

Biomedical Image Retrieval using Multimodal Context and Concept Feature Spaces

Md Mahmudur Rahman, Sameer K. Antani, Dina Demner Fushman, and
George R. Thoma

U.S. National Library of Medicine,
National Institutes of Health, Bethesda, MD, USA
{rahmanmm,santani,ddemner,gthoma}@mail.nih.gov

Abstract. This paper presents a unified medical image retrieval method that integrates visual features and text keywords using multimodal classification and filtering. For content-based image search, concepts derived from visual features are modeled using support vector machine (SVM)-based classification of local patches from local image regions. Text keywords from associated metadata provides the context and are indexed using the vector space model of information retrieval. The concept and context vectors are combined and trained for SVM classification at a global level for image modality (e.g., CT, MR, x-ray, etc.) detection. In this method, the probabilistic outputs from the modality categorization are used to filter images so that the search can be performed only on a candidate subset. An evaluation of the method on ImageCLEFmed'10 dataset of 77,000 images, XML annotations and topics results in a mean average precision (MAP) score of 0.1125. It demonstrates the effectiveness and efficiency of the proposed multimodal framework compared to using only a single modality or without using any classification information.

1 Introduction

The search for relevant and actionable information is key to achieving clinical and research goals in biomedicine. Biomedical information exists in different forms: as text, illustrations, and images in journal articles, documents, and other collections, and as patient cases in electronic health records. For example, in scientific publications, images are used to elucidate the text and can be easily understood in context. For example, Fig. 1 along with its caption are fairly informative in the context of the paper [1] “*Eosinophilic cellulitis-like reaction to subcutaneous etanercept injection*”. Taken out of context, the caption provides little information about the image, and the image does not provide enough information about the nature of the skin reaction. This example illustrates both the problem of finding text that provides sufficient information about the image without introducing irrelevant information, and the potential benefits of combining information provided by the text and image.

While there is a substantial amount of completed and ongoing research in both the text and content based image retrieval (CBIR) in medical domain, much



Figure 1: Reaction to intradermal adalimumab 1 to 2 days after the fourth dose

Fig. 1. Example image along with its caption in an article

remains to be done to see how effectively these two approaches can complement each other in an integrated framework. Biomedical image retrieval based on multimodal sources has been only recently gaining popularity due the large amount information sources [2, 3]. The results of the past medical retrieval tracks of ImageCLEF¹ suggest that the combination of visual and text based image searches provides better results than using the two different approaches individually.

Previous studies also have shown that imaging modality is an important aspect of medical retrieval [4]. In user-studies, clinicians have indicated that modality is one of the most important search filters that they would like to use. In fact, quality and speed of image retrieval from large biomedical collections can be improved by reducing the search space by filtering out irrelevant images and learning about the image categories. For example, to search “posteroanterior (PA) chest x-rays with enlarged heart”, automatically classified images in the collection could be organized according to modality (e.g., x-ray), body part (e.g., chest), and orientation (e.g., PA) criteria. Next, similarity matching can be performed between query and target images in the corresponding filtered subset to find “enlarged heart” as a distinct visual or textual concept. Some medical image search engines, such as Goldminer² and Yottalook³ allow users to limit

