J Digit Imaging (2013) 26:850–865
DOI 10.1007/s10278-013-9591-x

A Novel Similarity Learning Method via Relative Comparison for Content-Based Medical Image Retrieval

Wei Huang · Peng Zhang · Min Wan

Published online: 6 April 2013

Abstract Nowadays, the huge volume of medical images represents an enormous challenge towards health-care organizations, as it is often hard for clinicians and researchers to manage, access, and share the image database easily. Content-based medical image retrieval (CBMIR) techniques are employed to facilitate the above process. It is known that a few concrete factors, including visual attributes extracted from images, measures encoding the similarity between images, user interaction, etc. play important roles in determining the retrieval performance. This paper concentrates on the similarity learning problem of CBMIR. A novel similarity learning paradigm is proposed via relative comparison, and a large database composed of 5,000 images is utilized to evaluate the retrieval performance. Extensive experimental results and comprehensive statistical analysis demonstrate the superiority of adopting the newly introduced learning paradigm, compared with several conventional supervised and semi-supervised similarity learning methods, in the presented CBMIR application.

Keywords Content-based medical image retrieval · Similarity learning · Relative comparison

W. Huang (✉)
School of Information Engineering, Nanchang University, Nanchang, China
e-mail: n060101@e.ntu.edu.sg

P. Zhang (✉)
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
e-mail: zh0036ng@e.ntu.edu.sg

M. Wan
National Heart Centre, SingHealth, Singapore, Singapore
e-mail: wanmin1983@gmail.com

Introduction

Medical images are often of great importance in diverse clinical operations, including diagnosis, treatment planning, physicians education, etc. According to Frost and Sullivan's report, the number of medical images increases by 20–40 % each year worldwide, and it is estimated that there are no less than one billion medical images produced and stored in the USA alone by the year 2012 [1]. Although millions of medical images are available, these images represent an enormous challenge towards health-care organizations (such as hospitals, research institutes, medical colleges, etc.), as they struggle to manage, access, and share the extremely large scale of data in their daily affairs. In order to facilitate the above process, medical image retrieval techniques have been proposed and adopted in various applications for the past 20 years [2–19].

Based on different attributes of medical images employed in the retrieval procedure, most conventional medical image retrieval techniques can be categorized into two kinds: text-based medical image retrieval (TBMIR) and content-based medical image retrieval (CBMIR). For TBMIR, it aims to search similar or clinically relevant reference images from a large database based on annotated textual information of each image. For instance, in digital imaging and communication in medicine (DICOM), which is a popular standard for handling, storing, and transmitting information regarding medical images, such textual information (about patient data, imaging sequences, etc.) often exists in headers of image files. It is generally acknowledged that the retrieval performance of TBMIR systems highly depends on these textual information, in which high preciseness and accuracy is necessary. Although TBMIR is popular in several practical utilizations (e.g., Goldminer from ARRS [2], BioText from Berkeley [3], iMedline from NIH [4],

etc.), it has its own drawbacks. In the review paper of Müller et al. [5], it is mentioned that DICOM headers have proven to contain a fairly high error rate, which can be as high as 16 %. The performance of TBMIR based on such inaccurate textural information will certainly be badly influenced. Therefore, in order to improve the performance of medical image retrieval systems, much research effect nowadays has been put into CBMIR, which is often considered to be a good complement to the traditional TBMIR [5–19].

CBMIR incorporates visual attributes extracted from medical images, rather than textual information solely utilized Din TBMIR, for the retrieval purpose [5–8]. There are already many CBMIR studies proposed for a wide spectrum of clinical applications in recent years. For instance, CBMIR are vastly employed in digital mammography [6–16]. In [9], a computer-aided expert system named MAMMO/ICON is introduced for automated mammographic image retrieval, which is one of the earliest CBMIR studies in mammography. In [10], the authors concentrate on the relevance feedback issue in CBMIR and propose a new strategy to improve the performance of mammographic image retrieval via incremental learning and support vector machine (SVM) regression. In [11], the similarity problem is tackled in CBMIR and a hierarchal learning structure, in which neural networks (NN) and SVM are utilized as classifiers, is introduced. In [12], a variety of visual attributes extracted from breast regions on mammogram (including structural characteristics such as shape, size, etc., as well as texture information such as moment-based feature and Haralick's texture feature, etc.) are incorporated, and self-organizing map neural networks is employed to perform the retrieval operation. In [13], eight information-theoretic similarity measures (including various forms of entropy, mutual information, and divergence) are incorporated into the retrieval process, and the influence of adopting different information-theoretic similarity measures on the retrieval performance is analyzed. In [14], research effort has been put on the user interaction issue in CBMIR, and a k-nearest neighbor-based method is designed to improve the retrieval performance with prior knowledge received from user feedbacks. In [15], a genetic algorithm is introduced for finding the optimized importance weight of each mammography image through a pseudo natural evolution process. Besides the above-mentioned utilizations in mammography, CBMIR is also favored in other imaging modalities and clinical utilizations, such as lung images [16], CT images [17, 18], eye images [19], etc.

Although CBMIR systems are developed in a variety of ways, a few concrete factors that can influence the retrieval performance are common. It is summarized in [5, 6] that visual features extracted from medical images, measures encoding the similarity between the query image and retrieved images, retrieval algorithms, user interaction, etc. are often of great importance in a CBMIR system. Following the above understanding, much research effort has been emphasized on individuals of them. For instance, [12] focuses on different types of feature extracted from images; [11] and [13] tackle the similarity measure problem; while [10] and [14] concentrate more on the user interaction (relevance feedback) problem. In this study, the similarity measure problem in CBMIR is concentrated, and a novel similarity learning method via relative comparison is presented.

In order to automatically "find" a proper similarity measure from existing data, various similarity learning techniques have been proposed, and many of them fall into the category of machine learning studies. In machine learning, "similarity" is closely connected towards "distance," and a similarity learning problem is often addressed as a distance learning problem therein [20]. Depending on the availability of training data, most similarity (distance) learning methods can be mainly classified as supervised learning methods and unsupervised learning methods. For supervised learning, labels of training data are usually available. A parametric form of the similarity is usually predefined, and the unknown parameters within the similarity measure are automatically learned with the aid of available label information [21–25]. For unsupervised learning (a.k.a. manifold learning), such labels are unknown. The main idea is to learn an underlying low-dimensional manifold, where geometric relationship (e.g., distance) between data is preserved [20]. Recently, a new trend known as semi-supervised learning, which is in between supervised and unsupervised learning paradigms, begins to receive much attention [26, 27]. Supervisory information in semi-supervised learning can be constructed via either limited labeled data or specific constraints, in which side-information is popular [28]. Side-information is defined on data pairs, and it often contains a pairwise similarity constraint (e.g., data pair (A, B) are from the same but unknown group) and a pairwise dissimilarity constraint (e.g., data pair (A, B) are from different and unknown groups). Since these constraints can be derived from labeled data in supervised learning but not vice versa, supervisory information in semi-supervised learning is not as strong as that in supervised learning [26, 27].

Relative comparison is another kind of constraints in semi-supervised learning [29]. The main difference between it and side-information is that relative comparison can reflect the relationship between listwise data, instead of pairwise data in side-information, making it more suitable for image-retrieval applications. In this study, the idea of relative comparison is incorporated in similarity learning for CBMIR. The motivation of this study is as follows. For many existing CBMIR studies in which conventional supervised or semi-supervised similarity learning methods are adopted, problems of similarity learning and image retrieval are often tackled as classification tasks, and

established classifiers (e.g., SVM, NN, etc.) are vastly employed. However, in many up-to-date general image retrieval studies, there is a trend to consider the retrieval as a "ranking" process, which is believed to comply with the nature of image retrieval better (i.e., retrieval is in fact a ranking procedure) [30, 33]. Therefore, we follow this new trend and adopt the idea of relative comparison which belongs to the scope of ranking, into CBMIR for the first time in this study. It is also interesting to observe whether taking CBMIR as a ranking procedure can benefit its retrieval performance, compared with other conventional supervised or semi-supervised CBMIR studies.

The organization of the paper is as follows. In Section "Methodology," the technical part of this paper is elaborated. Section "Kendall's Tau Coefficient" introduces Kendall's Tau coefficient (KT) [34], which is a conventional rank correlation measure and the basis of this study. KT can reveal relative comparison between listwise images. However, the original KT definition cannot suit distance-based similarity learning well, as its way to represent relative comparison is discrete, making the whole KT infeasible to be optimized directly for similarity learning. Hence, a new rank correlation measure based on the original definition of KT, named surrogate Kendall's Tau coefficient (SKT), is presented in Section "A New Rank Correlation Measure." A corresponding distance-based similarity learning algorithm is introduced in Section "Distance-based Similarity Learning via Direct Optimization on SKT" for the new measure. In Section "Experiments and Discussion," the newly proposed similarity learning paradigm via relative comparison has been evaluated using a large database composed of 5,000 images for evaluating the retrieval performance. Five popular supervised and semi-supervised similarity learning methods are implemented for retrieval performance comparison. A statistical analysis is conducted based on retrieval results obtained by all methods. Extensive experimental and comprehensive statistical analysis reveal the superiority of incorporating the newly introduced similarity learning paradigm in CBMIR. In Section "Conclusion," the conclusion of this study is drawn.

