DOI 10.1007/s10916-016-0659-2

Medical Image Retrieval Using Vector Quantization and Fuzzy S-tree

Jana Nowaková[1] · **Michal Prílepok**[1] · **Václav Snášel**[1]

Received: 31 March 2016 / Accepted: 15 November 2016 / Published online: 15 December 2016

Abstract The aim of the article is to present a novel method for fuzzy medical image retrieval (FMIR) using vector quantization (VQ) with fuzzy signatures in conjunction with fuzzy S-trees. In past times, a task of similar pictures searching was not based on searching for similar content (e.g. shapes, colour) of the pictures but on the picture name. There exist some methods for the same purpose, but there is still some space for development of more efficient methods. The proposed image retrieval system is used for finding similar images, in our case in the medical area – in mammography, in addition to the creation of the list of similar images – cases. The created list is used for assessing the nature of the finding – whether the medical finding is malignant or benign. The suggested method is compared to the method using Normalized Compression Distance (NCD) instead of fuzzy signatures and fuzzy S-tree. The method with NCD is useful for the creation of the list of similar cases for malignancy assessment, but it is not able to capture the area of interest in the image. The proposed method is going to be added to the complex decision support system to help to determine appropriate healthcare according to the experiences of similar, previous cases.

Keywords Vector quantization · Image comparison · Image classification · TF-IDF · Fuzzy S-tree · Medical image · NCD

Introduction

The increasing amount of medical data stored in data sets is creating an urgent need for efficient algorithms for information retrieval. The bigger the databases are, the greater the chance that even a small relation or unobserved information at first sight will be found [24]. The types of data stored include not only plain text but also images, sounds and other types of digitalized information. While sophisticated methods for full-text retrieval have been created, the efficient retrieval of other types of data such as image, voice or video data remains a still open problem.

In the article, we present a method for fuzzy medical image retrieval (FMIR), concretely in mammography. In the medical area, there are some proposed approaches for image retrieval. Such as in [11] the authors proposed to use Wavelet transforms in a one and two-level decomposition for finding the normal regions and the regions containing the lesion, then they proposed using the coefficients obtained using the wavelet decomposition in a support vector machine classifier with a linear kernel. A very interesting approach is described in [50], the method enables the localization of the pathology area in medical images. The proposed approach utilizes the linguistic theory of pattern classification. The interpretation is built on cognitive processes, which is based on the human psychological

This article is part of the Topical Collection on *Image & Signal Processing*

✉ Václav Snášel
vaclav.snasel@vsb.cz

Jana Nowaková
jana.nowakova@vsb.cz

Michal Prílepok
michal.prilepok@vsb.cz

[1] Faculty of Electrical Engineering and Computer Science, Department of Computer Science, VŠB - Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava - Poruba, Czech Republic

processes of the understanding of the registered patterns in images. Our main inspiration was taken from information retrieval, document signatures, and NCD, the known approach is modified with the usage of vector quantization.

The article organisation is as follows: the workflow of the proposed FMIR method with respect to the necessity of the explanation of the theoretical principles of the proposed method. In the first part of this article, the theoretical principles of the methods used can be found. "Vector quantization" describes VQ, the following "TF-IDF" introduces TF-IDF. The next section explains signature files and the connection to fuzzy signatures and their organisation in computer memory ("Brief description of signature files" and "Fuzzy signatures"). The proposed method is compared to NCD, and it is defined in "Normalized compression distance". The following "Fuzzy medical image retrieval" and "Experiments" describe the proposed FMIR method and the setup of the experiments. The results of FMIR method and NCD are summarized in "Results". The last "Conclusion" sums up and concludes the experiment results.

Related work

Primarily, there are two ways how to search for similar images: querying the images is based on the keyword TBIR (Text-Base Image Retrieval) [45, 62] and based on the content CBIR (Content-Base Image retrieval) [13, 74].

The TBIR system can return the images with content that are not related to the request of querying since the nature of the keywords is independent of the content of the images. To overcome this problem, the CBIR system extracts the visual attributes of the images which are needed to query and then compare with the visual attributes of the other images. However, if the method of comparing the similarity of the content is ineffective, the results of querying put out the images with content which are not related to the requested query.

The similarity assessment method of images based on vector quantization, TF-IDF (Term Frequency–inverse Document Frequency), fuzzy signatures and the method of querying images based on the fuzzy S-tree (FS-tree), is built in the presented article. Vector quantization is used as a tool for the transformation of an image to a fuzzy signature and for creating the codebook. It was used in the past several times. A comparison of the vector quantization codebook was used for image retrieval in [58], Teng [67] uses vector quantization from image indexing and retrieval with the preservation of pixel position and Rahman et al. [55] uses vector quantization in fuzzy feature space for biomedical image retrieval. The proposed method is compared with the method, where Normalized Compression Distance (NCD) for similarity assessment is used.

NCD is based on Kolmogorov complexity. Calculation of Kolmogorov complexity requires high computing power. That is why in real applications it is approximated using compression algorithms [39]. The NCD has been used in text retrieval [21], text clustering, plagiarism detection [65], music clustering [21] and [49], music style modelling [17], automatic construction of the phylogeny tree based on whole mitochondrial genomes [42], the automatic construction of language trees [2] and [41], and the automatic evaluation of machine translations [16]. NCD has also been successfully applied in challenging areas, for example, spam detection [54] or EEG data for simple cognitive tasks [3]. The disadvantage of a method with NCD is that it is not able to capture the areas of interest in the image, which can help a physician. The distance for the assessment of the similarity of images is also proposed, which is used for the creation of the list of the similar images. This list is then utilized to determine if a medical finding is malignant or benign. This knowledge is obtained from the comparison of known cases and can be used for classification of the medical cases.

In the context of our search for new ways of looking at signature methods, we tried to apply a fuzzy approach to signature construction and manipulation. The result is the definition of fuzzy signatures which is presented. The side idea of the paper, which is going to be elaborated is to use fuzzy signatures to interface the TBIR and the CBIR.

The requirement for the widening of the palette of methods for image retrieval in our case is the growing usage of imaging methods in the medical area such as X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasonography (USG), positron emission tomography (PET) or single-photon emission computed tomography (SPECT) [29]. e.g. during the preoperative examination, during the surgery or the postoperative period, the amount of the picture, which needs to be assessed and compared is still rising [28, 70]. Nowadays, it is done using human power and requires a lot of time. One aspect of the power of the amount of images, which are stored in hospitals and research medical centers, is still neglected. In [73], the application of fuzzy signature for medical data is proposed. As fuzzy reasoning is used, the proposed method can be considered as bioinspired in a specific point of view.

Vector quantization (VQ) is a classical quantization technique which allows for the modeling of probability density functions through the distribution of prototype vectors. It was originally used for data compression. It works by dividing a large set of points (vectors) into groups with approximately the same number of points closest to them. Each group is represented by its centroid point, as in *k*-means or some other clustering algorithms. It has been used for image compression, CBIR, and image indexing for many years in many fields. In [1] Arya and Mount presented and compared three algorithms for nearest neighbor searching in high dimensions, within the framework of vector quantization.

Two of the algorithms provide drastic reductions in complexity with a negligible deterioration in performance. The papers [9, 10] presented VQ for image processing. The authors presented a multitude of various approaches for codebook construction and its application in the field of image processing. Park in [51] proposed a new block-transform-based image compression scheme by combining vector quantization (VQ) and two transformations, discrete cosine transform (DCT) and vector-embedded Karhunen-Loève transform (VEKLT) [66]. The author's proposed method performs better in peak signal-to-noise ratio if the bits per pixel (bpp) increase because the ratio of blocks which are processed by VQ and VEKLT increases. The visual quality of the proposed method is better than that of JPEG in high detail. The authors Huang and Harris in [25] investigated a performance comparison of several often-used VQ codebook generation algorithms. The discussed algorithms include the Linde-Buzo-Gray (LBG) binary splitting algorithm, the pairwise nearest neighbor algorithm, the simulated annealing algorithm, and the fuzzy c-means clustering analysis algorithm, and the new directed-search binary splitting method, which reduces the complexity of the LBG binary.

In [43], the authors presented an image indexing and retrieval scheme based on VQ. The results obtained by the authors show that it has a higher retrieval performance than the basic colour histogram based approach. In addition, it is computationally more efficient.

The increasing amount of digitally produced images in various areas, for example journalism, medicine, is requiring fast and effective ways of image accessing. The need for a fast and effective solution is growing. Deselaers et al. [15] discussed a large variety of features for image retrieval and a setup of five freely available databases. From the experiments conducted it can be deduced, which features perform well on which kind of task and which do not.

Vector quantization

Vector quantization has been used for image compression for many years. In this section, we will briefly review the basic concepts of VQ image compression. In most image compression techniques, the actual quantization or coding is done on scalars (e.g. on individual real value samples of waveforms or pixels of images). Transform coding does it by first taking the block transformation to a block of pixels and then individually coding the transform coefficients. Predictive coding does it by quantizing an error term formed as the difference between the new sample and a prediction of the new sample based on pasted coded outputs [67].

