

Biomedical Image Representation and Classification Using an Entropy Weighted Probabilistic Concept Feature Space

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

ABSTRACT

This paper presents a novel approach to biomedical image representation for classification by mapping image regions to local concepts and represent images in a weighted entropy based probabilistic feature space. In a heterogeneous collection of medical images, it is possible to identify specific local patches that are perceptually and/or semantically distinguishable. The variation of these patches is effectively modeled as local concepts based on their low-level features as inputs to a multi-class SVM classifier. The probability of occurrence of each concept in an image is measured by spreading and normalizing each region's class confidence score based on the probabilistic output of the classifier. Furthermore, importance of concepts is measured as Shannon entropy based on pixel values of image patches and used to refine the feature vector to overcome the limitation of the "TF-IDF"-based weighting. In addition, to take the localization information of concepts into consideration, each image each segmented into five overlapping regions and local concept feature vectors are generated from those regions to finally obtain a combined semi-global feature vector. A systematic evaluation of image classification on two biomedical image data sets demonstrates improvement of more than 10% for the proposed feature representation approach compared to the commonly used low level and visual word-based approaches.

1. INTRODUCTION

The number of medical images being generated, stored, and managed in clinics and hospitals is growing at a phenomenal rate due to the rapid advancement of software and hardware technology. These images of diverse modalities constitute an important source of anatomical and functional information for the diagnosis of diseases, medical research, and education.¹ Effectively and efficiently searching in these large image collections poses significant technical challenges as the characteristics of the medical images differ significantly from other general purpose images.

Quality and speed of retrieving medical images from large collections of multiple modalities can be improved by knowing image categories at first hand. A successful categorization of images would greatly enhance the performance of a retrieval system by filtering out irrelevant images and thereby reducing the search space. For example, to search Chest X-ray images with tuberculosis in a radiographic collection, database images at first can be pre-filtered with automatic categorization according to modality (e.g., X-ray) and body part (e.g., chest). The similarity matching can be performed between query and images in the filtered set with better accuracy and efficiency.²

Some medical image search engines, such as Goldminer^{*} and Yottalook[†] allow users to limit the search results to a particular modality. However, this modality is typically extracted from the caption and is often not correct or present. Modality classification recently has been getting significant attention and was introduced as a task in 2010 in the medical retrieval track of ImageCLEF[‡]. Many approaches have been explored in recent years to automatically classify medical image collections into multiple semantic categories using only low-level visual features.^{3,4} For example, the automatic categorization of 6231 radiological images into 81 categories is

Further author information: (Send correspondence to Md Mahmudur Rahman)

E-mail: rahmanmm@mail.nih.gov, Telephone: 1 301 435 3262

^{*}<http://goldminer.arrs.org/>

[†]<http://www.yottalook.com>

[‡]<http://imageclef.org>

examined by utilizing a combination of low-level global texture features with low-resolution scaled images and a K-nearest-neighbors (KNN) classifier.⁴

In an effort to minimize the semantic gap of low-level feature and image representation, the Bag-of-visual Words (BoW) based feature representation scheme become popular recently in medical domain.⁵ In this approach, generally the low-level visual features of local regions on points, such as color, texture, and so forth are vector quantized to generate the visual words. Although it has proved to be more effective for image representation similar to document representation in text retrieval, the unsupervised clustering to generate the words or dictionary largely neglects the semantic contexts of the local features. As a result, commonly generated visual words are still not as expressive as keywords in text documents.

This limitation also extends to image representation, where the quality of matching or correspondence between local feature to visual words is not always exact. During the image encoding process, a region is quantized with their nearest visual words by employing a general distance metric, such as Euclidean distance or L1-norm between their low-level features and the rest of the visual words are simply overlooked or ignored. In reality, there are usually several words with almost as closely match as the one detected for a particular image region. Although, two regions will be considered totally different if they match to different words even though they might be very similar or correlated to each other. Furthermore, since images are more complex than text documents for the task of classification and retrieval, only considering the visual word frequency at both image and collection level (such as, “TF-IDF” weighting) may not be sufficient to determine its importance for 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.⁶ The variation in these local patches can be effectively modeled as “concepts” by using supervised learning based classification techniques such as the Support Vector Machine (SVM).⁷ Here, concepts are more expressive semantically compared to the visual words generated by employing unsupervised clustering techniques.

