A Comparative Study of Similarity Measures for Content-Based Medical Image Retrieval

John Collins, Kazunori Okada

San Francisco State University,
1600 Holloway Avenue, San Francisco, CA 94132, USA
johncoll@mail.sfsu.edu
kazokada@sfsu.edu

Abstract. This note summarizes methodologies employed in our submissions for the medical retrieval subtask of 2012 ImageCLEF competition. Our work aims to provide a systematic comparison of various similarity measures in the Medical CBIR application context. Our system consists of the standard bag-of-words features with SIFT. Computed features are then compared by using various plug-in similarity measures, including diffusion distance and information-theoretic metric learning. This note provides the results of our experimental validation using the 2011 ImageCLEF dataset.

Keywords: ImageCLEF, CBIR, M-CBIR, Content-Based, Image Retrieval, Medical

1 Introduction

ImageCLEF[1–3] is a public standardized competition which focuses attention on, among other things, Medical CBIR (hereafter M-CBIR): CBIR[4–9] in which all images are taken from figures in medical publications. This note focuses on a subtask of M-CBIR 2012, the medical image retrieval task with image data alone without other text-based data. Previous work on M-CBIR has led to the development of an array of specific/general and local/global features. For examples, see SIFT [10, 11], SURF [12, 13] and Gabor Wavelets [14]. Despite the relative maturity of feature design studies, similarity measures in CBIR have not been investigated thoroughly. Previous studies in this regard [15–17] are still few and the lack is especially evident in the M-CBIR subfield.

Addressing this shortcoming, this paper presents a comparative study of M-CBIR with a comprehensive list of similarity measures of many types. Our study shows that well known measures tend to outperform more complex measures with the notable exception of the Diffusion Distance [18]. Further, we show that learning a metric from a set of training data is worthwhile, our best result coming from a combination of a metric learning transformation combined with the Diffusion Distance.

This paper is organized as follows. Sections 2 and 3 will outline, respectively, our methods of feature extraction and representation, and our comparative study of similarity measures. Sections 4 and 5 will summarize our results and their interpretation.

2 Feature Extraction and Image Representation

In this section we describe the process and the individual steps involved in transforming an image to a feature vector, which consists of the following three steps. First, we identify and extract SIFT features from all of the dataset images. Second, we create a codebook of K representative features using K-means clustering. Third, we generate a single vector per image as a normalized histogram of such representative features. Beyond this basic three-step procedure we experiment with a number of standard transformations on the feature codebooks for better retrieval performance.

2.1 Image Representation

From each image, we extract a variable number of features which we classify into K types using the codebook resulting from the bag-of-words model described below. An image is then represented by the frequency distribution of feature types in the image and is, by construction, a vector of length K. Before calculating similarities, each vector is normalized so that it is a probability distribution.

2.2 SIFT: Scale Invariant Feature Transform

SIFT [10, 11] is a proprietary algorithm that describes regions of interest within an image as a feature which is both scale and rotation invariant. The positions of these features, called keypoints, are determined by finding extrema of difference-of-Gaussian images which are robust across multiple scales. Such regions are then turned into 128-element SIFT feature vectors using local directional gradients around the keypoint. We include the 4 extra parameters consisting of the 2 spatial coordinates of the keypoint's position within the image, the scale parameter and the dominant-orientation parameter for a total of 132 dimensions.

2.3 Bag-of-Words

In order to generate an fixed-length vector for each image, we cluster all features together in space using K-means clustering with a predefined vector-length K. Before clustering, each SIFT feature-vector is centered and scaled using Z-Score

normalization. In our case we chose K to be 1000 where this number was taken from an earlier report in the same competition [19]. Each SIFT feature can then be matched with one of the 1000 labels, 0-999, corresponding to the cluster centers. We refer to this set of centers and the corresponding labels as a *codebook*. This bag-of-words method yields the frequency distribution of these labels, 0-999, which describes an image. The notion of a bag-of-words comes from textual data mining and was originally proposed as a way of representing a text document by it's word frequency distribution, ignoring order. In the analogy here, SIFT vectors are word instances and the K centers returned from K-means clustering are the true words. Instead of instances being exact copies of that word as in the text mining case, in the image context a word instance is ascribed to represent the center to which it is closest in distance.

2.4 Data Transformations

The following standard transformations were examined with the goal of improved performance.