A fundamental result of Shannon's rate-distortion theory [61], the branch of information theory devoted to data compression, is that a better performance can always be achieved by coding vectors (a group of values) instead of scalar (individual value). Thus, vector quantization can successfully be used for image and audio compression. A vector quantizer can be defined as a mapping Q of k-dimensional vector space $\mathbb{R}^k$ into a finite subset $\mathcal{C}$ of $\mathbb{R}^k$, that is $Q : \mathbb{R}^k \rightarrow C$, where $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$. $\mathcal{C}$ is the set of reproduction vectors and is called a codebook. $|\mathcal{C}| = n$ is the number of vectors in $\mathcal{C}$. At the encoder, each data vector x belonging to $\mathbb{R}^k$ is matched or approximated with a codeword in the codebook and the address or index of that codeword is transmitted instead of the data vector itself. To find the best match codeword for a data vector, we can use Euclidean or Manhattan distance.

Vector quantization can be used for image transformation to fuzzy vector.

TF-IDF

TF-IDF [40], is a numerical statistic that is meant to reflect how important a word, n-gram or other term is to a document in a set of documents. TF-IDF is a weighting factor in text information retrieval and text mining applications. TF-IDF assigns a higher value of weight to terms which occurs rarely in the collection of documents. The terms with a higher occurrence frequency are evaluated with this lower weight. This means that these terms are less important in the documents.

TF-IDF is the product of two statistics values, TF and IDF. TF (Term Frequency) is a value that represents how often a particular term appears in a concrete document. The more frequent the term is in the document, the more important the term in the document is. The common terms in the collection, that appear in many documents, have the highest frequency. For example, the word "patient" is frequently mentioned in medical documents, and its TF would be very high. In other documents, this word can rarely be found and its TF is low. Therefore, we need to control the importance of the word. This situation is adjusted by IDF (Inverse Document Frequency). DF (Document Frequency) is obtained by dividing the total number of documents by the number of documents containing the term. IDF is the inverse value of DF. TF-IDF is widely used in many text data mining and data analysis applications [40].

Brief description of signature files

In [34], an application of the signature for the more efficient processing of range queries is studied. This approach puts the signature into multi-dimensional data structures like R-tree [23] or UB-tree [56] but original functionalities are preserved, i.e. the range query algorithm for the general range query.

At first, we would like to introduce the basic principles of general signature methods. For example the details can be found in [4, 18, 30, 76]. As was mentioned, we used the inspiration in information and image retrieval [36, 37, 48]. For the usage of signatures for images, we inspired ourselves in the signatures of text documents, so the principle is going to be described, for easier understanding, on a text – document example.

The information retrieval methods are able to separate a finite set of all documents $D = \{D_1, D_2, \ldots, D_k\}$ into two subsets – a set of relevant documents and the rest of the documents. The extracted documents in the subset of relevant documents are determined for a given query Q. A certain required amount of work is needed before a query can be evaluated. During this work, auxiliary data structures are created. These structures are used for later evaluation of any given query. In methods where signatures are utilised, signatures are considered as auxiliary structures.

In general, a signature is a bit string $s_1 s_2 \ldots s_n$, of a fixed length n. Each document has its own signature and is used for query evaluation. Both signatures, signature S_i of a document D_i and query signature S_Q, have a length of n. Document D_i is relevant only if its signature S_i contains ones in all positions in which ones are encountered in the query signature S_Q.

We can divide the ways of creating signatures into two main principles: superimposed coding and concatenation. In our article, we use superimposed coding and we will discuss this principle in more detail.

Every signature with superimposed coding is created in the following way. In the beginning, the bit string of signature $S_i = s_1 s_2 \ldots s_n$ contains zeros in all positions. A document consists of a finite set of distinct words $\{w_1, w_2, \ldots, w_l\}$. For each word w_i belongs m bit positions in the signature S_i. These bit word positions should not overlap and should be distributed evenly throughout the full length of the signature. If the bit string contains zeros in these positions, they are replaced with ones. The word bit position is usually chosen as a result of a hash function which takes the word w_j as an argument. The result of this method is the same as if a signature were created separately for each word w_j. Also, this signature would contain up to m ones. The signature of the whole document is the result of superimposing all individual signatures $S_{w_1}, S_{w_2}, \ldots, S_{w_l}$. An example of signature superimposed coding is described in Algorithm 1.

Algorithm 1 Signature algorithm with superimposed coding

```
Data: document $D = \{w_1, \ldots, w_l\}$
Result: document signature $S = s_1 \vee s_2 \vee \cdots \vee s_m$
set all positions in signature $S$ to 0;
get list of distinct words from document $D$;
forall the words $w$ in document wordList do
    get bit positions for word $w$;
end
forall the word signatures $s$ do
    superimpose word signature $s$ with document
    signature $S$;
end
return S;
```

Fuzzy signatures

Another issue to be resolved is the way fuzzy signatures are organized in the signature file. The simplest way would be to store the fuzzy signatures in sequential order. This method is not efficient if the time required for query evaluation is concerned. That is why we have modified the data structure called S-tree, which is traditionally used to store ordinary signatures. In the next section, we will present the modification which enables using this data structure to store fuzzy signatures.

In this chapter, we describe persistent data structures for searching efficiently, i.e., fast, for images similar to a query image within a large database. Mapping images to fuzzy signature is an elegant way how to solve the problem. Since we use fuzzy signatures to represent images we address the efficiency issue by proposing the use of an enhanced fuzzy signature tree (FS-tree) [52, 53, 63, 64].

Fuzzy signatures are hierarchical representations of data structuring into vectors of fuzzy values. A fuzzy signature is defined as a special multidimensional fuzzy data structure, which is a generalization of vector valued fuzzy sets. Vector valued fuzzy sets are special cases of L-fuzzy sets which were introduced in [19]. Fuzzy signatures are studied in [33, 69].

The weighted aggregation concept is proposed [46] to provide additional expert knowledge to the fuzzy signature structure by introducing the weighted relevance of each branch to its higher branches of the fuzzy signature structure. In [32], vector valued fuzzy sets are introduced as the generalization of the original concept of fuzzy sets [31, 75].

Definition 1 The fuzzy signature F is a vector $f_1 f_2 \ldots f_n$, where $f_i \in \langle 0; 1\rangle$ $\forall i = 1, 2, \ldots, n$.

Provided that we have created fuzzy signatures for all documents in the set D and that we have the fuzzy signature

F_Q of the query Q, we can use the operation of conjunction to find the relevant documents. The operation is defined in the same way as in fuzzy logic.

Definition 2 The conjunction of fuzzy signatures F_i and F_j is the fuzzy signature

$$F_i \otimes F_j = (f_{i_1} \otimes f_{j_1})(f_{i_2} \otimes f_{j_2}) \ldots (f_{i_n} \otimes f_{j_n}). \qquad (1)$$

The operation $\otimes$ could be defined as one of the possible functions of the family of the functions, which are called the operation of triangular norm (t-norm) see [20]. The example of the possible interpretation of the t-norm function is provided in Table 1.

In order to find all documents relevant to the given query Q we have to find all documents D_i which satisfy the formula $F_i \otimes F_Q = F_Q$. This means that for a document to be relevant to the query, all the elements of its signature must be equal to or greater than all the corresponding elements in the query signature.

Definition 3 The disjunction of fuzzy signatures F_i and F_j is the fuzzy signature

$$F_i \oplus F_j = (f_{i_1} \oplus f_{j_1})(f_{i_2} \oplus f_{j_2}) \ldots (f_{i_n} \oplus f_{j_n}). \qquad (2)$$

The operation $\oplus$ could also be chosen from the set of the functions, which are in this case called triangular co-norm – t-conorm or known as the s-norm [20] (Table 1).

These definitions of logical operations with fuzzy signatures reflect the t-norm and s-norm [20], the common definitions of logical operations in fuzzy logic.

We have described the method of evaluating a query by examining the fuzzy signatures of individual documents and the signature of the query itself, but we have not discussed how fuzzy signatures are constructed. We assume that the elements f_i of the string F are degrees of membership of the sets M_i. That is why using fuzzy signatures depend on determining the sets M_i.

It is possible to construct these sets in the following way, so let the set A be the set of all distinct words that can appear in all the documents. Let δ be the distance on the set A. We can apply clustering methods to the set A using the distance δ. It will give a finite number of clusters which will become the sets M_i. For each word w in the document D_i, the degrees of membership in the sets $M_1, M_2, \ldots, M_n$ can be estimated using the distance δ and used as the elements of the fuzzy signature of this word. The fuzzy signature F_i of the whole document D_i is the disjunction of the fuzzy signatures of all the words.

$$F_i = F_{w_{i1}} \oplus F_{w_{i2}} \oplus \cdots \oplus F_{w_{il}}. \qquad (3)$$

We could use the set of all distinct words in the corpus of a natural language as the set A. We could try to apply some of the metrics used in linguistics which reflect relationships between words in the corpus. However, this concept of fuzzy signature construction has not yet been implemented.

