35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

LEARNING METRICS FOR CONTENT-BASED MEDICAL IMAGE RETRIEVAL

John Collins and Kazunori Okada

Computer Science Department, San Francisco State University, San Francisco, CA 94132
{johncoll,kazokada}@sfsu.edu

ABSTRACT

Application of content-based image retrieval (CBIR) to medical image analysis has recently become an active research field. While many previous studies have focused on the feature design, the metric design, another key CBIR component, has not been well investigated in this application context. This paper presents a medical CBIR that adapts its similarity metric from data by using information theoretic metric learning. Also we systematically compare our SIFT bag-of-words-based system with various plug-in similarity measures available in literature. The proposed systems are evaluated with the ImageCLEF-2011 benchmarking dataset. Our experimental results demonstrate the advantage of the proposed metric learning approach and $L1$ distance-based measures.

1. INTRODUCTION

In recent years, an application of content-based image retrieval (CBIR) [1, 2] to medical image analysis has become an active research field. Such medical CBIR (M-CBIR) focuses on retrieving medical images similar to a single or a set of query images without using semantic annotations and can be applied to various decision support problems in pathology and radiology [3]. Typically a M-CBIR involves a two-step procedure of feature extraction followed by similarity comparison, both of which are equally important for successful applications. Previous work on CBIR has led to the development of various feature designs, such as SIFT [4], SURF [5] and Gabor wavelets [6]. Despite the relative maturity of these feature designs, similarity measures in CBIR have not been investigated thoroughly. Previous studies on metric design in CBIR [7, 8, 9] are still few and the lack is especially evident for M-CBIR.

Addressing this shortcoming, we present our investigation on metric design for an M-CBIR application. Our contributions are two-fold. First, we propose an in-house M-CBIR system with information theoretic metric learning [10] that adapts its similarity measure according to known relevance side-information. Second, we report a comparative study with a comprehensive list of similarity measures of many types using a large dataset. Our experimental evaluation employs a public bench marking dataset available from ImageCLEF-2011, which includes various types (e.g., tomographic images, compositions, plots etc) of 2D digital photographs derived from figures in Radiology journal documents. Our results demonstrate the advantage of our metric learning approach and of $L1$-based measures that we tested.

This paper is organized as follows. Section 2 presents our metric learning method and other technical components of M-CBIR methods evaluated in this study. Section 3 outlines our experimental study: data, experimental design, and the results. Finally, Section 4 discusses our study's result and potential future work.

2. M-CBIR METHOD

2.1. Feature Design

A feature vector is extracted from each image in the well-known bag-of-words (BoW) scheme [11] with SIFT features [4] as described below.

Given a training dataset, we first construct a K-word *codebook* by 1) identifying and extracting SIFT features from all images in the dataset and 2) creating K representative features via K-means clustering over the extracted SIFT features. We use the standard algorithm for SIFT feature extraction, yielding a 128-element histogram of local gradient directions for 8 orientations in 16 tiles. We include 4 extra parameters of the 2 spatial coordinates of the SIFT keypoint, the scale parameter, and the dominant-orientation parameter, making the total dimension of our feature vector to be 132. Each SIFT feature vector is centered and normalized using Z-Score transform, and we randomly initialize K centers for the K-means clustering.

Given this codebook, the BoW method extracts a feature vector of length K for each new image by 1) performing the same SIFT feature extraction and 2) constructing a normalized histogram representing the frequency distribution of the extracted SIFT features with respect to the codebook. For each feature, we find the nearest-neighbor best match among the K representative codebook vectors that is closest to the input feature in Euclidean distance. Finally, we perform a number of standard feature transformation such as PCA and TF-IDF, for better retrieval performance.