Methodology

Kendall's Tau Coefficient

Kendall's Tau coefficient (a.k.a. Kendall tau rank correlation coefficient) is a traditional rank correlation measure named after the British statistician Sir Maurice Kendall. Its definition is as follows [34].

$$\mathrm{KT} = \frac{N}{N_n} = \frac{P-Q}{N_n} = \frac{P-Q}{\frac{1}{2}n(n-1)}, \tag{1}$$

where P and Q represent relative comparison; P and Q are numbers of concordant pairs and discordant pairs, respectively; N_n is a normalization term denoted by the number of image pairs in a retrieved list consisting of n images (N_n is equivalent to the number of two combinations from n images: $\mathrm{C}_n^2 = \frac{n!}{(n-2)!\cdot 2!} = \frac{1}{2}n(n-1)$). The idea of concordant/discordant pairs can be illustrated in Fig. 1. Provided two retrieved images (x, y), x is more relevant than y towards a given query, and both x and y are in a retrieved image list (normally in a descending relevance order in practice). If x is retrieved before y in a list of descending relevance order, which matches the relevance relationship between the two images, (x, y) forms a concordant pair. Otherwise, they forms a discordant pair. Generally speaking, the range of KT is within $[-1, +1]$, and higher KT values imply better retrieval performance within obtained retrieved lists. In this study, we intend to conduct similarity learning based on KT following the idea of relative comparison and ranking. However, optimization operations cannot be applied on KT directly for learning, as KT is neither continuous nor differentiable in terms of discrete pair-counting terms P and Q. Hence, a new continuous and differentiable rank correlation measure is necessary.

A New Rank Correlation Measure

We first represent terms P and Q in Eq. 1 mathematically. Given a query image $q = (\Re_q, \ell_q)$ and two of its retrieved images $x = (\Re_x, \ell_x)$ and $y = (\Re_y, \ell_y)$, where $\Re_{(x)}$ and $\ell_{(x)}$ denote extracted low-level features from image x and annotated information by clinicians to image x, respectively. P and Q in Eq. 1 can be represented via the following mathematical way:

$$\text{concordant pair (P):} \quad \mathrm{sgn}(s_{(q,x)} - s_{(q,y)}) \times \mathrm{sgn}(\ell_{(q,y)} - \ell_{(q,x)}) = 1 \tag{2}$$

$$\text{discordant pair (Q):} \quad \mathrm{sgn}(s_{(q,x)} - s_{(q,y)}) \times \mathrm{sgn}(\ell_{(q,y)} - \ell_{(q,x)}) = -1, \tag{3}$$

where $\mathrm{sgn}(\cdot)$ is a signum function, whose outcome is +1 when its variable is nonnegative and −1 otherwise; $s_{(q,x)}$ is a distance-based similarity between q and x, which is of

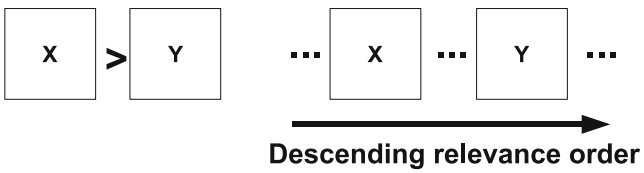


Fig. 1 An illustration of concordant and discordant pairs in KT measure (Eq. 1) for image retrieval

an exponential form based on a weighted L_1 distance in this study: $s_{(q,x)} = \exp(- < a, |\Re_q - \Re_x| >)$, where $<, >$ denotes an inner product between vector a and the absolute feature vector difference between images q and x; $\ell_{(q,x)}$ is the absolute difference of annotated information (in this study, this annotated information to each image is an integer grade) between q and x: $\ell_{(q,x)} = |\ell_q - \ell_x|$. Hence, the vector a performs a scaling on the feature space, and elements within the vector a are parameters to learn in this study (i.e., a multivariable learning problem). Given retrieved image x is more relevant than retrieved image y towards query image q, the similarity between x and q should be larger than that between y and q (i.e., $s_{(q,x)} > s_{(q,y)}$), the difference of annotated information (integer grades difference in this study) between q and x should be smaller than that between q and y (i.e., $\ell_{(q,x)} < \ell_{(q,y)}$). If the above conditions holds, (x, y) is a concordant pair, P increases by 1 as indicated by Eq. 2. Otherwise, (x, y) is a discordant pair, Eq. 3 holds, and Q increases by 1. After substituting terms P and Q into the original KT definition in Eq. 1, it becomes:

$$\mathrm{KT} = \frac{N}{N_n} = \frac{1}{N_n} \sum_{x,y \in D, x \neq y} \Big(\mathrm{sgn}\big(s_{(q,x)} - s_{(q,y)}\big) \times \mathrm{sgn}\big(\ell_{(q,y)} - \ell_{(q,x)}\big) \Big), \quad (4)$$

where, D denotes all images to retrieve regarding query q. We further overcome the step transition characteristics of signum functions in Eq. 4 by approximating them via continuous hyperbolic tangent functions. An illustration of this approximation is shown in Fig. 2. The approximation is depicted as follows:

$$\mathrm{sgn}(\xi) \simeq \tanh(\xi) = \frac{\sinh(\xi)}{\cosh(\xi)} = \frac{\frac{e^{\xi} - e^{-\xi}}{2}}{\frac{e^{\xi} + e^{-\xi}}{2}} = \frac{e^{\xi} - e^{-\xi}}{e^{\xi} + e^{-\xi}} = \frac{e^{2\xi} - 1}{e^{2\xi} + 1} \quad (5)$$

After incorporating Eq. 5 into Eq. 4, a new continuous and differentiable rank correlation measure, named SKT, can be written as:

$$\mathrm{SKT} = \frac{1}{N_n} \cdot \sum_{x,y \in D, x \neq y} \left(\frac{\exp\big(2(s_{(q,x)} - s_{(q,y)})\big) - 1}{\exp\big(2(s_{(q,x)} - s_{(q,y)})\big) + 1} \times \frac{\exp\big(2(\ell_{(q,y)} - \ell_{(q,x)})\big) - 1}{\exp\big(2(\ell_{(q,y)} - \ell_{(q,x)})\big) + 1} \right) \quad (6)$$

Distance-Based Similarity Learning via Direct Optimization on SKT

A distance-based similarity learning algorithm via direct optimization on SKT based on a gradient method is elaborated in Table 1. The most critical part here is to calculate the gradient of SKT with respect to the parameter to learn a $\big($i.e., $\nabla \mathrm{SKT}(a)\big)$ in Steps T4 and T5. The detailed derivation is demonstrated in the Appendix section at the end of this paper. After derivation, the gradient of SKT is as follows:

$$\nabla \mathrm{SKT}(a) = \frac{1}{N_n} \cdot \left(\sum_{x,y \in D, x \neq y} \frac{4 \cdot \left(\frac{\partial s_{(q,x)}}{\partial a} - \frac{\partial s_{(q,y)}}{\partial a} \right) \cdot \frac{\exp\big(2(\ell_{(q,y)} - \ell_{(q,x)})\big) - 1}{\exp\big(2(\ell_{(q,y)} - \ell_{(q,x)})\big) + 1}}{\exp\big(2(s_{(q,x)} - s_{(q,y)})\big) + \exp\big(2(s_{(q,y)} - s_{(q,x)})\big) + 2} \right) \quad (7)$$

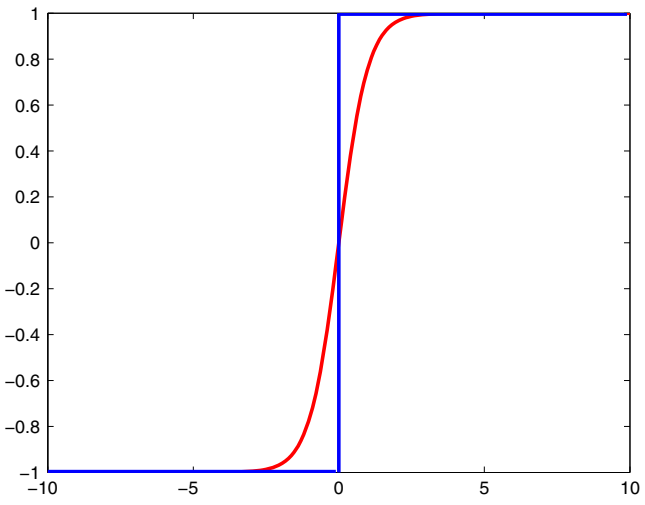


Fig. 2 An illustration of approximating a discrete signum function (*in blue*) via a continuous hyperbolic tangent function (*in red*)

For the utilized similarity (i.e., $s_{(q,x)} = \exp(- < a, |\Re_q - \Re_x| >)$), a is the parameter to learn and there are several elements within it (i.e., the extracted feature vector is 21-dimensional according to Table 2 in this study, hence there are also 21 elements within a to learn). Therefore, learning a in this study is actually solving a multidimensional optimization problem. It is commonly acknowledged that high-dimensional problems can be really tricky, and it is often hard to find the global optimum. In many studies, researchers look for a local minimum that is good enough for the specific application, instead. In Table 1, T iterations are executed in the training step and there are T learned parameters a obtained (Steps T2 to T7 in Table 1). After that, one optimal a with the highest KT value evaluated on the validation data is sorted out (Steps V1 to V5 in Table 1). The similarity measure with this optimal a will be utilized in the testing step for retrieval performance evaluation.

