characterizes a visual concept. The training set consists of local patches generated by a fixed-partition and represented by a combination of color and texture moment-based features. The input to the system is the feature vectors for patches along with their manually assigned concept labels. Concept labels are assigned by fixed partitioning each image I_j into l regions as $\{\mathbf{x}_{1_j}, \cdots, \mathbf{x}_{k_j}, \cdots, \mathbf{x}_{l_j}\}$, where each $\mathbf{x}_{k_j} \in \Re^d$ is a combined color and texture feature vector. For each $\mathbf{x}_{k_j}$, its category c_m is determined by the prediction of the multi-class SVM. Hence, instead of the low-level feature-based representation, an entire image is represented as a two-dimensional index linked to visual concepts. Based on this encoding scheme, an image I_j is represented as a vector of concepts

$$\mathbf{f}_j^{\text{concept}} = \; [w_{1_j}, \cdots, w_{i_j}, \cdots w_{L_j}]^{\mathrm{T}} \tag{2}$$

where each w_{i_j} denotes the *"tf-idf"* weight of a concept $c_i, 1 \leq i \leq L$ in image I_j, depending on its information content.

"Bag of Keypoints" Feature We also extract robust and invariant image features that are commonly termed affine region detectors [11]. These regions simply refer to a set of pixels or interest points, which are invariant to affine transformations as well as occlusion, lighting, and intra-class variations. We use the Harris-affine detector to locate interest points [10] as a large number of overlapping regions. We then associate with each interest point a vector descriptor invariant to viewpoint changes and, to some extent, illumination changes computed from the intensity pattern within the point. We use a local descriptor developed by Lowe [9] based on the Scale-Invariant Feature Transform (SIFT), to describe the information in a set of scale-invariant coordinates. The SIFT

[2] `http://freshmeat.net/projects/lirecbir/`

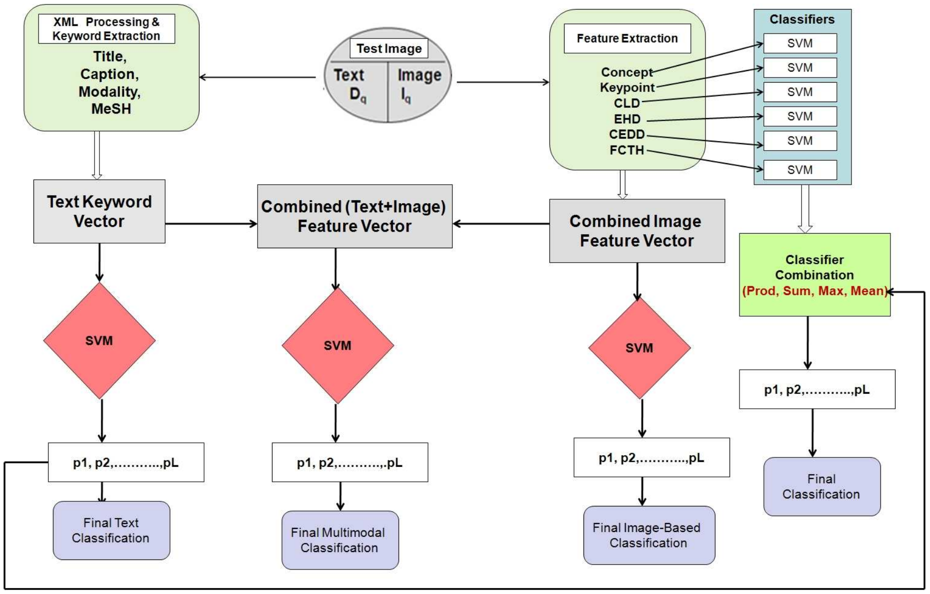
XML Processing & Keyword Extraction
Title, Caption, Modality, MeSH
Test Image
Text D_q
Image I_q
Feature Extraction
Concept
Keypoint
CLD
EHD
CEDD
FCTH
Classifiers
SVM
SVM
SVM
SVM
SVM
SVM
Text Keyword Vector
Combined (Text+Image) Feature Vector
Combined Image Feature Vector
Classifier Combination (Prod, Sum, Max, Mean)
SVM
SVM
SVM
p1, p2,...........,pL
p1, p2,..........,.pL
p1, p2,...........,pL
p1, p2,...........,pL
Final Text Classification
Final Multimodal Classification
Final Image-Based Classification
Final Classification

Fig. 1. Process flow diagram of the modality detection approach

descriptor is chosen to be invariant to viewpoint changes and, to some extent, illumination changes, and to discriminate between the regions. The above features are vector quantized by a self-organizing map (SOM)-based clustering. Finally, images are represented by a bag of these quantized features (i.e., a *bag of keypoints*). Hence, the model is applied to images by using a visual analogue of the *bag of words* model used in text retrieval [1].