¹ <http://imageclef.org>

² <http://goldminer.ars.org/home.php>

³ <http://www.yottalook.com>

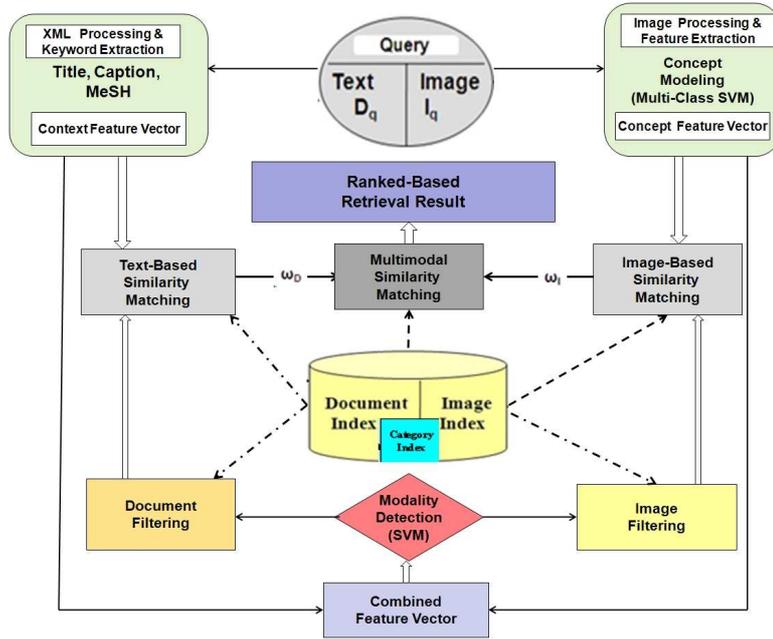


Fig. 2. Process flow diagram of the multimodal retrieval framework

the search results to a particular modality. However, this modality is typically extracted from the caption and is often not correct or present.

Studies have also shown that the modality can be extracted from the image itself using visual features. For example, in [5], the automatic categorization of 6231 radiological images into 81 categories is examined by utilizing a combination of low-level global texture features with low-resolution scaled images and a K-nearest-neighbors (KNN) classifier. In [6], the performances of two medical image categorization architectures with and without a learning scheme are evaluated on 10,322 images of 33 categories based on modality, body part, and orientation with a high accuracy rate of more than 95%. Although these approaches demonstrated promising results for medical image classification at a global level, they do not relate classification to retrieval in a direct manner, instead only stressed its value for image annotation and pre-filtering purposes.

To minimize limitations of low-level feature representations that result in the semantic gap and motivated by the successful use of machine learning in information retrieval (IR), we present a multimodal classification-based medical image retrieval method. We perform the multimodal search based on image classification and filtering using both textual and visual features. Text feature provide the context while the concept is derived from the visual features. In this framework, the modality specific information that is available as probabilistic outputs of SVM learning on the query and database images is used select the relevancy image subset. It is a primary goal of this work to develop improved informa-

tion retrieval techniques by moving beyond conventional text-based searching to combining both text and visual features extracted from collections of full-text biomedical journal articles, images and illustrations within these, and a collection of patient cases.

Fig. 2 shows the process flow diagram of the proposed multimodal retrieval approach. As can be seen from the top portion of Fig. 2, a search can be initiated simultaneously based on both text (left) and image parts (right) of a multimodal query and later the individual similarity scores are weighted combined (middle) for a final ranked result list. In addition, the text and image features are combined (bottom) to determine the query image modality from a SVM classification sub-system and based on that information only filtered images are accessed from the document and image indexes for further similarity matching.

The proposed approach and an evaluation of its efficacy are presented as follows: in Section 2, we briefly describe the image representation approach in concept and context feature spaces. Section 3 describes the multimodal search approach and Section 4 presents the modality detection and filtering approach based on the SVM classification. The experiments and the analysis of the results are presented in Section 5.

2 Image Feature Representation

The performance of a classification and/or retrieval system depends on the underlying image representation, usually in the form of a feature vector. The following feature vectors are generated at different levels of abstraction.