Table 1 A distance-based similarity learning algorithm via direct optimization on SKT

Inputs	1. Images for training: $(q, x \in D)$
	2. Images for validation: $(q_v, x_v \in D_v)$
	3. Iteration times: T
	4. Learning rate: η
Training	
T1.	Initialize parameter a of the similarity $s_{(q,x)}$ as a_0
T2.	For t = 1 to T
T3.	Set $a = a_{t-1}$
T4.	Feed $(q, x \in D)$ to Eq. 7 to calculate $\bigtriangledown \mathrm{SKT}(a)$
T5.	Update a via a *gradient approach* : $a = a + \eta \cdot \bigtriangledown \mathrm{SKT}(a)$
T6.	Set $a_t = a$
T7.	End for T2
Training Results	T learned similarities $s_{(q,x)}$ with their corresponding learned parameters a
Validation	
V1.	For j = 1 to T
V2.	Feed jth learned similarity to $(q_v, x_v \in D_v)$ to retrieve images
V3.	Calculate its corresponding KT value using Eq. 1
V4.	End for V1
V5.	Determine $s^{opt}_{(q,x)}$ as the one with the highest KT value
Outputs	Optimal learned similarity: $s^{opt}_{(q,x)}$

Table 2 Detailed descriptions of the extracted 21-dimensional local feature vector from each slit-lamp image

Dimension	Description
1	Mean intensity inside the detected lens contour
2–4	Color components (HSI) inside the detected lens contour
5	Entropy inside the detected lens contour
6	Neighborhood standard deviation inside the detected lens contour
7	Mean intensity inside the detected nucleus contour
8–10	Color components (HSI) inside the detected nucleus contour
11	Entropy inside the detected nucleus contour
12	Neighborhood standard deviation inside the detected nucleus contour
13	Intensity ratio between nucleus and lens
14	Intensity ratio between sulcus and nucleus
15	Intensity ratio between anterior lentil and posterior lentil
16	Intensity of sulcus
17–18	Strength of nucleus edge
19–21	Color components (HSI) on posterior reflex

Experiments and Discussion

Experimental Data and Feature Extraction

Our experimental evaluation was conducted using a large database composed of 5,000 slit-lamp images obtained from 5,000 different cases with nuclear cataract disease. Informed consent was obtained for all cases. A slit-lamp image focuses on the human lens in the eyeball, and it is the primary means to diagnose nuclear cataract. Nuclear cataract is observed as the clouding or opacity developed in the nucleus of the human lens, and it is the most common type of age-related cataract, which affects more than 20 million people worldwide and is reported as the leading cause of blindness [35]. For all images, they were captured by a Topcon DC-1 digital slit-lamp camera with FD-21 flash attachment. Each image was saved as a 24-bit color image of the size 2,048×1,536 pixels. For each image, a clinical grade was provided by senior ophthalmologists by consensus, indicating its severity of nuclear cataract disease, following the Wisconsin Cataract Grading System [36]. Figure 3 shows a set of four standard slit-lamp images used in the very grading system. These images together represent an increasing severity of cataract indicated by increasing integer grades (i.e., from 1 to 4), which construct annotated information ℓ introduced in Section "Methodology." In this database, the number of images with different grades is equivalent, which is 1,250 for each of the four grades. Hence, the utilized database is well balanced.

In each image, the lens and nucleus regions, which are believed to be more discriminative in identifying and diagnosing nuclear cataract disease in its conventional clinical diagnosis, were detected by a contour evolution method [37]. A 21-dimensional local feature vector was extracted from detected regions of each slit-lamp image as its visual attribute. The intensity, color, and texture within the nucleus region are calculated. The color of posterior subcapsular reflex (Fig. 4) is selected as it is the best location to judge the quality of opacity color. Detailed description of each local feature vector dimension is listed in Table 2.

Experiments Strategy and Implementation of Our Method

In this study, a 5-fold cross-validation [38] strategy is incorporated to evaluate the retrieval performance. For k-fold cross-validation, normally the original database is randomly partitioned into k equal-size subsets. Of the k subsets, a single subset is retained for testing the performance of the investigated model, while the remaining $(k-1)$ subsets are used for training (as well as validation in our study). The cross-validation process is then repeated k times (folds),

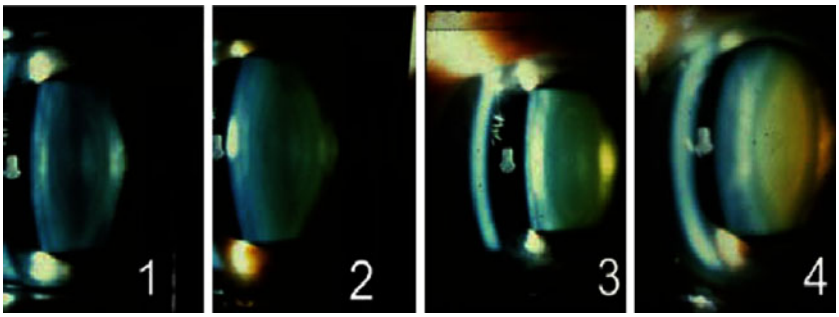


Fig. 3 Standard images with grades indicating the severity of nuclear cataract disease within slit-lamp images according to Wisconsin Cataract Grading System

with each of the k subsets used exactly once for testing. In this study, since a 5-fold cross-validation strategy is adopted, all 5,000 slit-lamp images are randomly partitioned into five equal-size subsets (i.e., 1,000 images per subset). One thing to emphasize here is that in each subset, the number of images with different grades is equivalent as well (in other words, there are 250 images per grade per subset). For onefold experiment, there are three subsets (including 3,000 images) used for training, one subset (including 1,000 images) used for validation, and one subset (including 1,000 images) used for testing. The subset used for testing in each fold is different, while the summation of them constructs the whole database (following the theory of k-fold cross-validation). In the testing subset of every fold, we utilize each one of the 1,000 testing images as a query image, and conduct retrieval on all other images in the testing subset.

For our method, we empirically set $T = 100$ and $\eta = 0.01$ as inputs in Table 1. Values of parameters T and η in this study are determined based on a trial-and-error scheme. When determining the learning rate η, we fixed the value of T, and switched the value of η within a range, in order to short-list values of η that can provide us the highest SKT based on multiple queries and their retrieved results. Figure 5a depicts a curve of SKT (averaged on multiple retrieval results) when selecting different η values. It can be observed that $\eta = 0.01$ can provide the highest SKT, and it is chosen as the predefined value of η in the implementation of our method. When determining the iteration times T, it is similar towards the above operation. From 5b, it can be

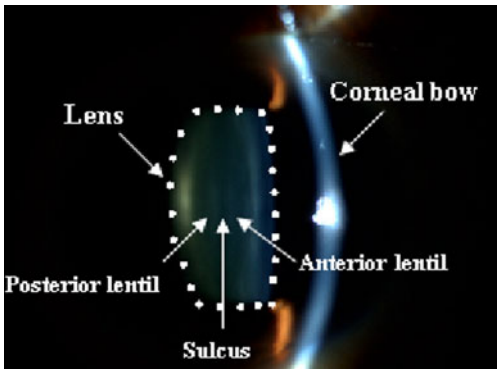


Fig. 4 An illustration of lens structure

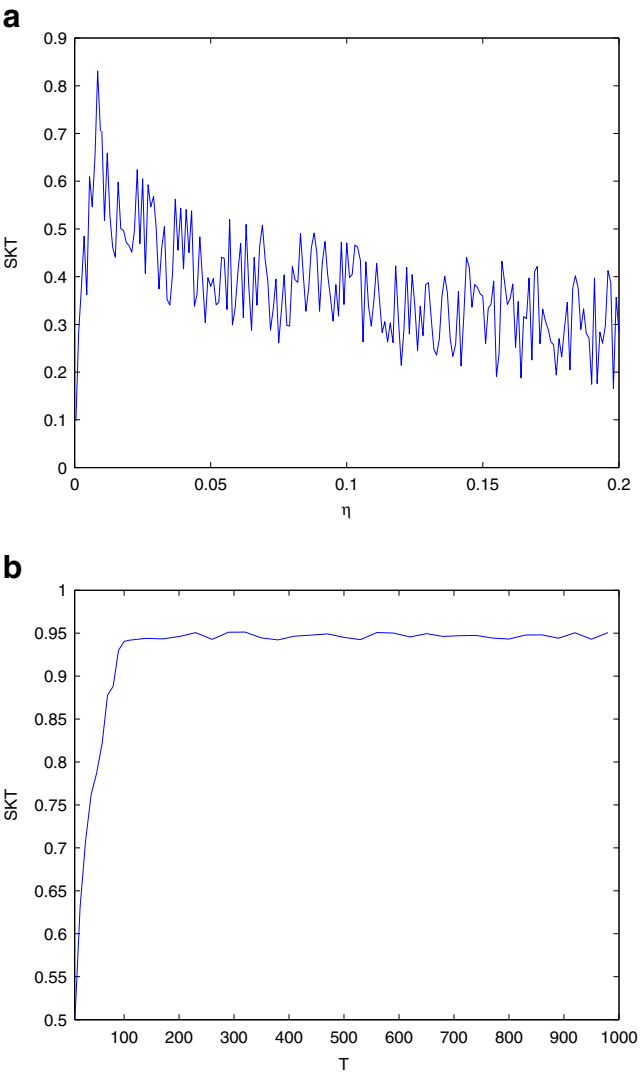


Fig. 5 **a** The influence of different values of learning rate η. **b** The influence of different values of iteration times T

concluded that $T = 100$ can provide good balance between effectiveness (high SKT) and efficiency (low iteration times, low computational cost in training therein).