To minimize the limitations of low-level and BoW-based feature representations that result in the semantic gap and motivated by the successful use of machine learning in information retrieval (IR), we present a concept-based feature representation scheme for medical image classification by exploiting the probabilistic outputs of the SVM classifier. In addition, to overcome the limitation of the “TF-IDF”-based weighting in commonly used BoW-based image representation, the importance of concepts is measured as Shannon entropy based on pixel values of image patches and used to refine the feature vector. Finally, to consider the localization information of concept distribution in images, we propose a semi-global concept based representation scheme. Each image each segmented into five overlapping regions based on the use of a grid of cells superimposed on the encoded images and local concept feature vectors are generated from those regions to finally form a combined multi-dimensional vector. The proposed feature extraction and representation scheme is robust against classification and quantization errors and some notion of concept location is taken into account.

2. IMAGE REPRESENTATION IN CONCEPT FEATURE SPACE

In order to perform image representation in a concept feature space, the first step is to generate a set of concepts from distinguished local image patches by employing a supervised learning technique. 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 frequency related features.

A multi-class SVM⁷-based classification method is used by combining all pairwise comparisons of binary SVM classifiers, known as *one-against-one* or pairwise coupling (PWC).⁸ 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.

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 local concept labels






| Patch Category | Visual Example | Avg. Entropy |
|----------------------|---|--------------|
| angio_coronary |  | 6.23 |
| background_black |  | 2.24 |
| background_grey |  | 0.71 |
| | | |
| xray_finger_bone |  | 6.62 |
| microscopy_cell_blue |  | 6.11 |

Figure 1. Average entropies of different example concept categories.

by partitioning each image I_j into an equivalent $r \times r$ grid of l vectors 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 d -dimensional combined feature vector.

For each \mathbf{x}_{k_j} , the local concept category probabilities are determined by the prediction of the multi-class SVMs as

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

Finally, the category label of \mathbf{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 concept labels. Based on this encoding scheme, an image I_j can be represented as a concept vector as

$$\mathbf{f}_j^{\text{Concept}} = [w_{1_j} \cdots w_{i_j} \cdots w_{L_j}]^T \quad (2)$$

where each w_{i_j} corresponds to the weighted concepts $c_i, 1 \leq i \leq L$ in image I_j , depending on its information content. In general, the popular “TF-IDF” term-weighting scheme is used, where the element w_{i_j} is expressed as the product of local and global weights.

However, this representation scheme captures only a coarse distribution of the concepts and is analogous to the distribution of quantized color in a global color histogram. The image representation based on the hard encoding scheme, i.e., to find only the best concept prototype for each region, is very sensitive to classification errors and ignores correlations among concepts. Two regions within an image will be considered different if their corresponding concept labels are predicted as different even though they might be very similar or correlated to each other.

To overcome this, we present a feature representation scheme by spreading each region’s membership values or confidence scores to all the local concept categories. During the image encoding process, the probabilistic membership values of each region to all concept prototypes are computed for an image I_j . Based on the probabilistic values of each region, an image I_j is represented as $\mathbf{f}_j^{\text{PConcept}} = [\hat{f}_{1_j} \cdots \hat{f}_{i_j} \cdots \hat{f}_{L_j}]^T$, where

$$\hat{f}_{i_j} = \sum_{k=1}^l p_{ik_j} P_k = \frac{1}{l} \sum_{k=1}^l p_{ik_j}; \quad \text{for } i = 1, 2, \dots, L \quad (3)$$

where p_{ik_j} is determined based on (1). Here, we consider each of the regions in an image being related to all the concepts via the membership values, such that the degree of association of the k_j -th region in I_j to the c_i concept is determined by distributing the membership values to the corresponding index of the vector. This