PCA: Principal Components Analysis PCA[20] is a technique used mainly for dimension reduction. For a space X, It seeks to find the linear combination $Y = \sum_{i=1}^{n} \lambda_i x^{(i)}$ for column vectors $x^{(i)}$ of X such that the dimensions of Y are not correlated (linearly independent). Moreover, dimensions in Y are ordered from most to least important, where importance is defined in terms of variance. In practice, the transformed data in Y is often used for dimension reduction since one gets a variance-maximal m-dimensional representation of X by taking the first m dimensions of Y. How small to make m is data dependent and is typically chosen to cover at least 95% or 99% of the data's variance.

We experimented by varying the number of dimensions in PCA with both 2011 and 2012 ImageCLEF competition datasets and the results are shown in Figure. 1. We found the variance spread of these two datasets to be quite large. Overall, using our image representation, the 2012 codebook captured more variance in fewer components than did the 2011 codebook. However, in both cases we found that it took most of the components to cover an adequate amount of variance.

Tf-Idf: Term Frequency - Inverse Document Frequency This idea, like bag-of-words, comes from textual data mining. The goal is to penalize a vector for words (features) whenever they are common across the entire data set. Term Frequency (Tf) for an observation x is just the value at term i's position, i.e. x_i. Inverse Document Frequency (Idf) is calculated by $Idf_t = \log \frac{|D|}{|\{d \in D : t \in d\}|}$ where D is the dataset of observations and $\{d \in D : t \in d\}$ is the number of observations which are non-zero in the i-th position. For Tf-Idf, we transform $d \in D$ by $d \cdot Idf$. In our case, we do not explicitly measure the presence or non-presence of a feature but rather the

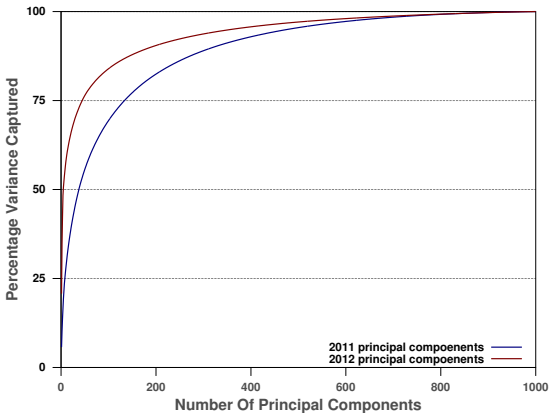

Fig. 1: Variance captured by principal components

count of each feature. Thus, Tf-Idf provides for us a weighting of our images which penalizes features if they are very common in the data set and awards features otherwise.

In the course of our study we experimented not just with PCA and Tf-Idf, but also with nestings of these operations. In short, for our dataset, X, we compute the following data transformations.

1. PCA(X)
2. Tf-Idf(X)
3. PCA(Tf-Idf(X))
4. Tf-Idf(PCA(X))

3 Database Ranking by Similarity Comparison

Given a query image, the goal here is to calculate the similarities or distances between it and each of the images in the database. Then the first image returned will be the most similar, the second return will be the second most similar, and so on. In some cases a query may consist of multiple images. In this case, we calculate the average similarity of the query parts to each database image as the representative score. The subjectivity inherent to the idea of similarity is reflected in the varying types of similarity measures which can be defined. In some cases below, e.g. cosine similarity, a measure has its natural expression as a similarity rather than a dissimilarity measure. However, in most cases the natural definition is as a dissimilarity measure. We shall use d when referring to a dissimilarity measure and s when referring to a similarity measure. The idea of calculating similarity as an additive inverse of distance comes from the idea of

a metric. A metric on a set X is a mapping $d : X \times X \to \mathbb{R}$ such that $\forall x, y \in X$, the following conditions all hold: $d(x, y) \geq 0$, $d(x, y) = 0$ if and only if $x = y$, $d(x, y) = d(y, x)$, and $d(x, z) \leq d(x, y) + d(y, z)$.

We use the broader term *measure* because in some cases what we use will fail in one or more of the conditions above. For example, the *Kullback-Liebler Divergence* is not symmetric since, in general, $d(x, y) \neq d(y, x)$. Finally, when a dissimilarity measure is being considered, it should be understood that we are using $1 - d(x, y)$ to calculate the similarity where x and y are appropriately scaled so that $d(x, y) \in [0, 1]$.