Methods to Compare

Besides our method, there are five other popular distance learning methods implemented for performance comparison in this CBMIR application, including four supervised learning methods: large margin nearest neighbor (LMNN) [25], SVM+evidence maximization (SVM+EV) [39], SVM+leave-one-out error minimization (SVM+LOO) [40], SVM+radius/margin bound minimization (SVM+RM) [41], and a popular semi-supervised learning method with side-information (Side+Semi) [28]. Their basic ideas as well as implementation strategies in our experiments are explained below:

Large Margin Nearest Neighbor

LMNN aims to learn a Mahalanobis distance to reflect the pairwise similarity for a kNN classification by semi-definite

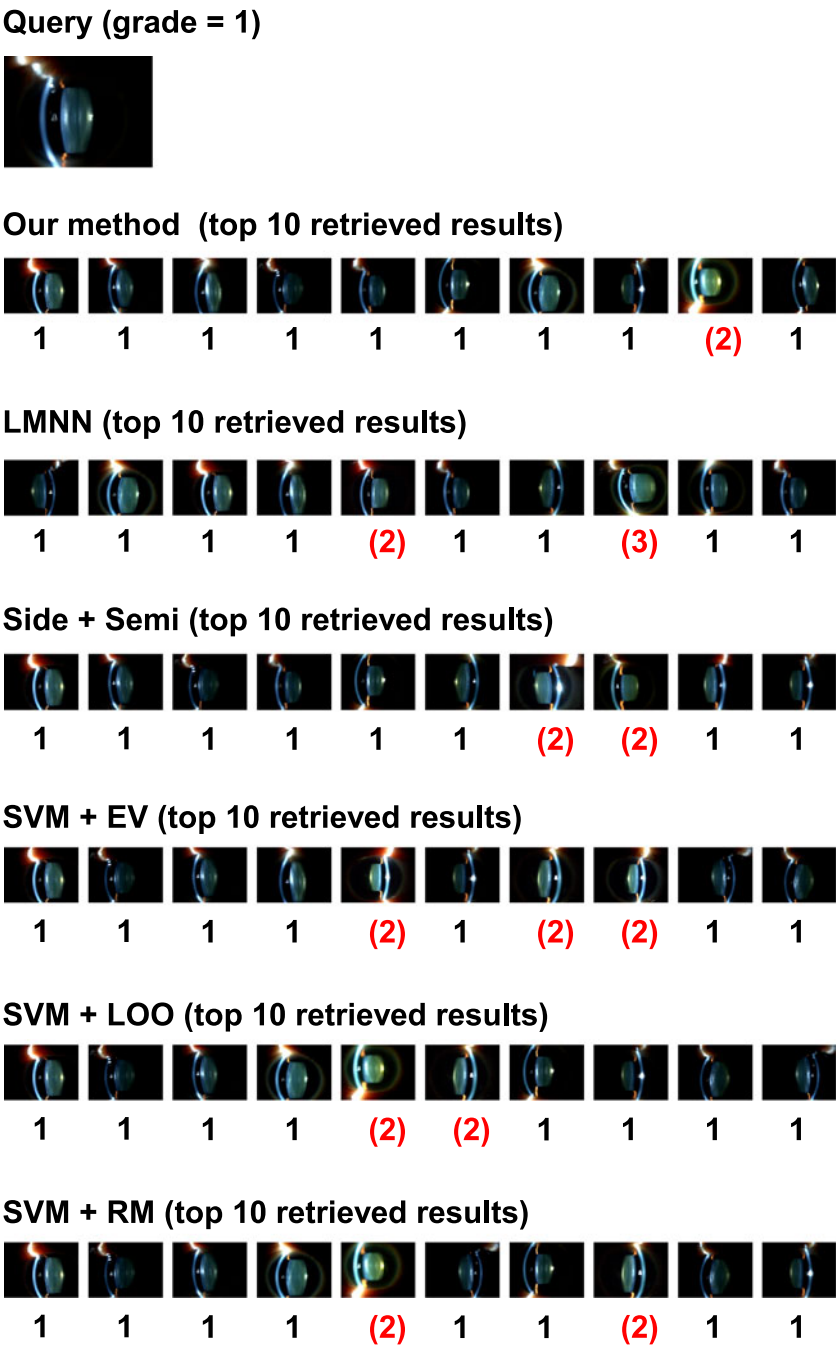


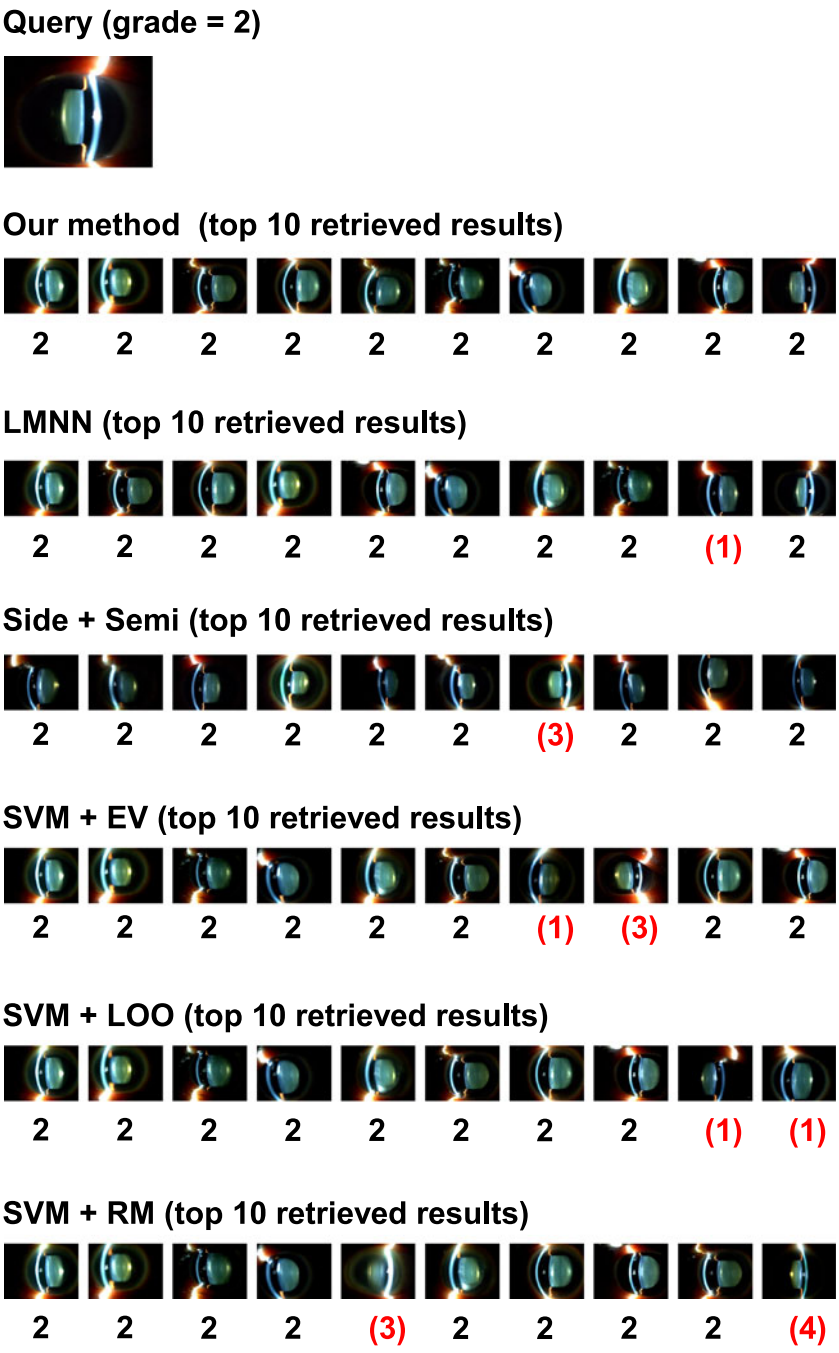
Fig. 6 Examples of retrieving slit-lamp images based on queries with diverse grades via different methods

programming [25]. Let $\{(x_i, y_i)\}$ denote a training set of labeled examples x_i with discrete class labels y_i. The cost function to optimize in LMNN is as follows.

$$\varepsilon(L) = \sum_{i,j} \eta_{i,j} \left\| L(x_i - x_j) \right\|^2 + c \sum_{i,j,l} \eta_{i,j} (1 - y_{i,l}) \left[1 + \left\| L(x_i - x_j) \right\|^2 - \left\| L(x_i - x_l) \right\|^2 \right]_+ , \quad (8)$$

where L denotes the parameter to learn. The distance between x_i and x_j: $\|L(x_i - x_j)\|^2$ is rewritten using $M = L^T L$ from a Mahalanobis distance $(x_i - x_j)^T M (x_i - x_j)$. $\eta_{i,j}$ is an indicator with binary outputs $\{0, 1\}$ to suggest whether x_i is in the nearest neighborhood of x_j or not. $y_{i,l}$ is a matrix with binary elements to indicate whether the label y_i of data x_i matches the label y_l of data x_l or not. $[\cdot]_+ = \max(\cdot, 0)$ is a hinge loss and c represents a positive constant. For the implementation of this method, we incorporate the code provided by the authors [42] and test