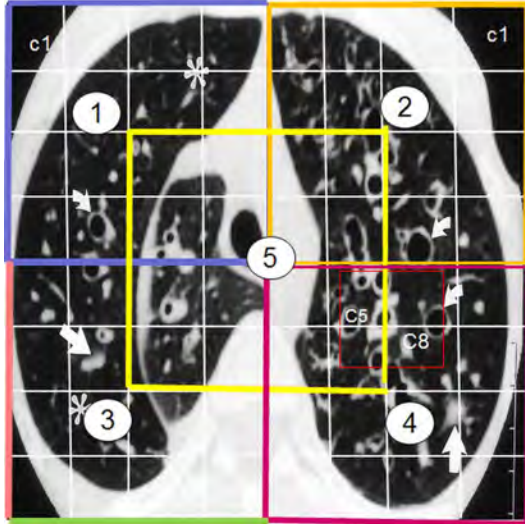


Figure 2. Segmentation of images with five overlapping regions.

vector representation considers not only the similarity of different region vectors from different concepts but also the dissimilarity of those region vectors mapped to the same concepts.

2.1 Entropy Weighted Concept Vector

In general, images are more complex than text documents and image patches contain more information about visual patterns of particular concepts than single keywords in text documents. Hence, measuring only the concept frequency at local image and global collection level (such as, “TF-IDF” weighting) might not be sufficient to determine their importance for image representation. Additional information is necessary.

An entropy of an image patch (e.g., concept) can be valuable to determine its importance for image representation. Entropy can be considered as the average number of bits one needs to represent a symbol in a stationary system, where the limited source symbols have fixed probabilities of occurrence.⁹ In other words, an entropy is a quantity which is used to describe the degree of randomness of an image. Low entropy image patches will have very little contrast and a relatively uniform color information, such as those appear in the background area of medical images, as shown in Figure 1 for the concepts black and grey background. On the other hand, high entropy image patches such as the foreground image region have a great deal of contrast from one pixel to the next and more important in capturing image content.

We observe that in medical images, various region-of-interest (ROI) in the foreground (such as, concepts like honeycomb, bronchi, etc.) appear as rough or fractal. So the entropy of image patches is helpful for their separation from uniform, and smoothly varying background. The concepts in the part of background usually appear more frequently in images (similar to “stopwords” in documents) than those in the part of foreground; hence they are less effective in representing images for classification or retrieval. Hence, we use the entropy as a measure of concept importance similar to the keyblock entropy in,¹⁰ which is the Shannon entropy of the patches based on pixel values. However, we only consider the intensity values of the grey-scale image patches instead of considering all three color channels in RGB space. The entropy of an image patch is expressed as

$$E = - \sum_{I=0}^{255} P(I) \log_2 P(I) \quad (4)$$

where $P(I)$ is probability of occurrence of the intensity value (grey value) I appears in a patch.

Now, to measure the entropy of each concept categories $c_i \in C$, we select all the training patches in each category, sum up their entropy values based on (5), and obtain the average values based on the number of

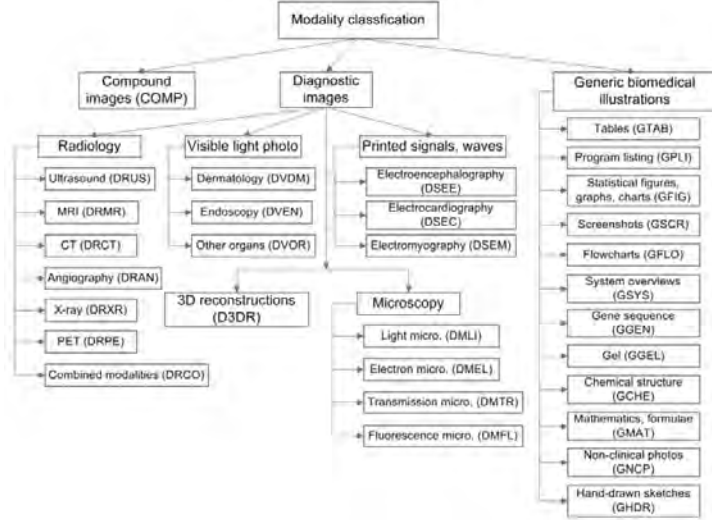


Figure 3. Modality classes in ImageCLEFmed'2012.³

training sample in each category as

$$e_i = \frac{\sum_{j=0}^{N_{c_i}} E_j}{N_{c_i}} \quad (5)$$

where, E_j is the entropy of a patch \mathbf{x}_j and N_{c_i} be the number of training sample in concept category c_i .