2.1 Context-Based Image Representation

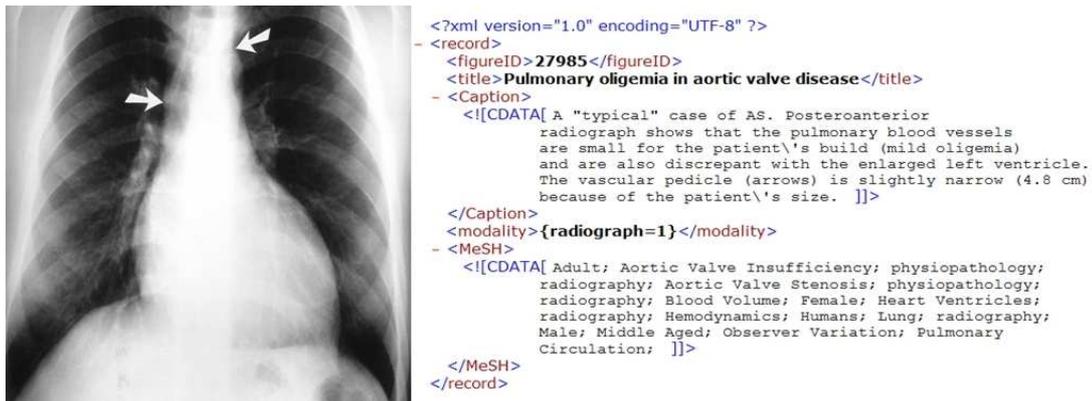


Fig. 3. Sample Chest x-ray image with annotation

For purposes of this research we use the ImageCLEFmed'10 dataset [4] that is provided to the participants of the evaluation. The collection comprises journal articles from two journals published by the Radiological Society of North America (RSNA), viz., *Radiographics* and *Radiology*. The collection includes full text from the articles and all images and figures within these. In all there are nearly 77,500 images from over 5,600 articles. The contents of this collection represent a broad and significant body of medical knowledge, which make the retrieval more challenging. The collection contains a variety of imaging modalities, image sizes, and resolutions and can be considered as a fairly a realistic set for evaluating medical image retrieval techniques.

Each image in the data set is represented as a structured document of image-related text, which is termed as *context* here. Now each image in the collection is attached to a manually annotated case or lab report in a XML file. It is necessary to index these annotation files into an easily accessible representation. There are a variety of indexing techniques which mostly rely on keywords or terms to represent the information content of documents [7]. In our case, information from only relevant tags are extracted and preprocessed by removing stop words that are considered to be of no importance for the actual retrieval process. Subsequently, the remaining words are reduced to their stems, which finally form the index terms or keywords of the annotation files. Next, the annotation files (document) are modeled as a vector of words based on the popular vector space model (VSM) of IR [7]. Our representation includes the title, and MeSH terms of the article in which the image appears as well as caption of the images. Fig. 3 shows an example chest x-ray image from the collection along with its annotation which is generated from the article where the image appears.

Let $T = \{t_1, t_2, \dots, t_N\}$ denote the set of terms in the collection. Then it can represent a document D_j as vector in a N -dimensional space as $\mathbf{f}_j^D = [w_{j1}, w_{j2}, \dots, w_{jN}]^T$. The element w_{jk} denotes the weight of term t_k in document D_j , depending on its information content. A weighting scheme has two components: a global weight and a local weight. The global importance of a term is indicating its overall importance in the entire collection, weighting all occurrences of the term with the same value. The popular *tf-idf* term-weighting scheme is used in this work, where the local weight is denoted as $L_{jk} = \log(f_{jk}) + 1$, f_{jk} is the frequency of occurrence of keyword t_k in document D_j . The global weight G_k is denoted as inverse document frequency as $G_k = \log(M/M_k)$, for $i = (1, \dots, N)$, where M_k be the number of documents in which t_k is found and M is the total number of documents in the collection. Finally, the element w_{jk} is expressed as the product of local and global weight as $w_{jk} = L_{jk} * G_k$. This weighting scheme amplifies the influence of terms, which occur often in a document (e.g., *tf* factor), but relative rarely in the whole collection of documents (e.g., *idf* factor) [7]. A query D_q is also represented as a vector of length N as $\mathbf{f}_q^D = [\hat{w}_{q1}, \dots, \hat{w}_{qi}, \dots, \hat{w}_{qN}]^T$.