Fig. 6 (continued)

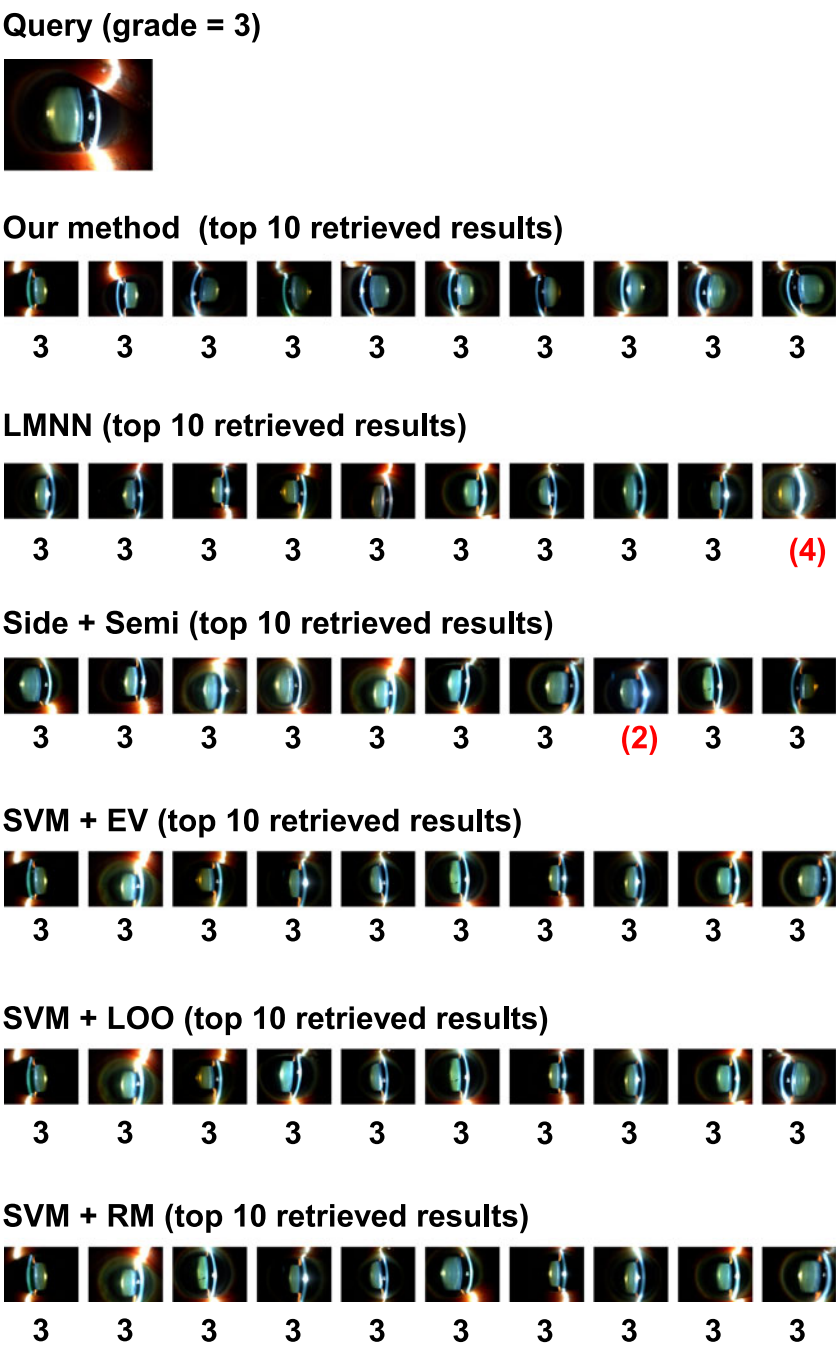


its performance in our CBMIR application. c is set as 0.05 in implementation.

Support Vector Machine

Kernel methods, which make SVM classifiers fit maximum-margin hyperplane in high-dimensional transformed feature space, are usually incorporated to separate data which may not be differentiated in their original low-dimensional feature space. A kernel function $K(x_i, x_j) =< h(x_i), h(x_j) >$ is usually computed as the inner product between transformed feature vectors $h(x_i)$ and $h(x_j)$, in which $h(\cdot)$ may not be explicitly known. Hence, the similarity between transformed feature vectors $h(x_i)$ and $h(x_j)$ is determined by their mapping function $h(\cdot)$, which is also directly influenced by the adopted kernel function $K(\cdot)$. Therefore, similarity learning in SVM can be realized via kernel learning, instead. There are various methods proposed for kernel learning in SVM in literature. In this study, we implemented three popular learning methods among them. The first one

Fig. 6 (continued)

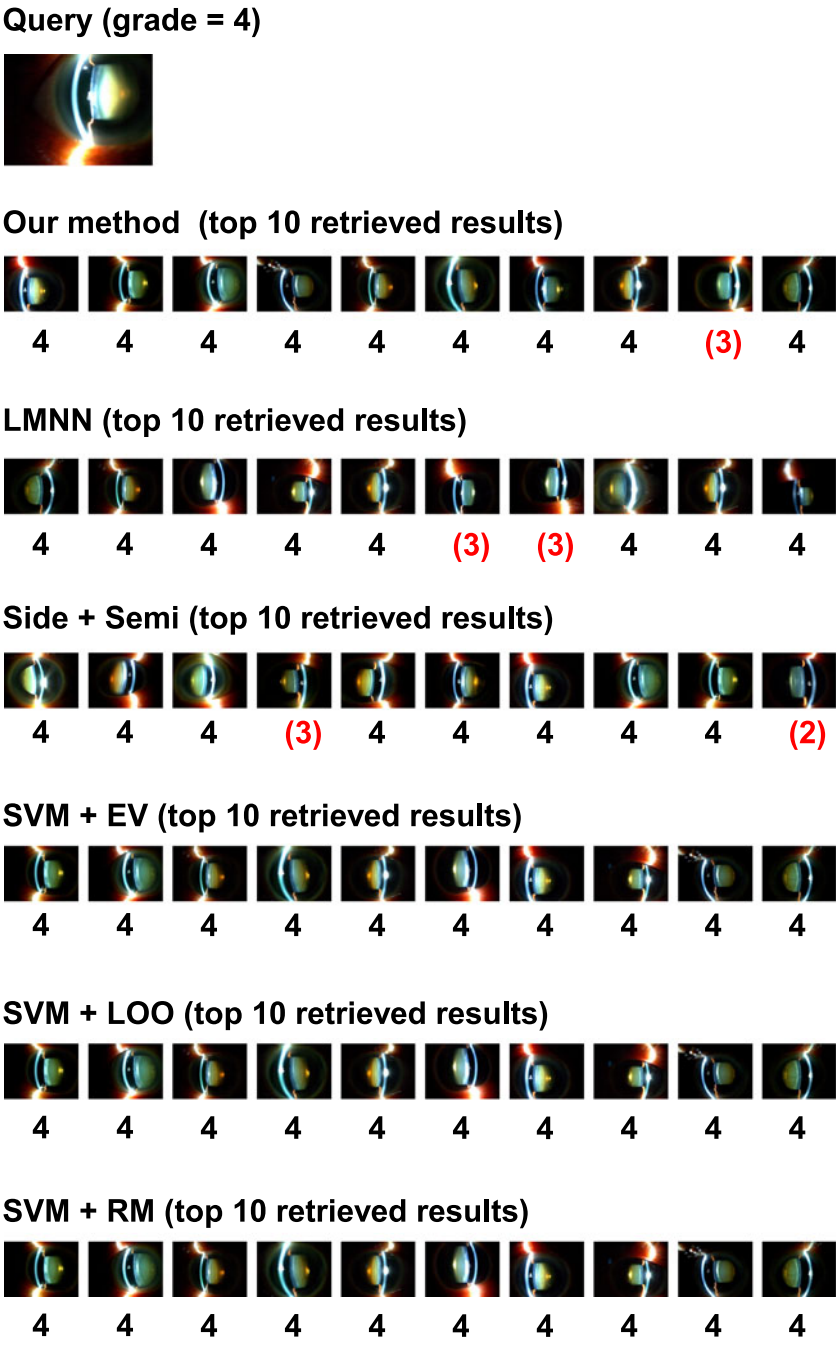


follows the leave-one-out (LOO) strategy [40]. The main idea of LOO is to remove one element from the whole training database, construct a decision rule based on the remaining training data, and test the rule on the removed element. Parameters which can provide the smallest error will be chosen. The second one is the radius–margin bound method (RM) [41], in which an upper bound based on the number of errors of a leave-one-out procedure is minimized for learning parameters. The third one is through evidence maximization [39]. In this study, we incorporate the code in [43] for implementing kernel learning in SVM. The adopted kernel function here is the popular Gaussian radius basis function. For implementations of SVM methods in this study, the tolerance of termination criterion in optimization is set as 0.001.