Finally, by using the entropy values of each concept categories, an image I_j is represented as a weighted probabilistic concept vector $\mathbf{f}_j^{\text{PConcept}(E)} = [\hat{f}_{1_j} \cdots \hat{f}_{i_j} \cdots \hat{f}_{L_j}]^T$, where, each \hat{f}_{i_j} is weighted as $\hat{f}_{i_j} * e_i$ to take into effect of entropy-based weight of each concept c_i .

2.2 Semi-Global Probabilistic Concept Feature

The concept-based feature representation technique described above only captures the global properties so that it cannot effectively characterize an image. Similar to the global color histogram, it does not take the localization information of concepts into consideration. To overcome this drawback, we present a semi-global feature representation scheme to capture information about concepts distribution in different regions.

In our method, each image is segmented into five overlapping regions based on the use of a grid of cells superimposed on the encoded images. These cells are obtained by first dividing the entire image space into 16 non-overlapping sub-images. From there, four connected sub-images are grouped to generate five different clusters of overlapping sub-regions. For example, the five overlapping regions (1-5) with different colors are shown in Figure 3 for a lung CT image. Each region is now considered as an independent image where a probabilistic concept vector $\mathbf{f}^{\text{PConcept}}$ is computed as described in Section 5. Finally, feature vectors of the five regions are combined to generate a multi-dimensional semi-global vector $\mathbf{f}^{\text{SG-PConcept}}$ of dimension of five times of the original global concept vector of entire image.

3. EVALUATION

We evaluate our feature representation method based on classification performances on two different biomedical image collections. The first collection (CLEF2012) is used for modality detection task (1,000 training and 1,000 test images) in ImageCLEFmed'12,³ where images are categorized modality wise to 31 different classes. Figure 4 shows the modality classes along with the class code in parenthesis.

The second collection (CLEF2007) comprises of 5,000 bio-medical images of 30 manually assigned disjoint global categories, which is a subset of a larger collection of six different data sets used for retrieval evaluation campaign in the medical image retrieval track in ImageCLEFmed'07.¹¹ In our collection, the images are classified

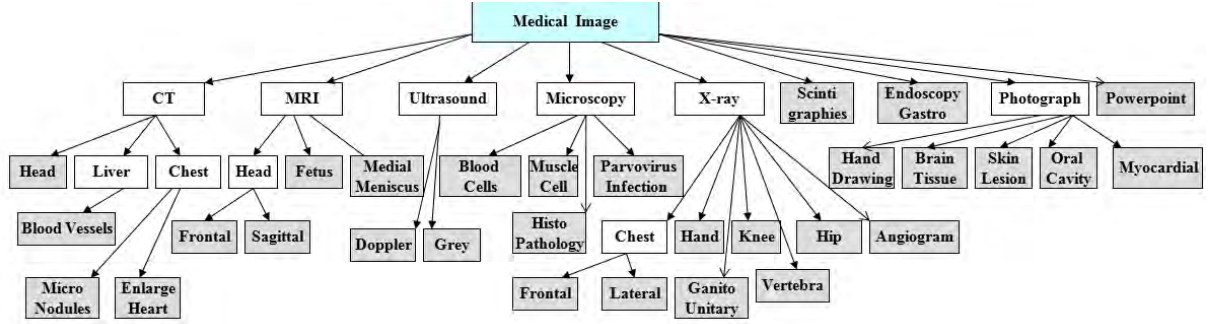


Figure 4. Classification structure of the ImageCLEFmed'07¹¹ data set.

into three levels as shown in Figure 4. In the first level, images are categorized according to the imaging modalities (e.g., X-ray, CT, MRI, etc.). At the next level, each of the modalities is further classified according to the examined body parts (e.g., head, chest, etc.) and finally it is further classified by orientation (e.g., frontal, sagittal, etc.) or distinct visual observation (e.g. CT liver images with large blood vessels). The disjoint categories are selected only from the leaf nodes (grey in color) to create the ground-truth data set. The categories are selected based on analyzing the visual and some mixed-mode query topics during the past several years of ImageCLEFmed retrieval campaign. Around 80% of the images are gray-level (e.g., x-ray, CT, MRI) and 20% are color images (e.g., microscopic pathology, histology, dermatology) with varying resolutions. We performed an 80/20 split of the data set to make separate training (4,000 images) and test sets (1,000 images) for classification performance evaluation. The split was made in such a way that the proportion of images for each class in both training and test sets was almost similar to the original distribution.