Fuzzy S-tree

Deppisch, in [14] proposed a B-tree like structure to facilitate fast access to the records (which are signatures) in a signature file. The leaf of an S-tree consists of similar (i.e. with small distance) signatures along with the document identifiers. The OR-ing of these k document signatures forms the key of an entry in an upper level node, which serves as a directory for the leaves. Higher-level nodes are constructed recursively, in the same way. Like a B-tree, the S-tree is kept balanced: when a leaf node overflows, it is split into two groups of similar signatures; the father node is changed appropriately to reflect the new situation. Splits may propagate upwards.

S-tree is a balanced tree which uses similar principles as the well-known B-tree or its variation, the B$^+$-tree. S-tree

Table 1 Example of possible interpretation of t-norms and t-conorms functions

Name	t-norms	t-conorms
Algebraic product & sum	$t_w(A, B) = AB$	$s_w(A, B) = A + B - AB$
Bold (Lukasiewicz, Bounded) product & sum	$A \otimes B = \max(0, A + B - 1)$	$A \oplus B = \min(1, A + B)$
Drastic product & sum	$t_w(A, B) = \begin{cases} \min(A, B) & \text{if } \max(A, B) = 1 \\ 0 & \text{else} \end{cases}$	$s_w(A, B) = \begin{cases} \max(A, B) & \text{if } \min(A, B) = 0 \\ 1 & \text{else} \end{cases}$
Einstein product & sum	$t_w(A, B) = \frac{AB}{2-(A+B-AB)}$	$s_w(A, B) = \frac{A+B}{1+AB}$
Hamacher product & sum	$t_w(A, B) = \frac{AB}{A+B-AB}$	$s_w(A, B) = \frac{A+B-2AB}{1-AB}$
Minimum & Maximum	$t_w(A, B) = \min(A, B)$	$s_w(A, B) = max(A, B)$

is a data structure which allows searching for, inserting and removing signatures [14, 47].

The modified S-tree [5] adds to the S-tree a third table (the count table) that stores, for each internal node, the number of leaves in its left subtree. The addition of the count table to the modified S-tree allows more efficient searches. In a modified S-tree, we can look up a given code without traversing the complete sequence, using the count values to jump forward in the sequence: if we want to go to the second child of a node, instead of traversing the complete first child we use the count to jump directly to the second child. It makes navigation faster but the space requirements are significantly higher.

The compact S-tree [6] is another variant of the S-tree that uses a single table (the linear-tree table) to represent the same information stored in the three tables of the modified S-tree. The compact S-tree is built from a bintree traversing the tree in preorder. For each internal node found, the number of leaves in its left subtree is added to the table. For each leaf node found, its color is added to the table. The final result is a sequence that contains a single symbol per node of the bintree: the number of leaves covered by each internal node and the color of each leaf node.

The major advantage of the S-tree structure is the reduction of the number of signatures which must be searched for during the evaluation of a query. In an ideal case, this number would be proportional to the height of the tree. However, this situation is very unlikely as we will explain in the section devoted to the splitting of tree pages.

FS-tree was introduced in [52, 63]. FS-tree is fuzzification of S-Tree [14, 68].

Fuzzy distance δ

$$\delta(S, S') = \gamma(S \oplus S') - \gamma(S \otimes S'), \tag{4}$$

where S and S' are signatures of the page and of the inserted document, and

$$\gamma(S) = \sum_{i=1}^{L} s_i \tag{5}$$

is the weight of the signature S, where L is the length of the signatures or instead of the weight function one of many aggregation functions could be used [20].

Normalized compression distance

NCD is a mathematical way of measuring the similarity of two documents D_x and D_y. The measuring of similarity is realized with the help of compression where repeating parts are suppressed by compression. NCD may be used for the comparison of different objects, such as images, music, texts or gene sequences. NCD has requirements to a compressor. The compressor meets the condition

$$C(D_x D_x) = C(D_x), \tag{6}$$

within logarithmic bounds [59]. NCD can be used for the detection of plagiarism and visual data extraction [7, 71]. The resulting rate of the probability distance between two documents D_x and D_y is calculated by the following formula

$$NCD\left(D_x, D_y\right) = \frac{C\left(D_x D_y\right) - min\left(C\left(D_x\right), C\left(D_y\right)\right)}{max\left(C\left(D_x\right), C(D_y)\right)}, \tag{7}$$

where

- $C(D_x)$ is the length (size) of compression of document D_x,
- $C(D_x D_y)$ is the length (size) of compression concatenation of documents D_x and D_y,
- $min(C(D_x), C(D_y))$ is the minimum of values $C\left(D_x\right)$ and $C\left(D_y\right)$,
- $max(C(D_x), C(D_y))$ is the maximum of values $C\left(D_x\right)$ and $C\left(D_y\right)$.

The NCD value is in the interval $0 \leq NCD(D_x, D_y) \leq 1 + \epsilon$. If $NCD(D_x, D_y) = 0$, then documents D_x and D_y are equal. They have the highest difference when the result value of $NCD(D_x, D_y) = 1 + \epsilon$. The constant ϵ describes the inefficiency of the used compressor. The NCD is not a metric. NCD approximates the computable Normalized Information Distance [72] with difficulty. The computation of the NCD is a very efficient way of measuring a distance between two documents because we do not need to create the output of the compression algorithm.

Fuzzy medical image retrieval

The fuzzy medical image retrieval (FMIR) method is a novel method for medical image similarity assessment and content retrieval using VQ and fuzzy signatures in conjunction with the FS-tree. The workflow of the whole method is depicted in Fig. 1. The method can be divided into the following main parts/tasks:

- Vector Quantization – in this block a codebook is created. The code book is used to convert medical images into vectors.
- Creation of Fuzzy Signatures – in this step the images vectors are converted into fuzzy vector. Each image is represented as a single fuzzy vector – signature.
- FS-tree – a data structure, which organizes fuzzy vectors – signatures in the tree form and allows to find similar fuzzy vectors to a given query.

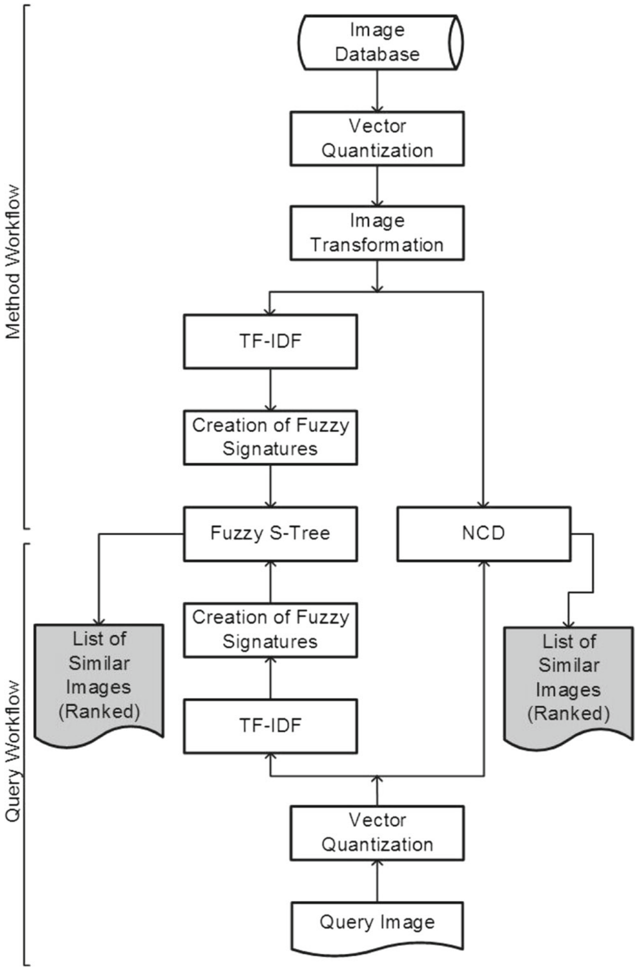


Fig. 1 The workflow diagram of FMIR

The presented FMIR method is compared with the method using NCD, with utilized Bzip2 compression algorithm [60], instead of fuzzy signatures and FS-tree (Fig. 1). In the workflow, it can be seen that for the initialization and running phase the algorithm parts are the same. Both phases share the same codebook.

For the evaluation purposes, the experiments collection of medical images is divided into two groups of images. The first group is the training group of images $\mathcal{X} = \{I_1, \dots, I_i\}$, (where I_i is a medical image), which is used for creating the codebook $\mathcal{C} = \{c_1, \dots, c_n\}$ by VQ and creating the FS-tree. The second group of images $\mathcal{Y}$, is used for the evaluation of the proposed method. The second group of images can be newly obtained images or images from different patients.