We consider two types of post-extraction feature transformation: principal component analysis (PCA) [12] and term frequency-inverse document frequency (TF-IDF) [13]. PCA is a standard dimension reduction method which computes an eigen subspace of (BoW) feature vectors derived from the training dataset. Each new feature vector is then projected onto this subspace before similarity comparison. TF-IDF originally comes from textual data mining, whose goal is to penalize common words (i.e., codebook feature vectors) across the training dataset. The BoW feature vector described above corresponds to TF. IDF is computed for each codebook vector as an inverse frequency of training images that include the codebook vector as a match. Each new feature vector is then multiplied with the resulting IDF filter. We experiment with four types of feature transformation including combinations of PCA and TF-IDF: 1) PCA($\cdot$), 2) TF-IDF($\cdot$), 3) PCA(TF-IDF($\cdot$)), and 4) TF-IDF(PCA($\cdot$)).

2.2. Database Ranking by Similarity Comparison

Given a query image, the goal here is to rank database images according to their distance or similarity to the query. In some cases a query may consist of multiple images. In this case, we calculate the average similarity of the query set to each database image and use this average for the ranking. Many standard similarity measures

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

exist, but most take the form of a distance/dissimilarity measure in their natural expression except for some rare cases (e.g., cosine similarity). When considering a dissimilarity measure $d(x, y)$, we calculate similarity with its additive inverse by $1 - d(x, y)$ where x and y are appropriately scaled so that $d(x, y) \in [0, 1]$. We also abuse the term *metric* to indicate both similarity and dissimilarity measures in this paper. Strictly speaking, a metric is a distance function that satisfies three conditions of positive definiteness ($d(x, y) \geq 0$; $d(x, y) = 0$ *iff* $x = y$), symmetry ($d(x, y) = d(y, x)$), and the triangle inequality ($d(x, z) \leq d(x, y) + d(y, z)$). However, some standard dissimilarity measures we considered violate these condition (e.g., Kullback-Liebler divergence is not symmetric).

2.3. Metric Design by Learning

Metric Learning [14] is the process of adapting a metric of a set S according to side-information about the similarity or dissimilarity of some known datapoints in S. Let $\mathbf{x} = (x_1, .., x_n)$ represent the query image and $\mathbf{y} = (y_1, .., y_n)$ represent another image against which to be compared. Let $\boldsymbol{\lambda}$ denote an n-dimensional vector in which λ_i determines the weight given to the i-th feature $x_i \in \mathbf{x}$. A weighted L^2 metric on S can then be defined as $d_\lambda(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{N} \lambda_i (x_i - y_i)^2}, \forall \mathbf{x}, \mathbf{y} \in S$. A more general form is given by Mahalanobis distance,

$$d_A(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_A = \sqrt{(\mathbf{x} - \mathbf{y})^T A (\mathbf{x} - \mathbf{y})} \qquad (1)$$

where A is a symmetric, positive semi-definite matrix and $\boldsymbol{\lambda} = diag(A)$.

One idea of metric learning is to learn the appropriate weights $\boldsymbol{\lambda}$ or A from training data [14]. Different approaches have been reported in literature for such metric learning [15].

We adopt information theoretic metric learning (ITML) [10] in our M-CBIR system. ITML is a popular metric learning algorithm that uses an information-theoretic cost model which iteratively enforces pairwise similarity/dissimilarity constraints, yielding a learned matrix A of the Mahalanobis distance as an output.

The Mahalanobis distance is a bijection to a Gaussian distribution with its covariance set as an inverse of A. Exploiting this bijective property, ITML poses the metric learning problem as a convex optimization of a relative entropy between a pair of Gaussian distributions with the unknown A and the identity matrix I under the similarity/dissimilarity constraints,

$$\begin{aligned} & \min_{A \succ 0} \quad KL(p(\mathbf{x}; \mathbf{m}, A) || p(\mathbf{x}; \mathbf{m}, I)) && (2) \\ & \text{Subject to:} \quad d_A(\mathbf{x}_i, \mathbf{x}_j) \leq u \;\; (i, j) \in S \\ & \qquad\qquad\quad\; d_A(\mathbf{x}_i, \mathbf{x}_j) \geq l \;\; (i, j) \in D \end{aligned}$$

where S and D are the sets of similar and dissimilar points, respectively. This formulation regularizes the optimization problem so as to seek a metric that satisfies the given constraints and is closest to the Euclidean distance.