For concept model generation based on the SVM learning, 60 local concept categories are manually defined from image patches. The local concepts are selected as the ones that exhibit some meanings to the physicians with distinct visual appearances, such as different lung tissue patterns of X-ray and CT images, microscopic images of different color and texture patterns, and so on. The training set used for this purpose consist of around 19,000 patches to represent 60 concept categories. To generate the local patches, each image in the training set (we used only 2% images of entire data set) is re-sized to 256×256 pixels and partitioned into an 8×8 grid generating 64 non-overlapping regions of size 32×32 pixels. Only the regions that conform to at least 80% of a particular concept category are selected and labeled with the corresponding category label. Color moment, Tamura, Auto-Correlation and Edge frequency-based features are extracted and combined to form a 77-dimensional feature vector for all training patches.

Table 1. Cross Validation (CV) accuracies of local concept classification.

| Kernel | C | γ | Degree | Accuracy |
|--------|-----|----------|--------|----------|
| RBF | 100 | 0.08 | | 77.06% |
| Poly | 100 | | 1 | 75.80% |
| Poly | 100 | | 2 | 74.14% |

For the SVM training, we utilized both the radial basis function (RBF) and the polynomial kernels. There are two tunable parameters while using RBF kernels: C and γ . It is not known beforehand which values of C and γ are the best for the classification problem at hand. Hence, a 10-fold cross-validation (CV) is conducted. Basically pairs of (C, γ) are used and the one with the best CV accuracy is picked. We also experimented with the polynomial kernel of degree 1 and 2 with $C = 100$. However, the best accuracies are achieved by using the RBF kernel as shown in Table 1. Hence, after finding the best values of parameters C and γ of the RBF kernel, they are utilized for the final training to generate the model file for the concept learning.

Table 2 shows our SVM classification results on both the test data sets. The results are initially compared with

Table 2. Classification accuracies for different feature representations.

| Feature | Accuracy (CLEF2012) | Accuracy (CLEF2007) |
|-----------------------|---------------------|---------------------|
| CLD | 29.65% | 58.11% |
| EHD | 35.57% | 58.94% |
| CEDD | 41.58% | 68.54% |
| Concept (TF-IDF) | 40.48% | 70.38% |
| Concept (Entropy) | 41.62% | 72.69% |
| PConcept | 46.79% | 75.54% |
| PConcept (Entropy) | 47.95% | 76.01% |
| SG-PConcept | 45.99% | 79.15% |
| SG-PConcept (Entropy) | 47.88% | 79.56% |

few well known low-level features, such as the MPEG-7 based Color Layout Descriptor (CLD), Edge Histogram Descriptor (EHD)¹² and the Color Edge Directional Descriptor (CEDD) from the Lucene image retrieval (LIRE) library.¹³

From Table 2, the initial observation is that there is a large variation in accuracies between the two data sets. Since, images are categorized based only modality in the first data set (CLEF2012), performance of it is comparatively lower. Compound images (COMP) are often confused with images in the “Illustration” category and images under the “Illustration” and “Photo” category are also confused with each other due to their close perceptual similarities based on color and edge-based information.

Our best accuracy (47.95%) in CLEF2012 dataset is also much lower compared to the best accuracy (69.7%) result in ImageCLEFmed’12³ evaluation campaign using visual feature only. However, we only compare performances by using only a single feature, whereas the majority approaches in ImageCLEFmed’12 used a combination of multiple features to boost performance further. The main goal of this work is to show the improvement in classification accuracy from traditional BoW-based feature representation by refining the feature vector based on using appropriate weighting and localization information. Based on the results in Table 2, we can conjecture that our feature enhancement assumption proved to be correct.