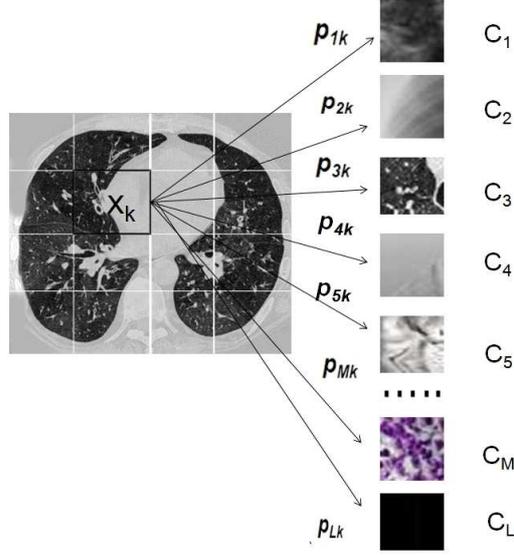


Fig. 4. Image encoding with probabilistic membership scores

2.2 Concepts-Based Image Representation

In a heterogeneous collection of medical images, it is possible to identify specific local patches that are perceptually and/or semantically distinguishable, such as homogeneous texture patterns in grey level radiological images, differential color and texture structures in microscopic pathology and dermoscopic images, etc. The variation in these local patches can be effectively modeled as visual keywords by using supervised learning based classification techniques, such as the support vector machine (SVM) [8]. In its basic formulation, the SVM is a binary classification method that constructs a decision surface and maximizing the inter-class boundary between the samples. A number of methods have been proposed for multi-class classification by solving many two-class problems and combining their predictions.

In this research, we utilize a multi-class classification method by combining all pairwise comparisons of binary SVM classifiers, known as *one-against-one* or pairwise coupling (PWC) [9]. PWC constructs binary SVM's between all possible pairs of classes. Hence, for L classes, this method uses $L * (L - 1)/2$ binary classifiers that individually compute a partial decision for classifying a data point (image). During the testing of a feature \mathbf{x} , each of the $L * (L - 1)/2$ classifier votes for one class. The winning class is the one with the largest number of accumulated votes.

In order to perform the learning, a set of L labels are assigned as $C = \{c_1, \dots, c_i, \dots, c_L\}$, where each $c_i \in C$ characterizes a visual concept. The training set of the local patches that are generated by a fixed-partition based

approach and represented by a combination of color and texture moment and edge histogram related features [10]. For SVM training, the initial input to the system is the feature vector set of the patches along with their manually assigned corresponding concept labels. Images in the data set are annotated with visual concept labels by fixed partitioning each image I_j into l regions as $\{\mathbf{x}_{1_j}, \dots, \mathbf{x}_{k_j}, \dots, \mathbf{x}_{l_j}\}$, where each $\mathbf{x}_{k_j} \in \mathbb{R}^d$ is a combined color and texture feature vector. For each \mathbf{x}_{k_j} , the visual concept probabilities are determined by the prediction of the multi-class SVMs as [9]

$$p_{ik_j} = P(y = i \mid \mathbf{x}_{k_j}), \quad 1 \leq i \leq L. \quad (1)$$

For example, Fig. 4 shows a particular region in a segmented image and its probabilistic membership scores to different local concept categories. Finally, the category label of x_{k_j} is determined as c_m , which is the label of the category with the maximum probability score. Hence, the entire image is thus represented as a two-dimensional index linked to the visual concept labels. Based on this encoding scheme, an image I_j is represented as a vector of visual concepts as

$$\mathbf{f}_j^I = [w_{j1}, \dots, w_{ji}, \dots, w_{jL}]^T \quad (2)$$

where each w_{ji} corresponds to the normalized frequency of a concept c_i , $1 \leq i \leq L$ in image I_j . Here, the vector dimension equals to the number of local concept categories.

3 Multimodal Image Search

Let us consider q as a multi-modal query, which has an image part as I_q and a text part as D_q . The similarity between q and a multi-modal item j , which also has also two parts (e.g., image (concept) I_j and text (context) D_j), is defined as

$$\text{Sim}(q, j) = \omega_I \text{Sim}_{\text{concept}}(I_q, I_j) + \omega_D \text{Sim}_{\text{context}}(D_q, D_j) \quad (3)$$

Here, ω_I and ω_D are normalized inter-modality weights within the concept and context feature spaces, which subject to $0 \leq \omega_I, \omega_D \leq 1$ and $\omega_I + \omega_D = 1$. The effectiveness of the linear combination depends mainly on the choice of the modality weights, which can be found out experimentally.