In the proposed method, an image I is transformed into the series of blocks $\mathcal{B}$ of $s \times s$ pixels using function $f : I \rightarrow \{b_1, \dots, b_m\}$, where m is the number of obtained blocks in I. Each block b of I is represented as a vector $b = (p_{11}, p_{12}, \dots, p_{1s}, p_{21}, \dots, p_{ss})$, where $p_{ij} \in< 0; 1 >$, is the grayscaled value of a pixel on i-th row and j-th column. These blocks are used in both experiments – FMIR and NCD.

As was mentioned in the previous paragraph, the $\mathcal{X}$ is transformed using the function f into a series of blocks $\mathcal{B}$. These blocks are used to create the codebook $\mathcal{C}$. The initial set of $\mathcal{B}$ is used to construct the desired $\mathcal{C}$ with the size n. To construct the $\mathcal{C}$ we utilized k-Means clustering (Algorithm 2) algorithm with n clusters [44]. The output of the Algorithm 2 are the centroids of the n clusters. These centroids represent the codes $c_1, \dots, c_n$ in the $\mathcal{C}$. We are not limited to using k-Means, we can choose many ways of how to construct the codebook.

Algorithm 2 Codebook construction algorithm

```
Data: set of image blocks $\mathcal{B}$, $n$ - codebook size
Result: codebook $\mathcal{C} = \{c_1, \dots, c_n\}$
get list of distinct blocks from $\mathcal{B}$;
sort the list of distinct blocks;
set initial positions of $n$ clusters;
while all blocks $\mathcal{B}$ are stable assigned to codes do
    assign blocks to closest code;
    update positions of codes;
end
return $\mathcal{C}$;
```

The obtained codebook $\mathcal{C}$ is used to convert database images $\mathcal{X}$ and query images $\mathcal{Y} = \{I_1, \dots, I_m\}$ into fuzzy signatures. Using the function $g(\{b_1, \dots, b_m\}, \mathcal{C}) \rightarrow \mathbb{R}^n$, where the function output is the fuzzy signature. This function maps each block b to a codec from the codebook $\mathcal{C}$ and creates the fuzzy signature. We choose the code c with the shortest distance between block b and code c. The function g calculates the membership value for each code. The size of the vector is the same. In our proposed method TF-IDF (see "TF-IDF" for more details) is utilized. The TF-IDF suppresses the membership value of blocks that occur very frequently in the images from the database. For example, it can be blocks of the I, which consists only of black color or other frequent blocks in I. This kind of block has no significant information. To calculate TF-IDF, we need occurrences of codes in the codebook, These counts are stored in vector G. The individual steps are described in Algorithm 3. We are not limited to use the TF-IDF, many other ways of how to set the membership degree of the code can be chosen, i.e. Okapi BM25 [57].

The fuzzy signatures, which are obtained from the set of images $\mathcal{X}$, are used to create the FS-tree. This tree-based data structure allows us to find similar fuzzy signatures in a short time. Similar images have similar fuzzy signatures and these fuzzy signatures are stored in close or the same nodes of the FS-tree.

The query images from $\mathcal{Y}$ are processed in the same way as images from $\mathcal{X}$. The function f converts the query image into blocks and assigns the code c to each block. Afterwards, the function g creates the fuzzy signature. As a result of a query to the FS-tree, the tree returns a list of similar medical images. The results are sorted from the shortest distance to longest. The shorter indicates the higher similarity between the query image and the resulting image.

Algorithm 3 Algorithm for creating a fuzzy signature from a image I

```
Data: Image $I$, codebook $\mathcal{C}$, Global code counts $G$
Result: Fuzzy signature $\mathcal{S}$
//code the image $I$ into vector $\mathcal{V}$
call function $\{b_1, \ldots, b_m\} = f(I, \mathcal{C})$;
// create fuzzy signature $\mathcal{S}$ from $\{b_1, \ldots, b_m\}$
call function $\mathcal{S} = g(\{b_1, \ldots, b_m\}, G)$;
return $\mathcal{S}$;
```

Experiments

The main goal of these experiments is to assess if, using the proposed method, it is possible to find medical images with a high similarity to the given query image. And the next aim is to assess the nature of the query image – if it is of a benign or malign nature. This is done using the list of similar images and the number of benign and malign nature images on the list of a predefined length. In the first step, we created a codebook $\mathcal{C}$ using VQ and k-Means clustering algorithm. The codebook was created using 1000 radiographs from the data set (500 benign and 500 cancer images) for both, CC and MLO views. The selected images belong to different patients and are of the same type. We set the size of the block to 8×8 pixels. This block size contains enough information to construct fuzzy signature that characterizes an image. In our experiment, we created a codebook $\mathcal{C}$ of size 200 codes. The similarity of fuzzy signatures was measured using fuzzy distance, for more details see "Fuzzy signatures". Some combination of t–norms and t–conorms was tested to define the fuzzy distance, so experiments taking into the account the different ways of fuzzy distance definition, are included. The discussion on whether it is possible to provide repeatability is also included.

Data description

To test our method, we first decided to use QIN Breast DCE-MRI dataset [8, 26]. This collection of breast dynamic contrast-enhanced (DCE) MRI data contains images from a longitudinal study to assess breast cancer response to neoadjuvant chemotherapy [27]. This collection of MRI images is freely available to use. The value of this collection is to provide clinical imaging data for the development and validation of quantitative imaging methods for the assessment of breast cancer response to treatment. The images from this database show the big area of the body. The usage of this type of images was limited, because parts of the image with other parts of the human body than the area of interest, influenced the results too much.

So the other database – Image Retrieval in Medical Applications (IRMA) Version of Digital Database for Screening Mammography (DDSM) LJPEG Data, with image in PNG format, which was found as very useful, was used [38]. The IRMA DDSM database consists of 9.852 radiographs (X-ray images) of breast, where normal, cancer and benign cases are included [11]. The normal cases, from our point of view, are not interesting, because they can be assessed very easily, the interesting problem is to assess if the radiographs with some problematic areas are cancerous or benign. Overall, we used 4000 radiographs.

The size of the radiographs varies from 1024 × 300 pixels to 1024 × 800 pixels. However, for our application the size of the images is not significant and the rotation of the image is also not significant. The database is composed of two types of views of the breast, of craniocaudal (CC) and mediolateral (MLO) views [12]. For our experiments, the groups were separated and the method was proven for both groups separately.

Description of used computational equipment

The experiments were run on a computer with two Intel Xeon Processor E5-2680 v2 (25M Cache, 2.80 GHz, 20 threads) CPUs installed with 768 GB of main memory. The application is capable of using the maximal number threads of both CPUs. Most parts of the suggested method are parallelized. The parallelization helps to reduce the required time which is needed to create the codebook, code input images, compute fuzzy signatures, and to evaluate the query. All necessary information is stored, the fuzzy signature tree and TF-IDF information, in the computer memory, which helps to minimize the time of the query evaluation. The most complex part of the suggested method is the codebook creation. This step has to be done only once, then the codebook is loaded and used. The query evaluation is processed very quickly.

Results

In this section, the obtained results with discussion are mentioned. To evaluate FRIM ("Fuzzy medical image retrieval") the data set described in "Data description" was used. The proposed FMIR method is compared to the method based on compression-based distance - NCD (Fig. 1). The approach with NCD takes into account the position of each block in the radiography image.

Results can be divided into two parts – the finding of similar pictures and picture classification.

Table 2 Results obtained using the proposed method for image classification for craniocaudal view (CC view)