Davis et al. [10] demonstrated the equivalence of this metric learning formulation and low-rank kernel learning problem [16], yielding an efficient solution to the problem in (2) based on Bregman's method [17]. This dual ascent optimization method iteratively projects onto one constraint at a time with a closed-form projection update without a need of numerical eigenvalue decomposition and is thus efficient.

Note that a pairwise distance computation by Eq.(1) can also be realized by first performing a linear transformation $S \mapsto T = A^{1/2} S$ and by computing the L^2 distance for the pair in T. This linear transformation makes similar datapoints in S closer together and dissimilar datapoints farther apart in T and yields more computationally efficient pairwise distance computation. Adopting this property, we treat the ITML's result A as a post feature transformation and evaluate it with different similarity measures in our experiment.

2.4. Standard Similarity Measures

The subjectivity inherent to the idea of similarity is reflected in the varying types of similarity measures which can be defined. Let $\bar{x}$ represent the mean value of $\mathbf{x}$ and $\bar{y}$ that of $\mathbf{y}$, while $\boldsymbol{\mu}$ denotes an average of $\mathbf{x}$ and $\mathbf{y}$: $\boldsymbol{\mu} = \frac{\mathbf{x} + \mathbf{y}}{2}$. Further, let $\mathbf{X}$ and $\mathbf{Y}$ represent the cumulative distributions of $\mathbf{x}$ and $\mathbf{y}$ when they are considered as probability distributions ($\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = 1$), respectively. That is $\mathbf{X} = (X_1, ..., X_n)$ where $X_j = \sum_{i=1}^{j} x_i$ and similarly for $\mathbf{Y}$ and $\mathbf{y}$. We use $\mathbf{z} = (z_1, z_2, \cdots, z_n)$ and $\mathbf{z}^{(l)}$ to denote the l-times iteratively Gaussian-smoothened, then 2-downsampled vector representation of $|\mathbf{X} - \mathbf{Y}|$. The following list of various similarity or dissimilarity measures were considered in our study.

$$L^2(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^t (\mathbf{x} - \mathbf{y})} \qquad (3)$$

$$L^1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \qquad (4)$$

$$L^\infty(\mathbf{x}, \mathbf{y}) = \max_{i=1}^{n} |x_i - y_i| \qquad (5)$$

$$CO(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| \, ||\mathbf{y}||} \qquad (6)$$

$$CC(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \qquad (7)$$

$$CS(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\mu_i} \qquad (8)$$

$$KL(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i \log \frac{x_i}{y_i} \qquad (9)$$

$$JF(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i \log \frac{x_i}{\mu_i} + y_i \log \frac{y_i}{\mu_i} \qquad (10)$$

$$KS(\mathbf{x}, \mathbf{y}) = \max_{i=1}^{n} |X_i - Y_i| \qquad (11)$$

$$CvM(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (X_i - Y_i)^2 \qquad (12)$$

$$EMD\text{-}L^1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |X_i - Y_i| \qquad (13)$$

$$DD(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\log_2 n} \sum_{j=1}^{n/2^j} \mathbf{z}_i^{(j)} \qquad (14)$$

where L^2: Euclidean distance, L^1: cityblock distance, L^∞: infinity distance, CO: cosine similarity, CC: Pearson correlation coefficient, CS: Chi-square dissimilarity [9], KL: Kullback-Liebler divergence [9], JF: Jeffrey divergence [9], KS: Kolmogorov-Smirnov divergence [9], CvM: Cramer-von Mises divergence [9], EMD-L^1: earth movers distance with L^1 ground distance [18] (EMD in 1D feature space is equivalent to the Mallows Distance [19]), DD: diffusion distance [20].