In our multimodal framework, the individual image $\text{Sim}_{\text{concept}}(I_q, I_j)$ and text $\text{Sim}_{\text{context}}(D_q, D_j)$ based similarities are computed based on the Cosine distance measure [7]. In particular, similar documents (images) are expected to have small angles between their corresponding vectors. In many cases, the direction or angle of the vectors are a more reliable indication of the semantic similarities of the objects than the distance between the objects in the term-document space. Hence, to compare a query and document vector, the cosine similarity measure is applied as follows [7]

$$\text{Sim}_{\text{context}}(D_q, D_j) = \cos(\mathbf{f}_q^D, \mathbf{f}_j^D) = \frac{\sum_{i=1}^N w_{qi} * w_{ji}}{\sqrt{\sum_{i=1}^N (w_{qi})^2} * \sqrt{\sum_{i=1}^N (w_{ji})^2}} \quad (4)$$

where w_{qi} and w_{ji} are the weights of the term \mathbf{t}_i in D_q and D_j respectively. In a similar way, cosine similarity measure is applied to the concept feature vector.

Due to the large number of images and vector size, it might take considerable amount of times to retrieve images from a collection. In the following section, we present a filtering approach based on multi-class classification on the multimodal input feature vector described earlier.

4 Modality Detection and Filtering

The variation of the medical image categories (e.g., modalities) at a global level can also be effectively modeled by the multi-class SVM as described in the previous section. For the SVM training, the input is a feature vector set of training images in which each image is manually annotated with a single modality label selected out of the M modalities. So, a set of M labels are defined as $\{\omega_1, \dots, \omega_i, \dots, \omega_M\}$, where each ω_i characterizes the representative image modality. In this context, given a multimodal feature vector \mathbf{x} , which is a simple concatenation of the context and concept feature vectors, the multi-class estimates the probability or confidence scores of each category as

$$p_m = P(y = \omega_m | \mathbf{x}), \text{ for } 1 \leq m \leq M \quad (5)$$

The final category of a feature is determined based on the maximum probability score.

Algorithm 1 Multimodal Image Filtering

(Off-line): Select a set training images (docs) of M categories with associated category label for SVM learning. Perform SVM learning based on the input of the combined multimodal feature vector $[\mathbf{f}^D \cdot \mathbf{f}^I]$ for each training images (docs).

(Off-line): Predict the category of each database image by applying SVM and store the category vectors (Equation 6) of N database images as a category index along with the feature indexes.

(On-line): For a multimodal query image of parts I_q and D_q , determine the category vector as $\mathbf{p}_q = [p_{q1}, p_{q2}, \dots, p_{qM}]^T$.

for $j = 1$ to N **do**

 Consider the top ranked ($n < M$) category labels for I_q and I_j after sorting the elements in the category vectors.

 Construct the category label sets as S_q and S_j for the top ranked categories of I_q and I_j respectively. Here, $|S_q| = n$ and $|S_j| = n$.

if $(S_q \cap S_j \neq \emptyset)$ **then**

 Consider $I(D)_j$ for further similarity matching (Equation 3)

end if

end for

We finally utilize the information about category prediction of query and database images for image filtering to reduce the search space. The output of

the above classification approach form a M -dimensional category vector of an image $I(D)$ as follows

$$\mathbf{p}_j = [p_{j_1}, \dots, p_{j_m}, \dots, p_{j_M}]^T \quad (6)$$

Here, p_{j_m} , $1 \leq m \leq M$, denotes the probability or class confidence score that an image $I(D)_j$ belongs to the category ω_m in terms of the multimodal feature vector.