3.1 Various Similarity Measures

The following lists similarity or dissimilarity measures we considered in our study. Let $\mathbf{x}$ denote the vector $(x_1, x_2, ..., x_n)$ representing the query image and $\mathbf{y}$ the vector $(y_1, y_2, ..., y_n)$ representing another image. Further, let $\bar{x}$ represents the mean of the values in the $\mathbf{x}$ vector and $\bar{y}$ the mean of $\mathbf{y}$. Further, let $\mathbf{X}$ and $\mathbf{Y}$ represent, respectively, the cumulative distributions of $\mathbf{x}$ and $\mathbf{y}$ when they are considered as probability distributions ($\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = 1$). That is $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ where $X_j = \sum_{i=1}^{j} x_i$ and similarly for $\mathbf{Y}$ and $\mathbf{y}$. Finally $\boldsymbol{\mu} = (\mu_1, .., \mu_n)$ is the mean vector such that $\boldsymbol{\mu} = \frac{\mathbf{x}+\mathbf{y}}{2}$.

- Minkowski and Standard Measures

  Euclidean Distance (L_2) $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

  Cityblock Distance (L_1) $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$

  Infinity Distance (L_∞) $d(\mathbf{x}, \mathbf{y}) = max_{i=1}^{n} |x_i - y_i|$

  Cosine Similarity (**CO**) $s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$

- Statistical Measures

  Pearson Correlation Coefficient (**CC**) $d(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$

  Chi-Square Dissimilarity (**CS**) $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\mu_i}$ [21]

- Divergence Measures

  Kullback-Liebler Divergence (**KL**) $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i \log \frac{x_i}{y_i}$ [22]

  Jeffrey Divergence (**JF**) $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i \log \frac{x_i}{\mu_i} + y_i \log \frac{y_i}{\mu_i}$ [21]

  Kolmogorov-Smirnov Divergence (**KS**) $d(\mathbf{x}, \mathbf{y}) = max_{i=1}^{n} |X_i - Y_i|$ [23]

  Cramer-von Mises Divergence (**CvM**) $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (X_i - Y_i)^2$ [21]

- Other Measures

Earth Mover's Distance (**EMD-L_1**) $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |X_i - Y_i|$ [24][1]

Diffusion Distance (**DD**) $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\log_2 n} \sum_{j=1}^{n/2^j} \mathbf{z}_i^{(j)}$ where $\mathbf{z} = (z_1, z_2, \cdots, z_n)$ and $\mathbf{z}^{(l)}$ is the l-times iteratively Gaussian-smoothened, then 2-downsampled vector representation of $|\mathbf{X} - \mathbf{Y}|$ [18].

3.2 Metric Learning

Metric Learning [26] is the process of using information about the similarity and/or dissimilarity of some dataset X, to learn a mapping to a new space $Y = A^{1/2}X$, in which similar data will be closer together and dissimilar data will be farther apart. Let $\boldsymbol{\lambda}$ denote an n-dimensional vector in which λ_i determines the weight given to the i-th variable. With such a $\boldsymbol{\lambda}$ we can define a weighted L_2 metric on X such that for each $\mathbf{x}$ and $\mathbf{y}$ in X we capture the distance between them by $d_\lambda(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{N} \lambda_i (x_i - y_i)^2}$. The idea of metric learning is to learn the appropriate weights $\boldsymbol{\lambda}$ from a training dataset. A less strict formulation of metric learning allows the weights to be described by a non-diagonal symmetric positive semi-definite matrix A such that $\boldsymbol{\lambda} = diag(A)$, leading to a more general Mahalanobis-type metric formulation:

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

Many algorithms [26–28] have been used to learn such a metric with Yang [29] giving a nice summary. We employ an algorithm called Information-Theoretic Metric Learning (hereafter ITML) which is widely used. ITML uses an information-theoretic cost model which iteratively enforces similarity/dissimilarity constraints with the input being a list of such pairwise constraints and the output being a learned matrix A. An equivalent and more computationally efficient formulation to the one above is to use the L_2 metric on the data after applying the data transformation $X \mapsto A^{1/2}X$. In this study, we employ the diagonal form of A for simplicity and information about similarity/dissimilarity attained from the 2011 ImageCLEF dataset as our training data.

3.3 Query Filtering

We used the Modality Classification results made available by ImageCLEF to filter out certain image types which are likely to be irrelevant to all queries. Table 1 indicates the filtering performed. In short, we included all and only diagnostic images.