3 Case Representation

We represent an article describing a patient's case by combining the textual features of each image contained in the article into a single surrogate document. Thus, each case representation consists of the article's title, abstract, and MeSH terms as well as the caption, mention and textual ROIs of each image contained in the article.

4 Modality Detection Task

Owing to their empirical success, we utilize multi-class SVMs to classifying images into eight image modalities [12] based on the above features. We compose multi-class SVMs by using the *one-against-one* method [5] for combining the pairwise classifications of each binary SVM.

Figure 1 shows the overall modality detection process. Textual and visual features can be used individually or combined to form a single feature vector, and the output of the multi-class SVMs can be used as separate predictions or “fused” to form a single classifier. We use the popular classifier combination techniques derived from Bayes’ theory (product, sum, maximum and mean rules) [4, 7] for fusing separate classifiers.

5 Ad-Hoc Image Retrieval Task

In this section we describe our text- and content-based approaches to image retrieval. The methods may be combined (e.g., by re-ranking retrieved images) to form multimodal approaches.

5.1 Text-Based Approach

We use the NLM-developed Essie [6] search engine to index our collection of structured image documents and retrieve relevant images. Key features of Essie that make it particularly well-suited to the medical retrieval track include its automatic expansion of query terms along synonymy relationships in the UMLS Metathesaurus and its ability to weight term occurrences according the location of the document in which they occur. For example, term occurrences in an image caption can be given a higher weight than those in the abstract of the article in which the image appears.

To construct queries for each topic, we organize each information request according to the well-formed clinical question framework, extracting UMLS concepts relating to problems, interventions, age, anatomy, drugs, and image modality. This procedure is identical to that described in Section 2.1

We use three methods of varying specificity for combining the extracted terms to form queries. First, the term-based method produces the `OR` of each extracted term. Second, the type-based method first applies the term-based method for each type (problem, intervention, etc.) and then `AND`s the result for each type group. Finally, the sentence-based method first applies the type-based method for each topic sentence and then `AND`s the result for each sentence.

Additionally, we may expand each query to include concepts taken from the definition of problems extracted from the original topic. Query expansion using problem definitions applies to each query construction method described above.

5.2 Image Content-Based Approach

Our content-based image retrieval approach is based on retrieving images that are visually similar to the given topic images. The similarity between a query image I_q and target image I_j is defined by

$$\mathrm{Sim}(I_q, I_j) = \sum_F \alpha^F \ \mathrm{Sim}^F(I_q, I_j) \tag{3}$$

where $F \in \{\mathrm{Concepts}, \mathrm{Keypoints}, \mathrm{EHD}, \mathrm{CLD}, \mathrm{CEDD}, \mathrm{FCTH}\}$ and α^F are the weights within the different image representations.

The feature weights are determined based on the 5-fold cross-validation (CV) accuracies of retrieval on the training set of images. The weights are normalized to $0 \leq \alpha^F \leq 1$ and $\sum \alpha^F = 1$ for $F \in \{\text{Concept}, \text{Keypoint}, \text{EHD}, \text{CLD}, \text{CEDD}, \text{FCTH}\}$. In addition, based on the online category prediction of a query image, precomputed category-specific feature weights (e.g., α^F) are utilized in the above linear combination of the similarity matching function.

6 Case-Based Retrieval Task

Our method for performing case-based retrieval is analogous to our text-based approach for ad-hoc image retrieval. Here, we use the Essie [6] search engine to index the structured case documents and construct queries for each case descriptions as described in Section 5.1.

7 Submitted Runs

In this section we describe each of our submitted runs for the modality detection, ad-hoc image retrieval, and case-based retrieval runs. Each run is identified by its (abbreviated) ID used with the `trec_eval` program and followed by a submission mode (textual, visual or mixed). All submitted runs are automatic.

7.1 Modality Detection Task

We submitted the following 10 runs for the modality detection task:

1. *result_image_combined (visual):* SVM classification combining an image's visual features (Concept, Keypoint, CLD, EHD, CEDD and FCTH) in a single feature vector.
2. *result_image_comb_cv (visual):* Classifier combination weighting the underlying classifiers according to their normalized cross validation accuracies. Visual features are each considered individually for SVM classification.
3. *result_image_comb_sum (visual):* Classifier combination using the "Sum" method of Bayes' theorem where an image's visual features are each considered individually for SVM classification.
4. *result_image_comb_max (visual):* Classifier combination using the "Maximum" method of Bayes' theorem where an images' visual features are each considered individually for SVM classification.
5. *result_text_title_caption_mod_mesh (textual):* SVM classification combining an image's textual features (*tf-idf* of keywords extracted from the title, caption, modality, and MeSH fields of an image's textual representation) as a single feature vector.
6. *result_text_image_combined (mixed):* SVM classification combining an image's textual and visual features as a single feature vector.
7. *result_text_image_comb_sum (mixed):* Classifier combination using the "Sum" method of Bayes' theorem where an image's textual and visual features are each considered individually for SVM classification.