3. EXPERIMENTS

3.1. Data

We use datasets made available by ImageCLEF [21]. ImageCLEF has offered standardized benchmark tests for a variety of language-neutral CBIR tasks since 2003. The data used in our experiments are from the medical image retrieval task of the ImageCLEF competition administered in 2011 [22]. Three types of datasets were available for this study: *training*, *query*, and *relevance judgment*.

Training data consist of $230,088$ images taken from Pubmed Central database (`www.ncbi.nlm.nih.gov/pmc/`) that contains more than 1 million images taken from published medical journals' figures. Images are therefore of a diverse set of types including those that have little relevance to our retrieval task, as shown in Figure 1. Query data consist of 30 distinct queries, each of which consists of 1-3 query images. These query images are of standard medical image types of different modalities and field of views. Some examples are shown in Figure 2.

Relevance judgment data provides our ground-truth information used in both metric learning and performance evaluation. For each query, a subset *pool* of the entire data set was first collected from the top N matches of an existing M-CBIR system by ImageCLEF organizers. These pooled images are then manually judged by physicians and medical students at Oregon Health and Science University using a web-based GUI tool to be either *relevant* or *irrelevant*. All images not in the pool are judged to be irrelevant. For all queries, these relevance scores are computed for all training images, without annotating the images, and stored in a file.

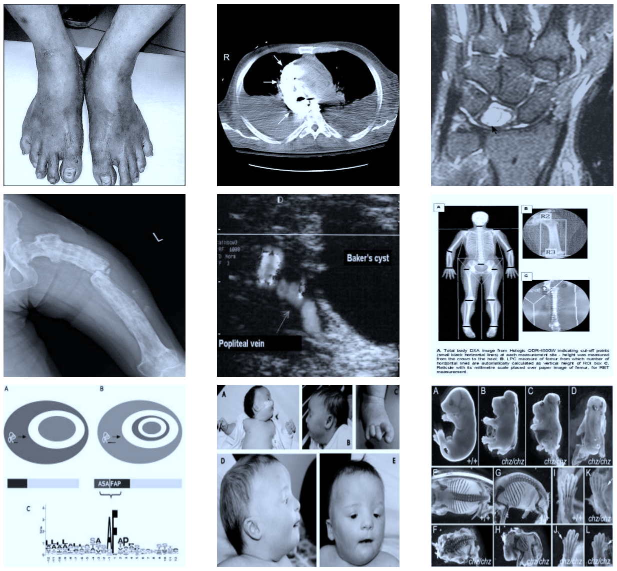

Fig. 1. Various training images for ImageCLEF-2011 displaying their diversity. Modalities of these images are, respectively: optical, CT, MRI, X-ray, ultrasound, DXA, graphical, optical and mixed.

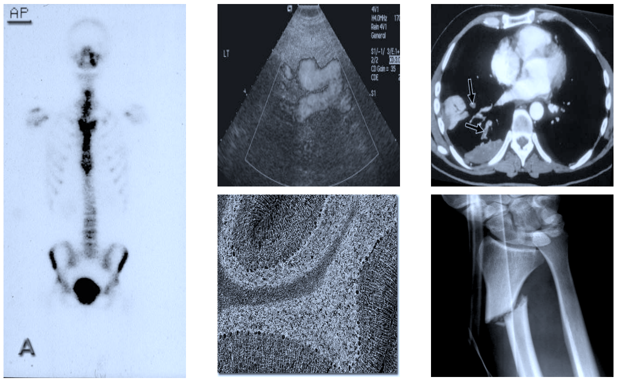

Fig. 2. Examples of query images for ImageCLEF-2011.

3.2. Performance Evaluation Measure

Mean average precision (MAP) is used as a measure to quantify performance of our M-CBIR systems. MAP is a popular performance measure in the information retrieval field, and is defined as the average of per-query precisions,

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} p_{\mu}(q) \qquad (15)$$

where Q is the number of queries and

$$p_{\mu}(q) = \frac{1}{R_q}\sum_{k=1}^{n} P(k)\cdot rel(k) \qquad (16)$$

where $P(k)$ is the precision at the k-th image, R_q is the number of retrieved images which are relevant to the q-th query, $rel(k)$ is a binary indicator function for relevance or lack thereof, and n is the total number of images retrieved for the q-th query.