Semi-Supervised Learning with Side-Information

Xing et al. proposed one popular semi-supervised learning study to conduct Mahalanobis distance learning with the help of side-information [28]. The pairwise similarity constraint in side-information is applied on a convex

Fig. 6 (continued)



optimization function and the Mahalanobis distance is learned via an iterative gradient descent method [28]. The termination tolerance is set as 0.001 in implementation of this method.

Experimental Evaluation and Statistical Analysis

For the above five compared methods, 4,000 images except for the 1,000 testing images in each fold were used for training as there were no validations required in these methods. In this study, four integers, i.e., 1–4, are annotated grades for all images, and higher grades indicate more severe nuclear cataract disease. For one query image with a specific grade, retrieved images with the same grade are considered as "relevant images," while others are "irrelevant images." In order to quantify the retrieval performance, we incorporate two well-known criteria commonly used in general image retrieval studies: precision and recall. Their definitions are as follows:

$$\text{precision} = \frac{\text{number of retrieved relevant image}}{\text{number of retrieved images}} \tag{9}$$

$$\text{recall} = \frac{\text{number of retrieved relevant image}}{\text{number of all relevant images}}. \tag{10}$$

Figure 6 illustrated four example query images with different grades (i.e., from 1 to 4) and their top 10 retrieved results provided by our method as well as all other compared methods. The number below each retrieved image depicts its annotated grade provided by ophthalmologists (utilized as ground truth when evaluating the retrieval performance). Images with grades highlighted in red within brackets indi-

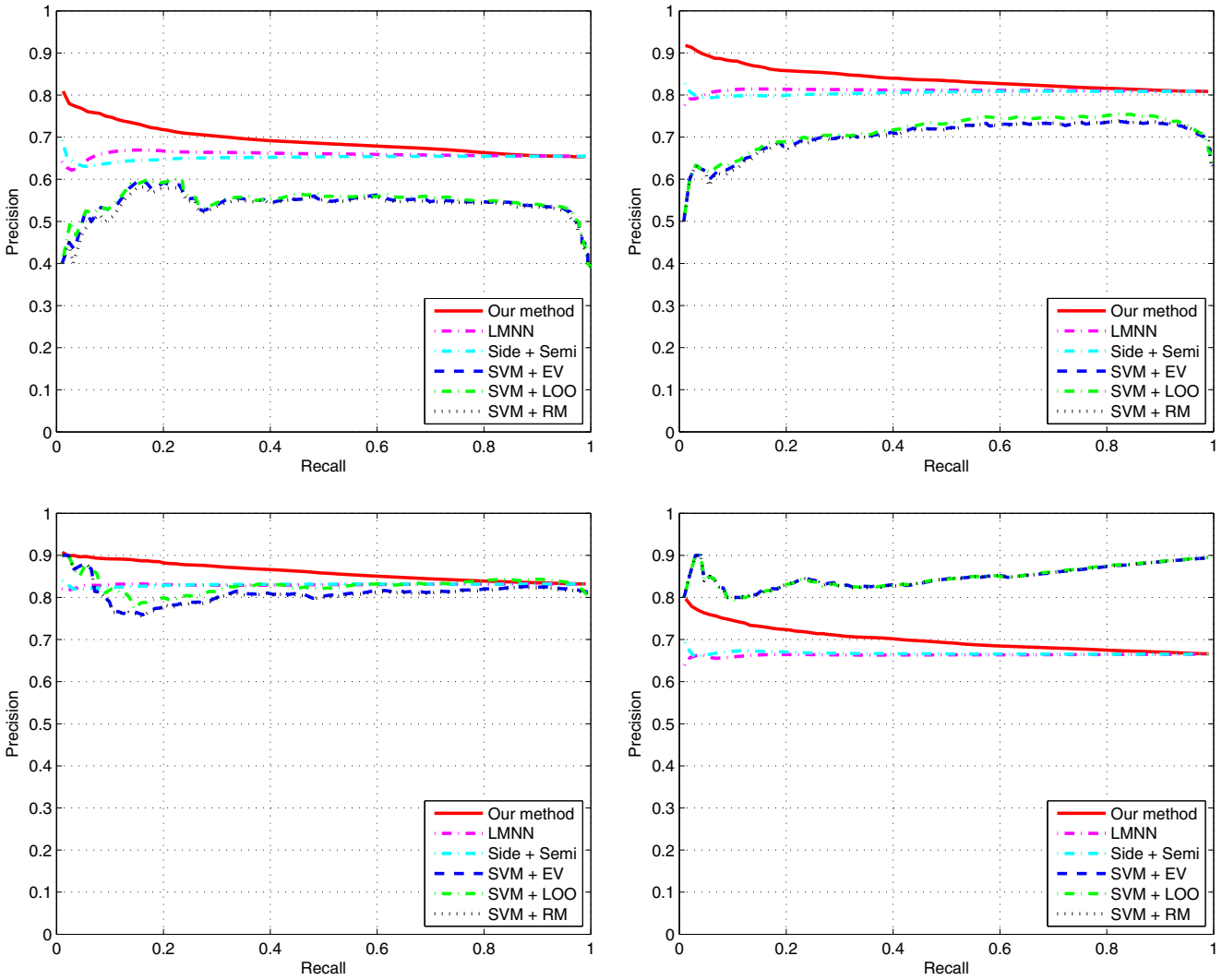


Fig. 7 Precision–recall curves of methods retrieving medical images depicting various degrees of severity of the nuclear cataract disease (from left to right, up to down: with grades 1, 2, 3, and 4)

cate retrieval errors. It can be observed that our method has the least retrieval errors when retrieving slit-lamp images with grades from 1 to 3 (in Fig. 6, the retrieval errors of our method are 1/ 0/ 0 for grade 1/ 2/ 3, respectively). When retrieving slit-lamp images of grade 4, the retrieval error of our method is larger than SVMs.

Figure 7 illustrates precision–recall curves of all methods when retrieving medical images with different degrees of disease severity. Precision and recall results depicted on each curve are averaged values based on retrieved results of all queries with the particular grade from five folds. It can be observed that when retrieving slit-lamp images with grades ranging from 1 to 3, the curves of our method are always above those of others, indicating that the mean average precision (MAP) of our method (represented by the geometrical area under the precision–recall curve) is the highest among all methods. When retrieving slit-lamp images with grade 4, our precision–recall curve is lower than those of SVMs. The reason for this is because that images with grade 1 are more or less like ordinary eye images; images with grades 2 and 3 are with medium nuclear cataract severity; images with grade 4 are with the highest severity, and they can be even identified via obvious color within the lens region since proteins for patients of the highest severity are prone to bind in the lens. It can be observed from example images in Fig. 3 that images with grade 4 look quite different from those with other grades. In this study, since such color information is also incorporated as components of extracted visual feature vectors (e.g., hue, saturation, intensity (HSI) inside the lens and nucleus regions in Table 2) and SVM is a discriminant-based method, it is expected to achieve better performance than the proposed relative comparison method, which is based on ranking and not discriminant.

Averaged precision–recall curves based on all retrieved results in this study are illustrated in Fig. 8a. It can be noticed that the averaged precision–recall curve of our method is above ones of other compared methods, indicating that the MAP of our method is higher than ones of others when taking all experimental results into consideration. Hence, the superiority of adopting the idea of ranking in similarity learning in this study is demonstrated. A box-and-whisker plot depicting all precision results based on all retrieved results is demonstrated in Fig. 8b. In each box, a red horizontal line is drawn across each box representing the median of precision of each method, while the upper and lower quartiles of precision in each box are depicted by blue lines above and below the median. A vertical dashed line is drawn up from the upper and down from the lower quartiles to their most extreme data points, which are within a 1.5 interquartile range (IQR) [38]. In each box, every data point beyond ends of a 1.5 IQR is marked by a red plus symbol. It can be observed that the box of our method is located higher than those of others, which suggests that our

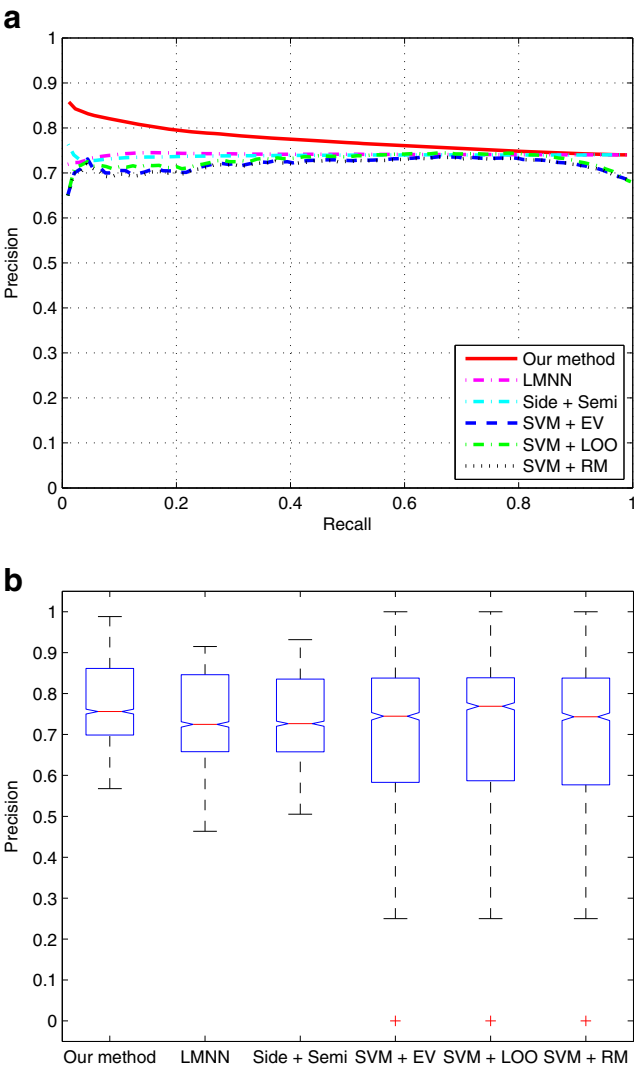


Fig. 8 **a** Averaged precision–recall curves on all experimental results. **b** Box plot of precision of all experimental results

method is superior to other compared methods for retrieving slit-lamp images in terms of precision. Also, the box range of our method is less than those of others, which suggests that our method is more stable and robust when handling medical images with various degrees of disease severity.