During the off-line indexing process, this output is stored as the category vector of the database images in a *category index* along with the feature indices. Similar feature extraction and category prediction stages are performed on-line when the system is searched using an unknown query image. The category vector of a query image $I(D)_q$ and the vectors of the database images from the category index are evaluated to identify candidate target images in the collection, thereby filtering out irrelevant images from further consideration. To minimize misclassification errors, instead of only considering the image categories based on the highest obtained probability values, $n < M$ nearest classes of the target images to the query image are considered.

The process validates for class overlap between the query and target images. Generally, the value of $n \ll M$ to prevent inclusion of distant classes and provide effective filtering. A target image is only selected for further matching if at least one common category is found out between the top n categories of the query image and itself. This further reduces the risk of searching wrong images due to misclassification. Steps of the filtering algorithm are presented in Algorithm 1.

5 Experiments & Results

To evaluate the retrieval effectiveness, experiments are performed on the ImageCLEFmed'10 benchmark medical image collection. The experimental results are generated based on the 16 ad hoc query topics (e.g., a short sentence or phrase describing the search request in a few words with one to three relevant images) that were initially generated based on a log file of Pubmed⁴. All topics were categorized with respect to the retrieval approaches expected to perform best, i.e., visual topics for CBIR, semantic topics for text retrieval and mixed topics for multi-modal retrieval. Each topic consisted of the query itself in three languages (English, German, French) and 2 to 3 example images for the visual part of the topic.

5.1 Training for SVM

A training set of about 2400 images provided by the ImageCLEFmed'10 is used for SVM training for modality detection. The images are classified into one of the 8 modalities (e.g., CT, MR, XR, etc.) as shown in Table 1.

⁴ <http://www.pubmed.gov>

Table 1. Image categories and number of training images

Modality	No. of Images
CT: Computerized tomography	314
GX: Graphics, typically drawing and graphs	355
MR: Magnetic resonance imaging	299
NM: Nuclear Medicine	204
PET: Positron emission tomography including PET/CT	285
PX: optical imaging including photographs, micrographs, gross pathology etc	330
US: ultrasound including (color) Doppler	307
XR: x-ray including x-ray angiography	296

Table 2. 10-Fold Cross Validation (CV) Accuracy

Feature	C	γ	Accuracy
Concept	100	0.0002	73.89%
Context (Caption)	20	0.0002	90.50%
Context (Caption+Title+MeSH)	20	0.0002	90.54%
Combined (Context + Concept)	200	0.00001	95.39%

For the SVM training, we utilized the radial basis function (RBF) as kernel. A 10-fold cross-validation (CV) is conducted to find the best values of the tunable parameters C and γ of the RBF kernel as shown in Table 2.

For the visual concept generation based on the SVM learning, 30 local concept categories are manually defined, such as tissues of lung or brain of CT or MRI, bone of chest, hand, or knee x-ray, microscopic blood or muscle cells, dark or white background, etc. The training set consists of less than 1% images of the entire collection. Each image in the training set is partitioned into an 8×8 grid generating 64 non-overlapping regions, which is proved to be effective to generate the local patches. Only the regions that conform to at least 80% of a particular concept category are selected and labeled with the corresponding category label due to the consideration of robustness to noise [10]. After finding the best values of the parameters $C = 200$ and $\gamma = 0.02$ of the RBF kernel with a 10-fold CV accuracy of 81.01%, they are utilized for the final training to generate the SVM model file. We utilized the *LIBSVM* software package [11] for implementing the multi-class SVM classifiers.

5.2 Performance Analysis

Results for different retrieval methods are computed using the latest version of TREC-EVAL⁵ software based on the relevant sets of all topics, which were created by the CLEF organizers by considering top retrieval results of all submitted runs of the participating groups. Results were evaluated using an interpolated (arithmetic) Mean Average Precisions (MAP) to test effectiveness, Geometric Mean Average Precision (GMAP) to test robustness, and Precision at rank 20 (P20).