From Table 2, it is observed that the accuracies of our baseline concept-based feature representation method based on “TF-IDF” weighting (e.g., Concept (TF-IDF)) even are better for both data sets compared to the low level features. However, when features are represented in the probabilistic concept feature space (e.g., PConcept) and entropy weighted (e.g., Concept (Entropy) and PConcept (Entropy)), we even achieved better classification accuracies compared to the baseline approach. Finally, when localization information of concepts is taken into consideration along with entropy in the semi-global feature space (SG-PConcept (Entropy)), we almost achieved a 10% improvement in accuracy from the baseline for both datasets.

4. CONCLUSION

This paper explored a new approach for improving accuracy of medical image classification by representing images in a weighted entropy based probabilistic feature space based on a soft annotation scheme. Instead of measuring the frequency of occurrence of each visual word (concept), we measure the probability of occurrence of each word in an image by spreading each concept class confidence score through the probabilistic output of the multi-class SVM classifier. In addition, the average entropy of each patch category is measured from the training samples and utilized that information to modify the feature vector. Finally, concept location information is taken into account by dividing each image into five overlapping regions. In summary, the proposed image representation schemes realize semantic abstraction via prior learning when compared to the representations based on the low-level features. Experimental results validated the assumption and showed that our approach is very effective

with around 10%-20% increase in accuracy compared to the well known low-level image features and BoW-based representation without any feature enhancement.

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 would like to thank the CLEF³ organizers for making the database available for the experiments.

REFERENCES

- [1] Müller H., Michoux N., Bandon D., and Geissbuhler A., “A Review of Content-Based Image Retrieval Systems in Medical Applications Clinical Benefits and Future Directions,” *International Journal of Medical Informatics* **73**(1) 1–23, (2004).
- [2] Rahman M. M., Antani S. K., and Thoma G. R., “A Learning-Based Similarity Fusion and Filtering Approach for Biomedical Image Retrieval Using SVM Classification and Relevance Feedback,” *IEEE Trans. On Information Technology in Biomedicine*, **15**(4) 640–646 (July 2011).
- [3] Müller H., de Herrera, A. G. S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., and Eggel, I., “Overview of the ImageCLEF 2012 Medical Image Retrieval and Classification Tasks,” in [CLEF (Online Working Notes/Labs/Workshop)], (2012).
- [4] Lehmann T. M., Güld M.O., Deselaers T., Keysers D., Schubert H., Spitzer K., Ney H., and Wein B.B., “Automatic categorization of medical images for content-based retrieval and data mining,” *Computerized Medical Imaging and Graphics* **29**, 143–155 (2005).
- [5] Juan C. Caicedo, Angel Cruz, and Fabio Gonzalez “Histopathology Image Classification Using Bag of Features and Kernel Functions,” *Artificial Intelligence in Medicine, AIME-09, Proc. LNCS*, **5651**, 126–135 (July 2009).
- [6] Rahman M. M., Antani S. K., and Thoma G. R., “A Medical Image Retrieval Framework in Correlation Enhanced Visual Concept Feature Space,” *22nd IEEE International Symposium on Computer-Based Medical Systems (CBMS)* (August 2009).
- [7] Vapnik V., [Statistical Learning Theory], Wiley, New York, NY (1998).
- [8] Wu T.F., Lin C.J., and Weng, R. C., “Probability Estimates for Multi-class Classification by Pairwise Coupling,” *Journal of Machine Learning Research* **5**, 975–1005 (2004).
- [9] Shannon C. E. and Weaver W., [The Mathematical Theory of Communication], University of Illinois Press, Urbana, IL (1949).
- [10] Zhu L., Tang C., and Zhang A., “Using Keyblock Statistics to Model Image Retrieval,” *Advances in Multimedia Information Processing PCM 2001, Proc. LNCS*, **2195**, 522–529 (2001).
- [11] H. Müller, T. Deselaers, E. Kim, C. Kalpathy, D. Jayashree, M. Thomas, P. Clough, and W. Hersh, “Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks,” *8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007), Proc. LNCS*, **5152** (2008).
- [12] Chang, S. F., Sikora, T., and Purl, A., “Overview of the MPEG-7 standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, **11**, 688–695 (Jun 2001).
- [13] Grubinger M., Clough P., Hanbury A., and Müller H., “Lire: lucene image retrieval: an extensible java CBIR library,” *Proc. of the 16th ACM international conference on Multimedia*, 1085–1088, (2008).