A statistical analysis is further conducted to substantiate the superiority of our method based on precision results from statistical point of view. After performing one-way analysis of variance [38], the p value is 0, which suggests that our method cannot share the same overall precision mean with others. Thus, a series of post hoc multiple comparison tests [38] are further incorporated to discern which method is superior. Results of multiple comparison tests between paired methods are listed in Table 3. It can be observed from entries in the third column that the preci-

Table 3 Multiple comparison test of all compared methods on precision

Method I	Method II	Precision Mean Difference (I-II)	A 95 % Confidence Interval
Our method	LMNN	0.0316	[0.0199, 0.0433]
Our method	Side + Semi	0.0337	[0.0220, 0.0454]
Our method	SVM + EV	0.0542	[0.0424, 0.0659]
Our method	SVM + LOO	0.0464	[0.0347, 0.0581]
Our method	SVM + RM	0.0562	[0.0445, 0.0679]
LMNN	Side + Semi	0.0021	[−0.0096, 0.0138]
LMNN	SVM + EV	0.0225	[0.0108, 0.0343]
LMNN	SVM + LOO	0.0148	[0.0031, 0.0265]
LMNN	SVM + RM	0.0246	[0.0129, 0.0363]
Side + Semi	SVM + EV	0.0204	[0.0087, 0.0321]
Side + Semi	SVM + LOO	0.0127	[0.0010, 0.0244]
Side + Semi	SVM + RM	0.0225	[0.0108, 0.0342]
SVM + EV	SVM + LOO	−0.0077	[−0.0195, 0.0040]
SVM + EV	SVM + RM	0.0021	[−0.0097, 0.0138]
SVM + LOO	SVM + RM	0.0098	[−0.0019, 0.0215]

sion mean difference of our method against others (i.e., our method minus others) are always positive, which indicates the superiority of our method from single-value estimation perspective. Another interval estimation of the precision mean difference is given by a 95 % confidence interval, which estimates a range that the mean difference between two compared methods is likely to be included. Entries in the fourth column of Table 3 show that the lower and upper bounds of each interval in comparisons between our method and other methods are positive as well. It gives a strong indication that our method is superior to others in terms of precision from both statistical single-value and interval estimation perspectives.

Discussion

For the utilized similarity $s_{(q,x)} = \exp(- < a, |f_q - f_x| >)$ in this study, a is the parameter to be learned and it is a 21-dimensional vector with 21 unknown elements corresponding to the 21-dimensional feature vector extracted from slit-lamp images according to Table 2. Therefore, the learned result a in this study can actually indicate weights (i.e., degree of importance) of different dimensions in the feature space, and the learned result a can help ophthalmologists concentrate more on those dimensions with higher weights (high relevance regarding the diagnosis of nuclear cataract disease severity). Figure 9 summarizes the distribution of all

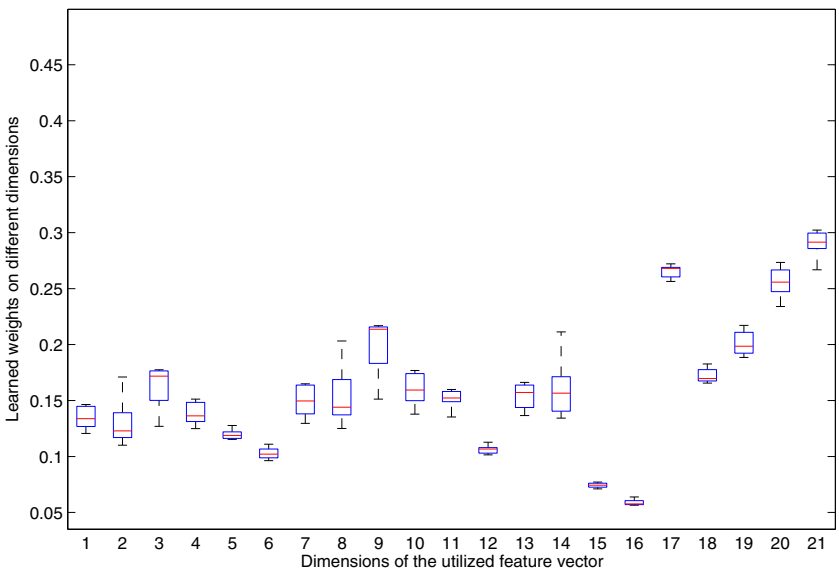


Fig. 9 Summary of learned weights on different dimensions of the constructed feature space according to Table 2

21 elements in the learned parameter a based on all retrieved results in our experiments. The box style in Fig. 9 is the same as the one in Fig. 8b. It can be noticed that the 9th, 17th, 19th, 20th, and 21st elements have more prominent weights compared with others. According to Table 2, the 9th element corresponds to the saturation (one of the HSI color components) inside the nucleus; the 17th element corresponds to the edge of nucleus; the 19th to 21st elements correspond to the HSI color components on the posterior reflex. It can be concluded that during the diagnosis of nuclear cataract, the color information is often of great importance and regions of nucleus as well as posterior reflex are of more relevance towards nuclear cataract disease. This learned result also complies well with literature in classic clinical ophthalmology books, in which nucleus, reflex, as well as their color information are highlighted in differentiating and diagnosing nuclear cataract disease [44, 45].

Conclusion

Extremely large number of medical images which are produced daily and commonly seen in hospitals, research institutes, medical colleges, etc. represent an enormous challenge towards these health-care organizations, as it is often hard to manage, access, and share the large scale of data efficiently. CBMIR is a popular way to facilitate the above process. In this paper, the similarity learning problem in CBMIR is mainly addressed. Inspired by recent studies to tackle information retrieval as a ranking procedure, a novel relative comparison-based similarity learning strategy is presented and its performance is evaluated using a large database composed of 5,000 images. Experimental results reveal that this newly presented learning paradigm via relative comparison is superior to other compared conventional supervised and semi-supervised similarity learning methods in the introduced CBMIR application. The effectiveness of incorporating the idea of ranking and relative comparison into CBMIR is also demonstrated.

Main contributions of this study can be summarized as follows: (1) Compared with conventional learning methods normally conducting the similarity learning task in CBMIR as a supervised classification problem, this study presents a novel semi-supervised learning paradigm via relative comparison. It is also the first attempt to incorporate the idea of relative comparison in CBMIR. (2) Technically, a new rank correlation measure is proposed and used for similarity learning via direct optimization. In the future studies, we will continue with the idea of ranking in similarity learning in CBMIR applications, try to propose more sophisticated learning methods and use other database made up of diverse medical imaging modalities for evaluation purpose.

Appendix

In Eq. 6, the second term $\frac{\exp\left(2(\ell_{(q,y)}-\ell_{(q,x)})\right)-1}{\exp\left(2(\ell_{(q,y)}-\ell_{(q,x)})\right)+1}$ of the right-hand side (RHS) of SKT can be viewed as a coefficient in the derivation, since it is not related to parameters a to learn in this study. Thus, the following derivation only focuses on the first term $\frac{\exp\left(2(s_{(q,x)}-s_{(q,y)})\right)-1}{\exp\left(2(s_{(q,x)}-s_{(q,y)})\right)+1}$ of RHS in SKT. For the ease of writing, let us denote the second term $\frac{\exp\left(2(\ell_{(q,y)}-\ell_{(q,x)})\right)-1}{\exp\left(2(\ell_{(q,y)}-\ell_{(q,x)})\right)+1}$ as term *coeff*. After differentiation, Eq. 6 becomes:

$$
\begin{aligned}
\nabla \mathrm{SKT}(a) = \frac{1}{N_n} \cdot \Bigg(\sum_{x,y \in D, x \neq y} & \frac{2 \cdot \exp\big(2(s_{(q,x)} - s_{(q,y)})\big)\big(s_{(q,x)} - s_{(q,y)}\big)' \cdot \Big(\exp\big(2(s_{(q,x)} - s_{(q,y)})\big) + 1\Big)}{\Big(\exp\big(2(s_{(q,x)} - s_{(q,y)})\big) + 1\Big)^2} \\
& \frac{-2 \cdot \exp\big(2(s_{(q,x)} - s_{(q,y)})\big)\big(s_{(q,x)} - s_{(q,y)}\big)' \cdot \Big(\exp\big(2(s_{(q,x)} - s_{(q,y)})\big) - 1\Big)}{\Big(\exp\big(2(s_{(q,x)} - s_{(q,y)})\big) + 1\Big)^2} \cdot coeff \Bigg) \\
= \frac{1}{N_n} \cdot \Bigg(\sum_{x,y \in D, x \neq y} & \frac{4 \cdot \exp\big(2(s_{(q,x)} - s_{(q,y)})\big)\big(s_{(q,x)} - s_{(q,y)}\big)'}{\exp^2\big(2(s_{(q,x)} - s_{(q,y)})\big) + 2\exp\big(2(s_{(q,x)} - s_{(q,y)})\big) + 1} \cdot coeff \Bigg) \\
= \frac{1}{N_n} \cdot \Bigg(\sum_{x,y \in D, x \neq y} & \frac{4 \cdot \big(s_{(q,x)} - s_{(q,y)}\big)'}{\exp\big(2(s_{(q,x)} - s_{(q,y)})\big) + \exp\big(2(s_{(q,y)} - s_{(q,x)})\big) + 2} \cdot coeff \Bigg)
\end{aligned}
\tag{11}
$$