8. *result_text_image_comb_prod (mixed):* Classifier combination using the "Product" method of Bayes' theorem where an image's textual and visual features are each considered individually for SVM classification.
9. *result_text_image_comb_max (mixed):* Classifier combination using the "Maximum" method of Bayes' theorem where an image's textual and visual features are each considered individually for SVM classification.
10. *result_text_image_comb_cv (mixed):* Classifier combination weighting classifiers according to their normalized cross validation accuracies. Textual and visual features are each considered individually for SVM classification.

7.2 Ad-hoc Image Retrieval Task

We submitted the following 10 runs for the ad-hoc image retrieval task:

1. *queries_terms (textual):* Essie search using term-based query construction.
2. *expanded_queries_terms (textual):* Essie search like run (1) but with query expansion using problem definitions.
3. *queries_terms_modalities (mixed):* Re-ranking of (1) according to topic image modality (determined by our modality detection approach) applied to the retrieved images' text-based modality class.
4. *fusion_cv_merge_max (visual):* Similarity matching using visual features (Concept, Keypoint, CLD, EHD, CEDD and FCTH) that are each weighted according to their normalized cross validation accuracy (from the modality detection task). All topic images produce individual result lists that are then merged based on the maximum score of each retrieved image.
5. *fusion_cv_merge_mean (visual):* Similarity matching according to run (4). All topic images produce individual result lists that are then merged based on the mean score of each retrieved image.
6. *fusion_cat_merge_max (visual):* Similarity matching using visual features that are each weighted according to online modality classification. All topic images produce individual result lists that are then merged based on the maximum score of each retrieved image.
7. *adhoc_queries_citations_cbir_cv_merge_max (mixed)*: Re-ranking of run (1) according to run (4).
8. *adhoc_exp_queries_citations_cbir_cv_merge_max (mixed)*: Re-ranking of run (2) according to run (4).
9. *adhoc_exp_queries_citations_cbir_cat_merge_max (mixed)*: Re-ranking of run (2) according to run (6).
10. *multimodal_rerank_roi_qe_merge (mixed)*: Re-ranking of run (1) according to visual Region of Interest (ROI) detection. Concept features from the ROIs of retrieved images are extracted and added to the Concept features of the original topic image (a form of query expansion). Similarity matching is then performed in the Concept space.

7.3 Case-based Retrieval Task

We submitted the following 10 runs for the case-based retrieval task:

ID	Mode	Accuracy (%)
result_text_image_combined	Mixed	92.00
result_text_image_comb_max	Mixed	91.00
result_text_image_comb_prod	Mixed	91.00
result_text_image_comb_cv	Mixed	89.00
result_text_title_caption_mod_mesh	Textual	89.00
result_text_image_comb_sum	Mixed	87.00
result_image_comb_cv	Visual	80.00
result_image_comb_sum	Visual	80.00
result_image_combined	Visual	79.00
result_image_comb_max	Visual	76.00

Table 1. Accuracy results for the modality detection task.

1. *queries_terms* (textual): Essie search using term-based query construction.
2. *queries_types* (textual): Essie search using type-based query construction.
3. *queries_backoff* (textual): Essie search using sentence-based query construction. If a query retrieves no results, it is sequentially relaxed using the type-based and term-based methods.
4. *expanded_queries_terms* (textual): Essie search like run (1) but with query expansion using problem definitions.
5. *expanded_queries_types* (textual): Essie search like run (2) but with query expansion using problem definitions.
6. *expanded_queries_backoff* (textual): Essie search like run (3) but with query expansion using problem definitions.
7. *queries_pico_backoff* (textual): Essie search matching terms extracted from the topic case with the structured captions in case representations. If a query retrieves no results, it is sequentially relaxed by removing all terms of a particular type (i.e., problem, intervention, etc.).
8. *queries_pico_ma* (textual): Essie search like run (7) but only considering terms related to image modality and anatomy.
9. *queries_cbir_without_case_backoff* (mixed): Essie search like run (3) but forming the query using the captions of the top 3 images retrieved using a content-based approach.
10. *queries_cbir_with_case_backoff* (mixed): Essie search like run (9) but also including the original topic case in the query.