Number of experiments		Cancer image			Benign image			Overall				
		500			500			1000				
	Number of listed similar images (length of list)	Median of images of same nature as query on list (%)	Lower-Upper quartile(IQR) (%)	Ratio of correctly classified images	Median of images of same nature as query on list (%)	Lower-Upper quartile(IQR) (%)	Ratio of correctly classified images	False negative rate	False positive rate	Sensitivity of test	Specificity of test	Accuracy of test (Ratio of correctly classified images)
Algebraic prod. & sum	10	100	70-100(30)	0.95	40	30-50(20)	0.24	0.05	0.76	0.95	0.24	0.595
	20	100	55-100(45)	0.89	45	40-60(20)	0.35	0.11	0.65	0.89	0.35	0.618
	30	93.3	50-100(50)	0.78	53	47-60(13)	0.63	0.22	0.37	0.78	0.63	0.705
	50	78	40-100(60)	0.60	62	60-64(4)	0.94	0.40	0.06	0.60	0.94	0.770
Bold prod. & sum	10	60	60-100(40)	1.00	40	40-40(0)	0.00	0.00	1.00	1.00	0.00	0.500
	20	50	50-85(35)	0.98	50	50-50(0)	0.00	0.02	1.00	0.98	0.00	0.490
	30	46.7	47-73(26)	0.42	53	53-53(0)	1.00	0.58	0.00	0.42	1.00	0.708
	50	34	34-57(23)	0.33	66	66-66(0)	1.00	0.66	0.00	0.33	1.00	0.667
Drastic prod. & sum	10	100	60-100(40)	0.80	90	70-100(30)	0.89	0.20	0.11	0.80	0.89	0.850
	20	90	50-100(50)	0.77	80	65-95(30)	0.90	0.23	0.10	0.77	0.90	0.834
	30	83.3	47-100(53)	0.74	77	63-93(30)	0.91	0.26	0.09	0.74	0.91	0.825
	50	72	44-100(56)	0.70	74	62-90(28)	0.89	0.30	0.11	0.70	0.89	0.794
Einstein prod. & sum	10	100	70-100(30)	0.98	30	30-40(10)	0.16	0.02	0.84	0.98	0.16	0.570
	20	100	55-100(45)	0.97	50	40-50(10)	0.17	0.03	0.83	0.97	0.17	0.570
	30	93.3	46-100(54)	0.74	53	50-57(7)	0.64	0.26	0.36	0.74	0.64	0.690
	50	82	36-100(64)	0.56	64	62-66(4)	0.98	0.44	0.02	0.56	0.98	0.770
Hamacher prod. & sum	10	100	70-100(30)	0.88	80	60-100(40)	0.78	0.12	0.22	0.88	0.78	0.831
	20	100	60-100(40)	0.84	80	63-90(27)	0.82	0.16	0.18	0.84	0.82	0.831
	30	90	53-100(47)	0.81	80	60-93(33)	0.84	0.19	0.16	0.81	0.84	0.825
	50	74	46-100(54)	0.70	76	62-88(26)	0.93	0.30	0.07	0.70	0.93	0.815
Minimum & Maximum	10	100	60-100(40)	0.85	90	70-100(30)	0.88	**0.15**	0.12	0.85	0.88	**0.866**
	20	85	55-100(45)	0.81	90	70-95(25)	0.91	0.19	0.09	0.81	0.91	0.860
	30	83	50-100(50)	0.76	97	70-93(23)	0.91	0.24	0.09	0.76	0.91	0.837
	50	74	44-100(56)	0.69	80	68-90(22)	0.94	0.31	0.06	0.69	0.94	0.816
NCD	10	90	60-100(40)	0.83	100	80-100(20)	0.96	**0.17**	0.04	0.83	0.96	**0.894**
	20	75	50-100(50)	0.79	90	75-100(25)	0.96	0.21	0.04	0.79	0.97	0.878
	30	70	50-100(50)	0.75	83.3	67-97(10)	0.97	0.25	0.03	0.75	0.97	0.863
	50	64	48-100(52)	0.74	80	48-100(52)	0.97	0.26	0.03	0.74	0.97	0.853

Table 3 Results obtained using the proposed method for image classification for mediolateral view (MLO view)

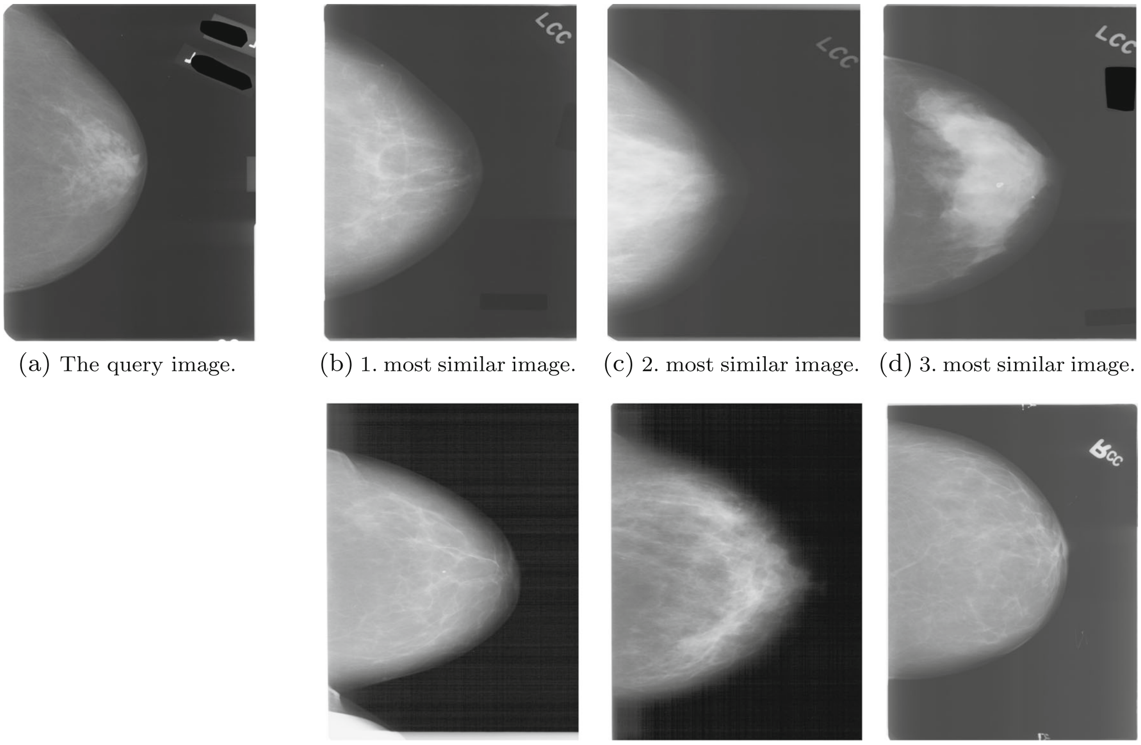
Number of experiments		Cancer image			Benign image			Overall				
		500			500			1000				
	Number of listed similar images (length of list)	Median of images of same nature as query on list (%)	Lower-Upper quartile(IQR) (%)	Ratio of correctly classified images	Median of images of same nature as query on list (%)	Lower-Upper quartile(IQR) (%)	Ratio of correctly classified images	False negative rate	False positive rate	Sensitivity of test	Specificity of test	Accuracy of test (Ratio of correctly classified images)
Algebraic prod. & sum	10	100	20-100(80)	0.64	20	20-30(10)	0.10	0.36	0.90	0.64	0.10	0.367
	20	92.5	20-100(80)	0.63	15	10-25(15)	0.10	0.37	0.90	0.63	0.10	0.363
	30	85	20-100(80)	0.60	16.7	13-23(10)	0.06	0.40	0.94	0.60	0.06	0.333
	50	70	22-100(78)	0.55	20	16-24(8)	0.03	0.45	0.97	0.55	0.03	0.291
Bold prod. & sum	10	20	20-20(0)	0.05	80	80-80(0)	1.00	0.95	0.00	0.05	1.00	0.527
	20	15	15-15(0)	0.03	85	85-85(0)	1.00	0.97	0.00	0.03	1.00	0.514
	30	20	20-20(0)	0.01	80	80-80(0)	1.00	0.99	0.00	0.01	1.00	0.506
	50	22	22-22(0)	0.01	78	78-78(0)	1.00	0.99	0.00	0.01	1.00	0.504
Drastic prod. & sum	10	90	60-100(40)	0.87	80	50-100(50)	0.74	0.13	0.26	0.87	0.74	0.804
	20	85	55-100(45)	0.86	75	55-95(40)	0.76	0.14	0.24	0.86	0.76	0.806
	30	76.7	57-100(43)	0.82	70	53-93(40)	0.77	0.18	0.23	0.82	0.77	0.795
	50	70	54-100(46)	0.81	66	54-90(36)	0.80	0.19	0.20	0.81	0.80	0.802
Einstein prod. & sum	10	100	70-100(30)	0.59	80	80-80(0)	0.95	0.41	0.05	0.59	0.95	0.769
	20	100	55-100(45)	0.54	85	80-85(5)	0.97	0.46	0.03	0.54	0.97	0.754
	30	93.3	46-100(54)	0.51	80	80-83(3)	0.98	0.49	0.02	0.51	0.98	0.750
	50	82	36-100(64)	0.50	80	80-82(2)	0.99	0.50	0.01	0.50	0.99	0.750
Hamacher prod. & sum	10	100	70-100(30)	0.83	90	70-100(30)	0.85	0.17	0.15	0.83	0.85	0.840
	20	100	60-100(40)	0.75	80	65-90(25)	0.87	0.25	0.13	0.75	0.87	0.813
	30	90	53-100(47)	0.69	76.7	63-87(24)	0.87	0.31	0.13	0.69	0.87	0.785
	50*	74	46-100(54)	0.62	72	62-84(22)	0.89	0.38	0.11	0.62	0.89	0.759
Minimum & Maximum	10	100	60-100(40)	0.86	90	70-100(30)	0.85	**0.14**	0.15	0.86	0.85	**0.855**
	20	85	55-100(45)	0.80	80	65-90(25)	0.88	0.20	0.12	0.80	0.88	0.841
	30	83.3	50-100(50)	0.76	76.7	63-90(27)	0.89	0.24	0.11	0.76	0.89	0.828
	50*	74	44-100(56)	0.68	72	62-84(22)	0.92	0.32	0.08	0.68	0.92	0.801
NCD	10	80	50-100(50)	0.80	100	80-100(20)	0.97	**0.20**	0.03	0.80	0.97	**0.886**
	20	70	45-100(55)	0.75	90	80-95(15)	0.97	0.25	0.03	0.75	0.97	0.860
	30	66.7	47-100(53)	0.75	86.7	73-93(20)	0.97	0.25	0.03	0.75	0.97	0.861
	50	62	48-100(52)	0.73	80	68-88(20)	0.97	0.27	0.03	0.73	0.97	0.835

Within the task of the finding of similar pictures, three picture sets are presented for illustration. The first two using FMIR – variant with Bold product & sum and Minimum & Maximum used for defining the fuzzy distance, the third one using the NCD. All images are rotated to the same side, because it could be hard for readers to assess the similarity of differently rotated images.