Table 3. Retrieval Results based on the Query Topics (CLEF'10)

Feature	MAP	GMAP	Rprec	Bpref	P(20)
Concept	0.0010	0.0001	0.0049	0.0144	0.0063
Context	0.1058	0.0133	0.1261	0.1441	0.1906
Multimodal	0.0958	0.0133	0.1150	0.1605	0.1781
Multimodal (Filter)	0.1125	0.0159	0.1292	0.2176	0.1875

It is clear from Table 3 that the best MAP score (0.1125) is achieved when a multimodal search is performed in a filtered image set. Although, we achieved a lower MAP score compared to the text only search approach when no filtering is applied based on multimodal search. This result might be an indication that the query topics are more semantic in nature and mixing with image features only lower the precision when search is performed on the entire collection. The other scores (e.g., GMAP, Rprec, and Bpref) also slightly improved when we compare filtering and without filtering approaches as shown in Table 3. Finally, from the results, we can conjecture that the pre-filtering approach is indeed an effective one as the performances are always better when compared to the searched which were performed on the entire collection.

Further, an important benefit of searching on a filtered image set is gain in computation time. We tested the efficiency of the multimodal search scheme by comparing the average retrieval time for 16 query topics with and without applying the filtering scheme. The experiment was performed in an Intel Pentium Dual-Core CPU at 3.40 GHz with 3.5 GB of RAM running Microsoft Windows XP SP2 Professional operating system. The linear search time without filtering was twice as much as search on the filtered image set, suggesting that the proposed method is both effective and efficient.

⁵ <http://trec.nist.gov/trec-eval/>

6 Conclusions

In this paper, a novel framework for multi-modal interaction and integration is proposed for a diverse medical image collection with associated annotation of the case or lab reports. Unlike in many other approaches, where the search is performed with a single modality and without any classification information, we proposed to use the classification result directly in the retrieval loop and integrate the results obtained from both the text and imaging modalities. A standard image dataset has provided enough reliability for objective performance evaluation that demonstrates the efficacy of the proposed method.

Acknowledgment

This research is supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC). We thank the ImageCLEFmed [4] organizers for making the dataset available for the experiments.

References

1. Winfield, W., Lain, E., Horn, T., Hoskyn, J. : Eosinophilic cellulitislike reaction to subcutaneous etanercept injection. *Arch. Dermatol.* **142** (2) (2006) 218–220
2. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A. : A Review of Content-Based Image Retrieval Systems in Medical Applications Clinical Benefits and Future Directions. *Int. J. of Med. Inform.* **73** (1) (2004) 1–23
3. Wong, T.C. : *Medical Image Databases*. New York, LLC: Springer Verlag, 1998.
4. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Reisetter, J., Jr., C.E.K., Hersh, W.R.: Overview of the CLEF 2010 Medical Image Retrieval Track. In *CLEF (Notebook Papers/LABs/Workshops)*(2010)
5. Lehmann, T.M., Güld, M.O., Deselaers, T., Keysers, D., Schubert, H., Spitzer, K., Ney, H., Wein, B.B. : Automatic categorization of medical images for content-based retrieval and data mining. *Comput. Med. Imag. and Graph.* **29** (2005) 143–155
6. Florea, F., Müller, H., Rogozan, A., Geissbuhler, A., Darmoni, S. : Medical image categorization with MedIC and MedGIFT. *Proc. Med. Inform. Europe (MIE 2006)*, Maastricht, Netherlands, (2006) 3–11
7. Yates R.B., Neto, B.R. : *Modern Information Retrieval*. Addison Wesley, 1999.
8. Vapnik, V. : *Statistical Learning Theory*. New York, NY, Wiley, 1998.
9. Wu, T.F., Lin, C.J., Weng, R.C. : Probability Estimates for Multi-class Classification by Pairwise Coupling. *J. of Mach. Learn. Research.* **5** (2004) 975–1005
10. Rahman, M.M., Antani, S.K., Thoma, G.R. : A Medical Image Retrieval Framework in Correlation Enhanced Visual Concept Feature Space. *Proc. 22nd IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, August 3-4, 2009, Albuquerque, New Mexico, USA.
11. Chang C.C., Lin, C.J. : LIBSVM : a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2001.