[1] EMD for 1D features is equivalent to the Mallows Distance [25]

Table 1: Filtering of Modality Types

Included Modalities	**Modalities Filtered Out**
Ultrasound	Compound or multipane images
Magnetic Resonance	Tables and Forms
Computerized Tomography	Program Listing
X-Ray, 2D Radiography	Statistical figures, graphs, charts
Angiography	Screenshots
PET	Flowcharts
Combined Modalities in one image	System overviews
Dermatology, skin	Gene sequence
Endoscopy	Chromatography
Other organs	Chemical structure
Electroencephalography	Mathematics, formulae
Electrocardiography	Non-clinical photos
Electromyography	Hand-drawn sketches
Light microscopy	
Electron microscopy	
Transmission microscopy	
Fluorescence microscopy	
3D reconstructions	

4 Experimental Results

Using the relevance judgments from 2011 ImageCLEF, we validate our proposed system. Table 2 shows the Mean Average Precision (hereafter MAP) scores for various permutation of our system components computed using the relevance judgment file from the 2011 results.

We used this table to select our best potential measure/transformation combinations for 2012 ImageCLEF competition. In the end we submitted the following seven runs to the 2012 ImageCLEF medical retrieval competition.

1. L_1 on the untransformed data (reg_cityblock)
2. DD on the untransformed data (reg_diffusion)
3. L_2 on the Tf-Idf(PCA) transformed data (tfidf_of_pca_euclidean)
4. CO on the Tf-Idf(PCA) transformed data (tfidf_of_pca_cosine)
5. PC on the Tf-Idf(PCA) transformed data (tfidf_of_pca_correlation)
6. L_1 on the ITML data (itml_cityblock)
7. DD on the ITML data (itml_diffusion)

These selected runs are identified in Table 2 as highlighted items. Submissions to ImageCLEF medical retrieval[30, 31] are text files containing a ranked list of at most 1000 images for each of the competition queries, along with information

Table 2: Result of similarity measure comparison using the MAP score with 2011 ImageCLEF data. PC_M: codebook constructed using the first M principal components.. PC: all principal components. Tf-Idf(PC) is the Tf-Idf transformation of the PC transformed data. Tf-Idf is the data under the Tf-Idf transformation and PC(Tf-Idf) is the PC transformation of the Tf-Idf transformed data. ITML is $A^{1/2}X$ where A is a metric learned from similarity/dissimilarity information about X.

Measure	**Data Transformation**								
	None	PC_{75}	PC_{200}	PC_{500}	PC	Tf-Idf(PC)	Tf-Idf	PC(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

such as the rank, query number and score. These submission files are constructed in the TREC-style submission format [32].

5 Discussion

We have presented a systematic comparison of various plug-in (dis-)similarity measures for M-CBIR with a standard bag-of-words feature method. Our validation results with the last year 2011 dataset indicates both ITML and diffusion distance to be promising choices for the ad-hoc image-based retrieval task for medical images. Based on this result, we have entered seven runs (combinations of three top performing measures with different feature transformations). The results were disappointing. All the runs were placed at the last of this category with very low MAP scores for this year competition. The reasons for this performance may include a potentially suboptimal choice of our feature extraction/representation and query filtering employed. Investigation of this and a rerun of our study with a better base-CBIR system is our important future work. Among our 2012 results, we observe the consistent trend of the diffusion and cityblock distances to perform best among other submitted runs. This indicates the virtue of distance measures based on L_1 metric. The run with metric learning (ITML) was placed the last in our list. This may indicate significant change of data characteristics between the 2011 and 2012 data, which would naturally cause this reduced performance. Investigating the true advantage of the metric learning approach in M-CBIR remains another future work.

References

[1] H. Muller, P. Clough, T. Deselaeres, and B. Caputo, eds., *ImageCLEF: Experimental Evaluation in Visual Information Retrieval (The Information Retrieval Series)*. Springer, 1st edition. ed., Aug. 2010.

[2] P. Clough, H. Müller, and M. Sanderson, "Seven Years of Image Retrieval Evaluation," in *ImageCLEF* (H. Müller, P. Clough, T. Deselaers, B. Caputo, and W. B. Croft, eds.), vol. 32 of *The Information Retrieval Series*, Springer Berlin Heidelberg, 2010.

[3] H. Müller and J. Kalpathy–Cramer, "The Medical Image Retrieval Task," in *ImageCLEF* (H. Müller, P. Clough, T. Deselaers, B. Caputo, and W. B. Croft, eds.), vol. 32 of *The Information Retrieval Series*, Springer Berlin Heidelberg, 2010.