3.3. Results

We evaluate the MAP score of our ITML systems in comparison with the twelve standard similarity measures applied on the four types of post feature transformations. We set the codebook size K to be 1000, following the previous M-CBIR report using similar feature design [23]. We did not observe benefits in learning a fully parameterized A in our pilot study and so, for computational simplicity, we utilize a metric learning formulation with a diagonal A.

Table 1 summarizes the resulting MAP scores. The highest scores of 0.0227 were achieved by the proposed ITML transformation with the L^1 and diffusion distances. The next highest scores of 0.0214 were achieved with L^1 and diffusion distances without using any post feature transformation. Among the cases with post feature transformations, TF-IDF(PCA) performed best at 0.0209 with the correlation coefficients used as similarity measure.

4. DISCUSSION

This paper proposed a metric learning-based medical CBIR method and presented a systematic experimental comparison of various similarity measures by using a large public database. Our experimental results demonstrated an advantage of the proposed ITML approach which outperformed other CBIR metrics we tested.

In ImageCLEF's medical image retrieval task in 2011, the best MAP score achieved by using only visual information was 0.0338 [22]. Our ITML-based score would have been at the 10th

Measure	**Data Transformation**								
	None	PCA_{75}	PCA_{200}	PCA_{500}	PCA	TF-IDF(PCA)	TF-IDF	PCA(TF-IDF)	**ITML**
L^2	0.0169	0.0207	0.0168	0.0194	0.0203	**0.0208**	0.0157	0.0172	0.0126
L^1	**0.0214**	0.0183	0.0091	0.0196	0.0182	0.0180	0.0207	0.0180	**0.0227**
L^∞	0.0029	0.0032	0.0011	0.0012	0.0029	0.0016	0.0034	0.0097	0.0023
CO	0.0169	0.0207	0.0168	0.0194	0.0203	**0.0208**	0.0157	0.0173	0.0126
CC	0.0184	0.0207	0.0168	0.0194	0.0203	**0.0209**	0.0201	0.0172	0.0185
CS	0.0133	0	0	0	0	0	0.0163	0	0
KL	0.0004	0	0	0	0	0	0.0004	0	0
JF	0	0	0	0	0	0	0.0008	0	0
KS	0.0010	0.0176	0.0003	0.0020	0.0107	0.0176	0.0008	0.0008	0.0005
CvM	0.0011	0.0047	0.0017	0.0014	0.0091	0.0104	0.0009	0.0008	0.0006
EMD-L^1	0.0011	0.0031	0.0016	0.0014	0.0089	0.0098	0.0009	0.0006	0.0006
DD	**0.0214**	0.0183	0.0091	0.0196	0.0140	0.0137	0.0207	0.0177	**0.0227**

Table 1. Result of similarity measure comparison using the MAP score with ImageCLEF 2011 data. None: without feature transformation. PCA_M: codebook constructed using the first M principal components. PCA: all principal components. TF-IDF(PCA) is the TF-IDF transformation of the PCA transformed data. PCA(TF-IDF) is the PCA transformation of the TF-IDF transformed data. ITML is $A^{1/2}S$ where A is a covariance of Mahalanobis metric learned for S. Bold-typed numbers indicate the best performing combinations.

place among 26 submissions in the competition. These scores were relatively much lower than retrieval performance with text annotative information exploited, indicating difficulty of the visual M-CBIR task we tackled.

Note also that the best performing standard metrics in our experiment were both based on the L^1 metric since the diffusion distance is also based on L^1. This seems sensible because L^1 distance tends to outperform L^2 distance in a high-dimensional space. Since overfitting in our metric learning was a concern, we chose to alleviate this by using a simpler form of the metric, forcing A to be diagonal. A small difference in the MAP score between the ITML results and the others supports this choice.