Then, after replacing the term *coeff* with its original mathematical form, the gradient can be rewritten as:

$$\nabla \mathrm{SKT}(a) = \frac{1}{N_n} \cdot \left(\sum_{x,y \in D, x \neq y} \frac{4 \cdot \left(\frac{\partial s_{(q,x)}}{\partial a} - \frac{\partial s_{(q,y)}}{\partial a} \right) \cdot \frac{\exp\left(2(\ell_{(q,y)} - \ell_{(q,x)})\right) - 1}{\exp\left(2(\ell_{(q,y)} - \ell_{(q,x)})\right) + 1}}{\exp\left(2(s_{(q,x)} - s_{(q,y)})\right) + \exp\left(2(s_{(q,y)} - s_{(q,x)})\right) + 2} \right) \tag{12}$$

which is the same as Eq. 7.

References

1. Frost, Sullivan: US Data storage management markets for healthcare. Online resource: http://www.frost.com
2. ARRS GoldMiner: Online resource: http://goldminer.arrs.org/home.php
3. Hearst M, Divoli A, Guturu H, Ksikes A, Nakov P, Wooldridge M, Ye J: BioText search engine: beyond abstract search. Bioinformatics 23(16):2196–2197, 2007
4. iMedline: Online resource: http://archive.nlm.nih.gov/iti/
5. 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 Med Inform 73(1): 1–23, 2004
6. Müller H, Deselaers T: Tutorial on medical image retrieval: content-based image retrieval. Med Inf Europe 2005. Online resource: http://thomas.deselaers.de/teaching/files/tutorial_mie06/ContentBasedImageRetrieval.pdf
7. Tagare H, Jaffe C, Duncan J: Medical image databases: a content-based retrieval approach. J Am Med Inform Assoc 4(3):184–198, 1997
8. Ghosh P, Antani S, Long R, Thoma G: Review of medical image retrieval systems and future directions. IEEE Int Symp Computer-Based Med Syst 1:1–16, 2011
9. Sweet H, Mutalik P, Neklesa V, Horvath L, Lee C, Richter J, Tocino I, Fisher P: Voice-activated retrieval of mammography reference images. J Digit Imaging 11(2):65–73, 1998
10. Oh J, Yang Y, El Naqa I: Adaptive learning for relevance feedback: application to digital mammography. Med Phys 37(8):4432–4444, 2010
11. El Naqa I, Yang Y, Galatsanos N, Nishikawa R, Wernick M: A similarity learning approach to content-based image retrieval: application to digital mammography. IEEE Trans Med Imaging 23(10):1233–1244, 2004
12. Kinoshita S, de Azevedo-Marques P, Pereira R, Rodrigues J, Rangayyan R: Content-based retrieval of mammograms using visual features related to breast density patterns. J Digit Imaging 20(2):172–190, 2007
13. Tourassi G, Harrawood B, Singh S, Lo J, Floyd C: Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammography. Med Phys 34(1): 140–150, 2007
14. Zheng B, Lu A, Hardesty L, Sumkin J, Hakim C, Ganott M, Gur D: A Method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment. Med Phys 33(1):111–117, 2006
15. Mazurowski M, Habas P, Zurada J, Tourassi G: Decision optimization of case-based computer-aided decision systems using genetic algorithms with application to mammography. Phys Med Biol 53(4):895–908, 2008
16. Dy J, Brodley C, Kak A, Broderick L, Aisen A: Unsupervised feature selection applied to content-based retrieval of lung images. IEEE Trans on Pattern Anal Mach Intell 25(3):373–378, 2003
17. Shyu C, Brodley C, Kak A, Kosaka A, Aisen A, Broderick L: ASSERT: a physician-in-the-loop content-based retrieval system for HRCT image databases. Comp Vision Image Underst 75(1–2):111–132, 1999
18. Lehmann T, Güld M, Thies C, Fischer B, Spitzer K, Keysers D, Ney H, Kohnen M, Schubert H, Wein B: Content-based image retrieval in medical applications. Methods Inf Med 43:354–361, 2004
19. Huang W, Chan K, Li H, Lim J, Liu J, Wong T: A computer assisted method for nuclear cataract grading from slit-lamp images using ranking. IEEE Trans Med Imaging 30(1):94–107, 2011
20. Yang L, Jin R: Distance metric learning: a comprehensice survey. Online resource: http://www.cs.cmu.edu/liuy/frame_survey_v2.pdf.
21. Guo G, Jain A, Ma W, Zhang H: Learning similarity measure for natural image retrieval with relevance feedback. IEEE Trans Neural Netw 13:811–820, 2002
22. Hertz T, Bar-Hillel A, Weinshall D: Learning distance functions for image retrieval. IEEE Conf Comput Vis Pattern Recognit 2:570–577, 2004
23. Si L, Jin R, Hoi S, Lyu M: Collaborative image retrieval via regularized metric learning. Mulitmedia Systems 12:34–44, 2006
24. Globerson A, Roweis S: Metric learning by collapsing classes. Adv Neural Inf Process Syst 1:451–458, 2005
25. Weinberger K, Blitzer J, Saul L: Distance metric learning for large margin nearest neighbor classification. Adv Neural Inf Process Syst 1:1473–1480, 2006
26. Zhu X: Semi-supervised learning literature survey, Online resource: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.9681&rep=rep1&type=pdf
27. Chapelle O, Scholkopf B, Zien A: Semi-supervised learning. MIT Press, 2006
28. Xing E, Ng A, Jordan M, Russell S: Distance metric learning, with application to clustering with side-information. Adv Neural Inf Process Syst 1:505–512, 2003
29. Schultz M, Joachims T: Learning a distance metric from relative comparisons. Adv Neural Inf Process Syst 1:41–48, 2004
30. Schroff F: Harvesting image databases from the web. IEEE Trans Pattern Anal Mach Intell 33(4):754–766, 2011
31. Jing Y: VisualRank: applying pagerank to large-scale image search. IEEE Trans Pattern Anal Mach Intell 30(11):1877–1890, 2008

32. Siddiquie B: Image ranking and retrieval based on multi-attribute queries. IEEE Conf Comput Vis Pattern Recognit 1:801–808, 2011
33. Datta R, Joshi D, Li J, Wang J: Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40(2):5:1–5:60, 2008
34. Kendall M: A new measure of rank correlation. Biometrika 30: 81–93, 1938
35. World Health Organization: The world health report: life in the 21st century - a vision for all. Online resource: http://www.who.int/whr/1998/en/whr98_en.pdf, 1998
36. Klein B, Klein R, Linton K, Magli Y, Neider M: Assessment of cataracts from photographs in the Beaver dam eye study. Ophthalmology 97(11):1428–1433, 1990
37. Li C, Kao C, Gore J, Ding Z: Implicit active contours driven by local binary fitting energy. IEEE Conf Comput Vis Pattern Recognit 1:1–7, 2007
38. Rice J: Mathematical Statistics and Data Analysis, 2nd edition. Duxbury Press, 2007
39. Gold C, Sollich P: Model selection for support vector machine classification. Neurocomputing 55:221–249, 2003
40. Weston J: Leave-one-out support vector machines. Int Joint Conf Artif Intell 1:727–733, 1999
41. Chapelle O, Vapnik V, Bousquet O, Mukherjee S: Choosing multiple parameters for support vector machines. Mach Learn 46(1–3):131–159, 2002
42. Weinberger K: Large margin NN classifier - matlab toolbox. Online resource http://www.weinbergerweb.net/Downloads/LMNN.html
43. Chapelle O: Selection of kernel parameters - matlab toolbox. Online resource http://olivier.chapelle.cc/ams/
44. Congdon N, Chang N, Botelho P, Stark W, Datiles III M: Cataract: clinical types. Duane's Ophthalmology, 2006 edition, vol 1, chapter 73(B). Lippincott: Williams & Wilkins, 2006
45. Zigler Jr J, Datiles III M: Pathogenesis of cataracts. Duane's Ophthalmology, 2006 edition, vol 1, chapter 72(B). Lippincott: Williams & Wilkins, 2006