[4] 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, no. 12, 2000.

[5] 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, no. 1, 2004.

[6] M. Rahman, T. Wang, and B. C. Desai, "Medical image retrieval and registration: towards computer assisted diagnostic approach," in *Proc. IDEAS Workshop on Medical Information Systems: The Digital Hospital*, 2004.

[7] T. Deserno, S. Antani, and R. Long, "Ontology of Gaps in Content-Based Image Retrieval," *Journal of Digital Imaging*, vol. 22, 2009.

[8] T. M. Lehmann, B. B. Wein, J. Dahmen, J. Bredno, F. Vogelsang, and M. Kohnen, "Content-based image retrieval in medical applications: a novel multistep approach," in *SPIE* (M. M. Yeung, B.-L. Yeo, and C. A. Bouman, eds.), vol. 3972, 1999.

[9] A. Marchiori, C. Brodley, J. Dy, C. Pavlopoulou, A. Kak, L. Broderick, and A. M. Aisen, "CBIR for medical images - an evaluation trial," in *IEEE Workshop on Content-based Access of Image and Video Libraries*, 2001.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Computer Vision*, vol. 2, 1999.

[11] D. G. Lowe, "Distinctive Image Features from Scale-invariant Keypoints," *Int. J. Computer Vision*, vol. 60, 2004.

[12] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Proc. European Conf. Computer Vision* (A. Leonardis, H. Bischof,

and A. Pinz, eds.), vol. 3951 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2006.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, 2008.

[14] T. S. Lee, "Image representation using 2D gabor wavelets," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 18, 1996.

[15] O. Pele and M. Werman, "The Quadratic-Chi Histogram Distance Family," in *Proc. European Conf. Computer Vision* (K. Daniilidis, P. Maragos, and N. Paragios, eds.), vol. 6312 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2010.

[16] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proc. IEEE Int. Conf. Computer Vision*, 1998.

[17] J. Puzicha, J. M. Buhmann, Y. Rubner, and C. Tomasi, "Empirical evaluation of dissimilarity measures for color and texture," in *Proc. IEEE Int. Conf. Computer Vision*, vol. 2, 1999.

[18] H. Ling and K. Okada, "Diffusion Distance for Histogram Comparison," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, 2006.

[19] U. Avni, J. Goldberger, and H. Greenspan, "Medical image classification at Tel Aviv and Bar Ilan Universities," in *ImageCLEF* (H. Müller, P. Clough, T. Deselaers, B. Caputo, and W. B. Croft, eds.), vol. 32 of *The Information Retrieval Series*, Springer Berlin Heidelberg, 2010.

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

[21] J. Puzicha, T. Hofmann, and J. M. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997.

[22] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, 1996.

[23] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary detection by constrained optimization," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 12, no. 7, 1990.

[24] 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, no. 5, 2007.

[25] E. Levina and P. Bickel, "The Earth Mover's distance is the Mallows distance: some insights from statistics," in *Proc. IEEE Int. Conf. Computer Vision*, vol. 2, 2001.

[26] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," *Learning*, vol. 15, no. 15, 2003.

[27] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. Intl. Conf. Machine learning*, (New York, NY, USA), ACM, 2007.

[28] B. McFee and G. Lanckriet, "Metric Learning to Rank," in *Proc. Intl. Conf. Machine learning*, 2010.

[29] L. Yang and R. Jin, "Distance Metric Learning: A Comprehensive Survey," tech. rep., Department of Computer Science and Engineering, Michigan State University, 2006.

[30] H. Müller, A. G. S. de Herrera, J. Kalpathy-Cramer, D. D. Fushman, S. Antani, and I. Eggel, "Overview of the ImageCLEF 2012 medical image retrieval and classification tasks," *CLEF 2012 working notes*, Sept. 2012.

[31] J. Kalpathy–Cramer, S. Bedrick, and W. Hersh, "Relevance Judgments for Image Retrieval Evaluation," in *ImageCLEF* (H. Müller, P. Clough, T. Deselaers, B. Caputo, and W. B. Croft, eds.), vol. 32 of *The Information Retrieval Series*, Springer Berlin Heidelberg, 2010.

[32] N. Stokes, "TREC: Experiment and Evaluation in Information Retrieval," *Computational Linguistics*, vol. 32, pp. 563–567, Nov. 2006.