As a future work, we plan to improve the overall retrieval performance by improving our feature design and to compare different metric learning algorithms to better understand the role of metric design in this M-CBIR application.

5. REFERENCES

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 22, pp. 1349–1380, 2000.

[2] T. Deserno, S. Antani, and R. Long, “Ontology of Gaps in Content-Based Image Retrieval,” *Journal of Digital Imaging*, vol. 22, pp. 202–215, 2009.

[3] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, “A review of content-based image retrieval systems in medical applications—clinical benefits and future directions,” *Intl. J. Medical Informatics*, vol. 73, pp. 1–23, 2004.

[4] D. G. Lowe, “Distinctive Image Features from Scale-invariant Keypoints,” *Int. J. Computer Vision*, vol. 60, pp. 91–110, 2004.

[5] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, “Speeded-up robust features (SURF),” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[6] T. S. Lee, “Image representation using 2D Gabor wavelets,” *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 18, pp. 959–971, 1996.

[7] O. Pele and M. Werman, “The Quadratic-Chi Histogram Distance Family,” in *Proc. European Conf. Computer Vision*, 2010, vol. 2, pp. 749–762.

[8] Y. Rubner, C. Tomasi, and L. J. Guibas, “A metric for distributions with applications to image databases,” in *Proc. Int. Conf. Computer Vision*, 1998, pp. 59–66.

[9] J. Puzicha, J. M. Buhmann, Y. Rubner, and C. Tomasi, “Empirical evaluation of dissimilarity measures for color and texture,” in *Proc. Int. Conf. Computer Vision*, 1999, vol. 2, pp. 1165–1172.

[10] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in *Proc. Int. Conf. Machine learning*, 2007, pp. 209–216.

[11] R. T. Dattola, “FIRST: Flexible Information Retrieval System for Text,” *J. Am. Soc. Info. Sci.*, vol. 30, pp. 9–14, 1979.

[12] R. O. Duda, D. G. Stork, and P. E. Hart, *Pattern classification*, Wiley, 2 edition, 2000.

[13] G. Salton, E. A. Fox, and H. Wu, “Extended Boolean Information Retrieval,” *Comm. of the ACM*, vol. 26, pp. 1022–1036, 1983.

[14] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, “Distance metric learning with application to clustering with side-information,” in *Proc. Advances in Neural Information Processing Systems*, 2003, vol. 15, pp. 505–512.

[15] L. Yang and R. Jin, “Distance Metric Learning: A Comprehensive Survey,” Tech. Rep., Department of Computer Science and Engineering, Michigan State University, 2006.

[16] B. Kulis, M. Sustik, and I. S. Dhillon, “Learning Low-rank Kernel Matrices,” in *Int. Conf. on Machine Learning*, 2006, pp. 505–512.

[17] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press, 1997.

[18] H. Ling and K. Okada, “An Efficient Earth Mover’s Distance Algorithm for Robust Histogram Comparison,” *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 29, pp. 840–853.

[19] E. Levina and P. Bickel, “The Earth Mover’s distance is the Mallows distance: some insights from statistics,” in *Proc. Int. Conf. Computer Vision*, 2001, vol. 2, pp. 251–256.

[20] H. Ling and K. Okada, “Diffusion Distance for Histogram Comparison,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, vol. 1, pp. 246–253.

[21] H. Müller, P. Clough, T. Deselaeres, and B. Caputo, Eds., *ImageCLEF: Experimental Evaluation in Visual Information Retrieval (The Information Retrieval Series)*, vol. 32, Springer, 2010.

[22] J. Kalpathy-Cramer, H. Müller, S. Bedrick, I. Eggel, A. G. S. de Herrera, and T. Tsikrika, “Overview of the CLEF 2011 Medical Image Classification and Retrieval Tasks,” in *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[23] U. Avni, J. Goldberger, and H. Greenspan, “Medical image classification at Tel Aviv and Bar Ilan Universities,” in *ImageCLEF*, vol. 32, pp. 435–451. 2010.