The assessment of classification efficiency is done for 2000 images – 1000 of the craniocaudal (CC) view (500 benign and 500 malign) images and 1000 of the mediolateral (MLO) view (500 benign and 500 malign) images, to be able to assess the repeatability. Experiments are performed in this way: first of all the query image is chosen (randomly). For the query image, the most similar pictures are found and presented as pictures. Then the classification, which is used to find the nature (malignity or benignity) of the query image was done. For every image in query, the list of the most similar images is written. The list is cut to 10, 20, 30 and 50 of the most similar images. Using this, we want to find the appropriate cut off of the list to be able to determine the membership in one of the classes of the query image - whether it is the malign or benign class. For every cut, the amount of the pictures of the benign and malign class are counted. The class (nature), which is bigger is significant for the assessment of the query image nature. If the ration of benign and malign pictures on the list is equal, the image is assessed as malign, because in breast cancer, it is more dangerous to determine an ill patient as a healthy patient, it means, that the smaller the false negativity is, the better. The fuzzy distance was defined in six ways, so experiments were repeated six times. All combinations of t–norms and t–conorms, which were used for the definition of the fuzzy distance, can be seen in Table 1. Only the combination listed in rows were used, but many other combinations can be defined and used.

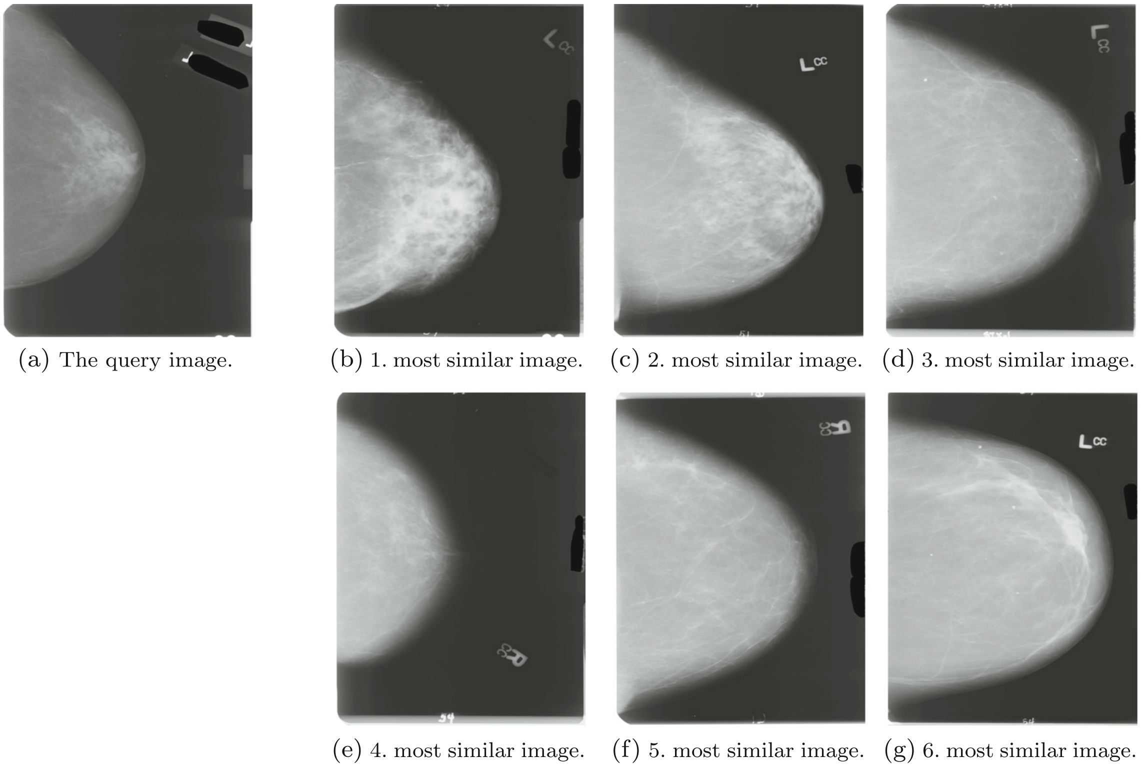
For the classification part, one of the main tasks is to declare the repeatability for every variant of fuzzy distance definition and for every image class (benign and malign) 500 experiments were done, followed by the performing of the statistical assessment. As it is necessary to be able to confidently identify both image classes – benign and malign, the results are presented separately for benign and malign to be able easily assess if there are any differences in the classification of the images of both classes.

The classification results are depicted in Table 2 for the CC view and in Table 3 for the MLO view. As the proposed method is applied in the medical area, standard test parameters such as false positive rate, false negative rate, sensitivity, specificity and accuracy of tests are calculated [35, 77]. All data (except data indicated with *) comes from non-normal distribution, so the basic statistics were done


(a) The query image. (b) 1. most similar image. (c) 2. most similar image. (d) 3. most similar image.

(e) 4. most similar image. (f) 5. most similar image. (g) 6. most similar image.

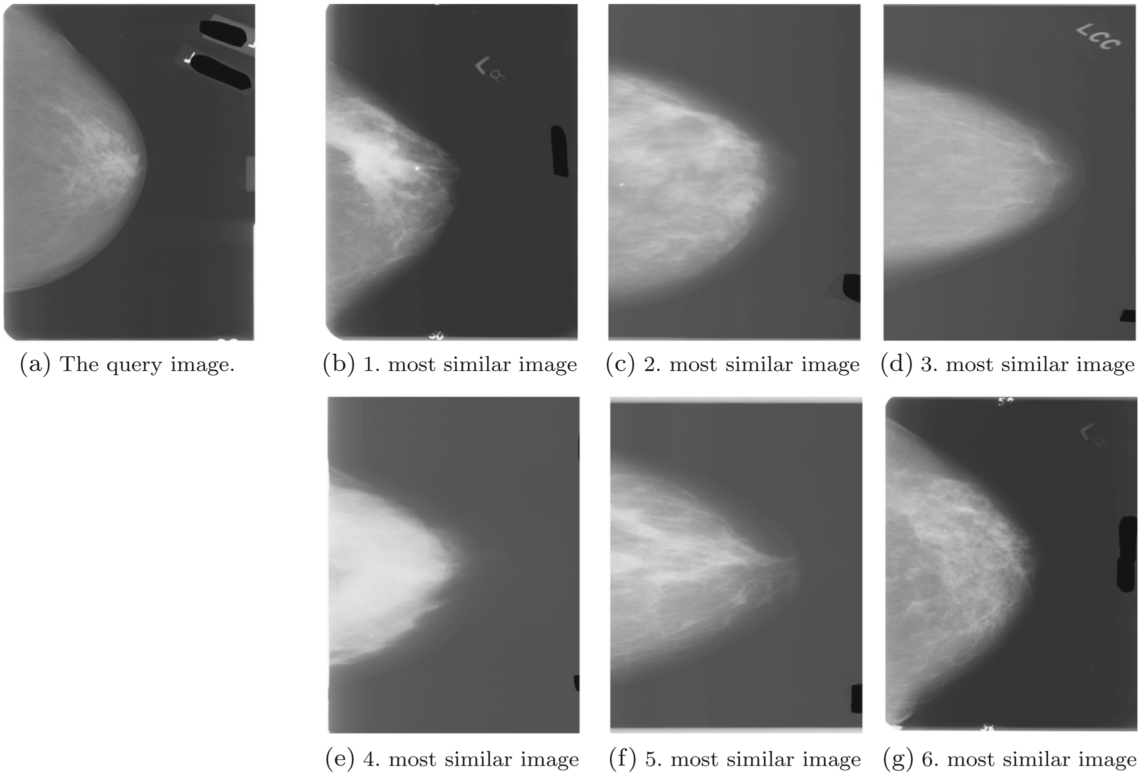
Fig. 2 Example of results using FMIR, where fuzzy distance is defined using Bold (Lukasiewicz, Bounded) product & sum for CC view



(a) The query image. (b) 1. most similar image. (c) 2. most similar image. (d) 3. most similar image.