8 Results

Table 1 presents the classification accuracy of our submitted runs for the modality detection task. *result_text_image_combined*, a multimodal approach, achieved the highest accuracy (92%) of our submitted runs and was ranked 2nd overall. Additionally, our text-based approach performed surprising well, achieving a classification accuracy of 89%, whereas our content-based approaches alone performed poorest. This result indicates that the combination of textual and

ID	Mode	Type	MAP
expanded_queries_terms	Textual	Automatic	0.19
queries_terms	Textual	Automatic	0.16
queries_terms_modalities	Mixed	Automatic	0.11
expanded_queries_terms_cbir_cv_merge_max	Mixed	Automatic	0.06
expanded_queries_terms_cbir_cat_merge_max	Mixed	Automatic	0.06
queries_terms_cbir_cv_merge_max	Mixed	Automatic	0.06
multimodal_rerank_roi_qe_merge	Mixed	Automatic	0.05
fusion_cv_merge_mean	Visual	Automatic	0.01
fusion_cv_merge_max	Visual	Automatic	0.00
fusion_cat_merge_max	Visual	Automatic	0.00

Table 2. Retrieval results for the ad-hoc image retrieval task

ID	Mode	Type	MAP
expanded_queries_backoff	Textual	Automatic	0.15
queries_pico_backoff	Textual	Automatic	0.14
queries_backoff	Textual	Automatic	0.13
expanded_queries_types	Textual	Automatic	0.12
queries_pico_MA	Textual	Automatic	0.11
queries_types	Textual	Automatic	0.10
expanded_queries_terms	Textual	Automatic	0.06
queries_terms	Textual	Automatic	0.05
queries_cbir_with_case_backoff	Mixed	Automatic	0.04
queries_cbir_without_case_backoff	Mixed	Automatic	0.03

Table 3. Retrieval results for the case-based retrieval task.

visual features can be leveraged to significantly improve the automatic modality classification of images found in biomedical articles.

Table 2 presents the results of our submitted runs for the ad-hoc image retrieval task. While our text-based submissions performed better than either the multimodal or content-based submissions, the Mean Average Precision (MAP) was much lower than expected given our prior experience [14]. We have determined that this discrepancy is likely due to noise in our text-based image representation, specifically concerning the extraction of image mentions and ROIs.

Our text-based submissions that expand queries to include concepts extracted from problem definitions show an improved MAP compared to submissions that do not perform query expansion in this way. This improvement is a promising result for use in future work.

Finally, Table 3 presents the results of our submitted runs for the case-based retrieval task. Given that our case representation is derived from our text-based image representation, the MAP of our case-based retrieval runs are also lower than expected. However, the submissions utilizing query expansion again show an improved in MAP, providing further evidence of its benefit.

9 Conclusion

This article describes the methods and results of the ITI group at the Communications Engineering Branch, NLM, for the ImageCLEF 2010 medical retrieval track. We submitted 10 runs each for the modality detection task and the ad-hoc and case-based retrieval tasks. For the modality detection task, our multimodal approach achieved a classification accuracy of 92%, which was the 2nd ranked submission overall. For the retrieval tasks, our results demonstrate that query expansion using the definitions of extracted terms is a promising direction for improving retrieval. While our results show no benefit in combining textual and visual features for the retrieval tasks, modality detection is improved when utilizing both text- and content-based approaches.

References

1. Baeza-Yates, R., Ribiero-Neto, B.: Modern Information Retrieval. Addison Wesley (1999)
2. Chang, S.F., Sikora, T., Puri, A.: Overview of the MPEG-7 standard. IEEE Transactions on Circuits and Systems for Video Technology 11(6), 688–695 (2001)
3. Demner-Fushman, D., Lin, J.: Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics 33(1), 63–103 (Mar 2007)
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons Ltd. (2001)
5. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. The Annals of Statistics 26(2), 451–471 (1998)
6. Ide, N.C., Loane, R.F., Demner-Fushman, D.: Essie: A concept-based search engine for structured biomedical text. Journal of the American Medical Informatics Association 1(3), 253–263 (2007)
7. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis 20(3), 226–2329 (1998)
8. Lindberg, D., Humphreys, B., McCray, A.: The unified medical language system. Methods of Information in Medicine 32(4), 281–291 (1993)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
10. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Proceedings of the European Conference on Computer Vision. pp. 128–142 (2002)
11. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. International Journal of Computer Vision 65(1–2), 43–72 (2005)
12. Mller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Charles E. Kahn, Jr., C.E., Hersh., W.: Overview of the clef 2010 medical image retrieval track. In: Working Notes of CLEF 2010 (2010)
13. Rahman, M.M., Antani, S., Thoma, G.: A medical image retrieval framework in correlation enhanced visual concept feature space. In: Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems (2009)
14. Simpson, M., Rahman, M.M., Demner-Fushman, D., Antani, S., Thoma, G.R.: Text- and content-based approaches to image retrieval for the imageclef 2009 medical retrieval track (2009)