(e) 4. most similar image. (f) 5. most similar image. (g) 6. most similar image.

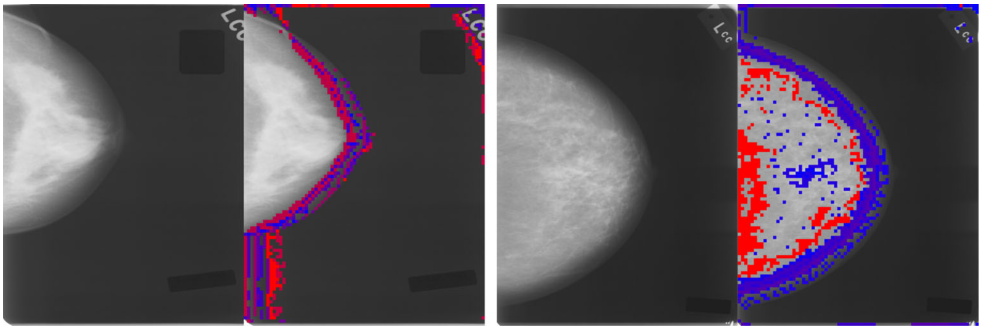
Fig. 3 Example of results using FMIR, where fuzzy distance is defined using Minimum & Maximum for CC view



(a) The query image. (b) 1. most similar image (c) 2. most similar image (d) 3. most similar image

(e) 4. most similar image (f) 5. most similar image (g) 6. most similar image

Fig. 4 Example of results using NCD for CC view



(a) A benign radiograph.

(b) A cancer radiograph.

Fig. 5 Example of the area of the interest in a benign and cancer radiographs for craniocaudal view

using the median, lower and upper quartile and interquartile range (IQR) [22] for all experiments.

Results obtained by FMIR

For the illustrative example, the query image was randomly chosen, then the list of similar images was created. In the Figs. 2 and 3 the same query image with the six most similar images are depicted (six images with the smallest value of the fuzzy distance) for the FMIR variant using the CC view in comparison to method with NCD (Fig. 4). The resulting images are sorted by the fuzzy distance. All resulting images belong to different patients. According to the results in Table 2 the experiment for the variant of FMIR with the best and with the worst accuracy were chosen.

The next part is dedicated to image classification – nature determination. The efficiency of FMIR in six variants was assessed, where fuzzy distance is defined in six ways and for the CC and MLO view separately. For the classification of radiographs with the CC view, the FMIR in the variant with Minimum & Maximum was found as the best one. And the same for the MLO view. The variant of FMIR with fuzzy distance defined with Minimum & Maximum shows the greatest accuracy with good false negativity. In the comparison with NCD, NCD gives better results with little difference between all cuts. But FMIR has a smaller false negativity in the variant with Minimum & Maximum, significantly (P-value=0.00578) for the MLO view in the 10 cut than NCD. Generally, 10 cuts give a better result, because the most similar images, should be on the list at the top.

Due to the usage of TF—IDF, FMIR can determine the areas of interest in the image. The Fig. 5 illustrates the position of the most important blocks, with the highest value of TF-IDF, in a radiograph. The top 20 most important blocks were chosen. In the Fig. 5a a benign image can be seen. The selected blocks occur on the edge of the breast. The situation is different on a cancer radiograph. In the cancer radiographs, Fig. 5b, the most important parts are located inside the breast. The meaning of the colors is as follows: the red colored blocks have the highest TF-IDF values and the blue colored block is to the 20-th highest TF-IDF value. The method with NCD does not provide this feature.

Results Obtained by NCD

In this section, results obtained using a simpler method with NCD are presented. The process is similar, but with one difference. Instead of the FS-tree, we measure the similarity between two images using NCD (“Normalized compression distance”) with the utilized Bzip2 compression algorithm. Again, it was chosen randomly the query image and similar images were found. In Fig. 4 can be seen of the six most similar images to the same query image as in the cases with FMIR. The FMIR variant brought better results in both presented variants.

Classification results were presented and compared with the results of FMIR in 9.1.

Conclusion

In the article, the novel method fuzzy medical image retrieval (FMIR) for image retrieval and classification was presented. FMIR is based on vector quantization, fuzzy signatures and fuzzy S-trees. The proposed method was compared in six variants (six variants of fuzzy distance definition) with the method using NCD. It was demonstrated, that FMIR is useful for finding similar images and classification – determination of the nature of query image, if it is malign or benign. In comparison with NCD in classification, it gives us comparable, but a little bit worse results. But FMIR brings significantly smaller false negativity in variant with Minimum & Maximum in fuzzy distance definition, which is very dangerous in breast cancer. Moreover, FMIR has a useful feature, it is able to determine the areas of interest in the picture. FMIR use for finding the area of interest and for the finding of similar images and the method with NCD for classification.

The proposed methods are going to be parts of the complex decision support system to help determine appropriate health care according to the experiences obtained from similar, previous cases. The method was verified on medical data concretely on the breast cancer radiographs (X-ray images). According to the results of the experiments, it was found, that from the clinical point of view, the methods are more useful for finding pathological structures in the part of the body, which are more homogenous in the population such as the liver or other internal organs, where the influence of body shape, or other parts of human body is small. The usage of the proposed method is dependent upon the existence of a large dataset of images, the next work is going to be focused on the enabling and simplification of work with such a huge collection.

Acknowledgments This work was supported by the Czech Science Foundation under grant No. GACR GJ16-25694Y and in part by the grant of SGS No. SP2016/68 VŠB - Technical University of Ostrava, Czech Republic. X-ray images were used with courtesy of TM Deserno, Dept. of Medical Informatics, RWTH Aachen, Germany.

References

1. Arya, S., and Mount, D.M., Algorithms for fast vector quantization. In: Data Compression Conference, 1993. DCC '93, pp. 381–390, 1993.
2. Benedetto, D., Caglioti, E., and Loreto, V., Language trees and zipping. *Physical Review Letters* 88:048702-1–048702-4, 2002.
3. Berek, P., Prílepok, M., Platos, J., and Snášel, V., Classification of EEG signals using vector quantization. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). 8468 LNAI (PART 2). pp. 107–118, 2014.
4. Chen, Y.J., and Chen, Y.B., On the signature tree construction and analysis. In: IEEE Transactions on knowledge and data engineering. vol. 18(9), pp. 1207–1224, 2006.
5. Chung, K.L., and Wu, C.J., A fast search algorithm on modified S-trees. In: Pattern recognition letters. vol. 16(11), pp. 1159–1164, 1995.
6. Chung, K.L., Wu, J.G., and Lan, J.K., Efficient search algorithm on compact S-trees. In: Pattern recognition letters vol. 18(14), pp. 1427–1434, 1997.
7. Cilibrasi, R., and Vitányi, P. M. B., Clustering by compression. In: IEEE Transactions on information theory. vol. 51(4), pp. 1523–1545, 2005.
8. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., and Prior, F., The cancer imaging archive (TCIA): Maintaining and operating a public information repository. In: Journal of digital imaging. vol. 26(6), pp. 1045-1057, 2013.
9. Cosman, P.C., Gray, R.M., and Vetterli, M., Vector quantization of image subbands: a survey. In: IEEE Transactions on image processing. vol. 5(2), pp. 202–225, 1996.
10. Cosman, P.C., Oehler, K.L., Riskin, E.A., and Gray, R.M., Using vector quantization for image processing. In: Proceedings of the IEEE. vol. 81,(9), pp. 1326–1341, 1993.
11. De Oliveira, J.E., Deserno, T.M., and Araujo A.D.A., (2008) Breast lesion classification applied to a reference database. In: Proceedings of the 2nd international conference on e-medical systems, Sfax, Tunisia. pp. 29–31.
12. De Oliveira, J.E., Machado, A.M., Chavez, G.C., Lopes, A.P.B., Deserno, T.M., and Araujo, A.D.A., Mammosys: A content-based image retrieval system using breast density patterns. In: Computer methods and programs in biomedicine. vol. 99(3), pp. 289–297, 2010.
13. Depeursinge, A., Duc, S., Eggel, I., and Muller, H., Mobile medical visual information retrieval (Review). In: IEEE Transactions on information technology in biomedicine. vol. 16(1), pp. 53–61, 2012.
14. Deppisch, U., S-tree: A Dynamic Balanced Signature Index for Office Retrieval. In: Proceedings of ACM research and development in information retrieval, pisa, Italy. Sept. 8-10. pp. 77–87, 1986.
15. Deselaers, T., Keysers, D., and Ney, H., Features for image retrieval: an experimental comparison. In: Information Retrieval. vol. 11(2). pp. 77–107, 2008.
16. Dobrinkat, M., Vayrynen, J., Tapiovaara, T., and Kettunen, K., Normalized Compression Distance Based Measures for MetricsMART 2010. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, WMT '10. pp. 343–348, 2010.
17. Dubnov, S., Assayag, G., Lartillot, O., and Bejerano, G., Using machine-learning methods for musical style modeling. In: IEEE Computer society. vol. 36(10), pp. 73–80, 2003.
18. Faloutsos, C., Signature files. In: Information retrieval: Data structures & algorithms, W.B. Frakes and r. Baeza-Yates, eds. Prentice Hall, New Jersey, pp. 44–65, 1992.
19. Goguen, J.A., L-fuzzy sets. In: Journal of mathematical analysis and applications. vol. 18(1), pp. 145–174, 1967.
20. Grabisch, M., Marichal, J.L., Mesiar, R., and Pap, E., Aggregation functions cambridge univ, Press,Cambridge, 2009.
21. Granados, A., Analysis and study on text representation to improve the accuracy of the normalized compression distance. In: AI Communications. vol. 25(4), pp. 381–384, 2012.
22. Gupta, B.C., and Guttman, I. *Statistics and Probability with Applications for Engineers and Scientists*. New Jersey: Wiley, 2013.
23. Guttman, A., R-trees a dynamic index structure for spatial searching. In: Proceedings ACM SIGMOD international conference on management of data. vol. 14(2), pp. 47–57, 1984.
24. Hill, T., and Lewicki, P. *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. Tulsa: StatSoft, Inc., 2006.
25. Huang, C.M., and Harris, R.W., A comparison of several vector quantization codebook generation approaches. In: IEEE Trans image process. vol. 2(1), pp. 108–12, 1993.
26. Huang, W., Li, X., Chen, Y., Li, X., Chang, M. C., Oborski, M. J., Malyarenko, D. I., Muzi, M., Jajamovich, G. H., Fedorov, A., Tudorica, A., Gupta, S. N., Laymon, C. M., Marro, K. I., Dyvorne, H. A., Miller, J. V., Barbodiak, D. P., Chenevert, T. L., Yankeelov, T. E., Mountz, J. M., Kinahan, P. E., Kikinis, R., Taouli, B., Fennessy, F., and Kalpathy-Cramer, J., Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge. *The Cancer Imaging Archive*, 2014. doi:10.7937/K9/TCIA.2014.A2N1IXOX.
27. Huang, W., Li, X., Chen, Y., Li, X., Chang, M.C., Oborski, M.J., Malyarenko, D.I., Muzi, M., Jajamovich, G.H., Fedorov, A., Tudorica, A., Gupta, S.N., Laymon, C.M., Marro, K.I., Dyvorne, H.A., Miller, J.V., Barbodiak, D.P., Chenevert, T.L., Yankeelov, T.E., Mountz, J.M., Kinahan, P.E., Kikinis, R., Taouli, B., Fennessy, F., and Kalpathy-Cramer, J., Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis

challenge. In: Translational oncology. vol. 7(1), pp. 153–166, 2014.

28. Ihnat, P., Gunkova, P., Peteja, M., Vavra, P., Pelikan, A., and Zonca, P., Diverting ileostomy in laparoscopic rectal cancer surgery: high price of protection. In: Surgical Endoscopy. pp. 1–8, 2016. doi:10.1007/s00464-016-4811-3.
29. Ihnat, P., Vavra, P., and Zonca, P., Treatment strategies for colorectal carcinoma with synchronous liver metastases: Which way to go? In: World journal of gastroenterology. vol. 21(22), pp. 7014–7021, 2015. doi:10.3748/wjg.v21.i22.7014.
30. Kent, A.J., Sacks-Davis, R., and Ramamohanarao, K., A signature file scheme based on multiple organizations for indexing very large text databases. In: Journal of the american society for information science. vol. 41(7), pp. 508–534, 1990.
31. Klir, G.J., S.t. Clair, U.H., and Yuan, B., Fuzzy set theory: foundations and applications. Prentice-Hall Inc., Upper Saddle River, NJ, 1997.
32. Koczy, L. T., Vector valued fuzzy sets. In: BUSEFAL-BULL STUD EXCH FUZZIN APPL. pp. 41–57, 1980.
33. Koczy, L.T., Vamos, T., and Biro, G., Fuzzy signatures. In: Proceedings of the 4th meeting of the euro working group on fuzzy sets and the 2nd international conference on soft and intelligent computing (EUROPUSE-SIC 1999), Budapest, Hungary. pp. 210–217, 1999.
34. Kratky, M., Snášel, V., Pokorny, J., and Zezula, P., Efficient processing of narrow range queries in multi-dimensional data structures. In: Proceedings of the International Database Engineering and Applications Symposium, IDEAS 2006. pp. 69–79, 2006.
35. Lalkhen, A.G., and McCluskez, A., Storage and Retrieval: Signature File Access. Clinical tests: sensitivity and specificity. In: Continuing education in anaesthesia, critical care & pain. vol. 8(6), pp. 221–223, 2008.
36. Le, T.M., and Van, T.T., Clustering binary signature applied in Content-Based image retrieval. In: New advances in information systems and technologies, advances in intelligent systems and computing. vol. 444, pp. 233–242, 2016.
37. Le, T.M., and Van, T.T., Image retrieval system based on EMD similarity measure and S-Tree. In: Intelligent technologies and engineering systems, lecture notes in electrical engineering. vol. 234, pp. 139–146, 2013.
38. Lehmann, T. M., Oliveira, J.E.E., Güld, M.O., and Welter, P., IRMA Version of DDSM LJPEG Data, 2010. https://ganymed.imib.rwth-aachen.de/irma/datasets_en.php?SELECTED=00010#00010.dataset.
39. Lempel, A., and Ziv, J., On the complexity of finite sequences. In: IEEE Transactions on information theory. vol. 22(1), pp. 75–81, 1976.
40. Leskovec, J., Rajaraman, A., and Ullman, J.D. *Data mining of massive datasets*. Cambridge: Cambridge University Press, 2014.
41. Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., and Zhang, H., An information-based sequence distance and its application to whole mitochondrial genome phylogeny. In: Bioinformatics. vol. 17(2), pp. 149–154, 2001. doi:10.1093/bioinformatics/17.2.149.
42. Li, M., Chen, X., Li, X., Ma, B., and Vitányi, P.M.B., The similarity metric. In: IEEE Transactions on information theory, vol. 50(12), pp. 3250–3264, 2002.
43. Lu, G., and Teng, S., A novel image retrieval technique based on vector quantization. In: Proceedings of International Conference on Computational Intelligence for Modelling, Control and Automation. pp. 36–41, 1999.
44. MacKay, D., An example inference task: Clustering. In: Information theory, inference and learning algorithms. Cambridge University Press, Cambridge. pp. 284–292, 2003.
45. Malik, F., and Baharudin, B.B., Feature analysis of quantized histogram color features for Content-Based image retrieval based on laplacian filter. In: Proceedings of the International Conference on System Engineering and Modeling. vol. 34, pp. 44–49, 2012.
46. Mendis, B.S.U., Gedeon, T.D., and Koczy, L.T., Investigation of aggregation in fuzzy signatures. In: Proceedings of 3rd international conference on computational intelligence, Robotics and Autonomous Systems, Singapore. vol. 411, 2005.
47. Nardelli, E., and Proietti, G., S^*-Tree: An Improved S^+-Tree for Coloured Images. In: Proceedings of the ADBIS'99, Springer Verlag. pp. 156–167, 1999.
48. Nascimento, M.A., Tousidou, E., Chitkara, V., and Manolopoulos, Y., Image indexing and retrieval using signature trees. In: Data & knowledge engineering. vol. 43(1), pp. 57–77, 2002.
49. Niblack, C. W., Barber, R., Equitz, W., Flickner, M., Glasman, E. H., Petkovic, D., Yanker, P., Faloutsos, C., and Taubin, G., The QBIC project: Querying images by content, using color, texture, and shape. In: Storage and Retrieval for Image and Video Databases (SPIE). pp. 173–187, 1993. doi:10.1117/12.143648.
50. Ogiela, L., Cognitive informatics in image semantics description, identification and automatic pattern understanding. In: Neurocomputing. vol. 122, pp. 58–69, 2013. doi:10.1016/j.neucom.2013.06.001.
51. Park, K., Hybrid Image Compression by Using Vector Quantization (VQ) and Vector-Embedded karhunen-loève Transform (VEKLT). In: Data compression conference (DCC), 2015, pp. 466, 2015.
52. Platos, J., Kromer, P., Snášel, V., and Abraham, A., Searching similar images - Vector quantization with S-tree. In: IEEE CASoN, pp. 384–388, 2012.
53. Pozna, C., Minculete, N., Precup, R.E., Koczy, L.T., and Ballagi, A., Signatures: definitions, operators and applications to fuzzy modelling. In: Fuzzy sets and systems. vol. 201, pp. 86–104, 2012.
54. Prílepok, M., Berek, P., Platos, J., and Snášel, V., Spam Detection using Data Compression and Signatures. In: Cybernetics and systems. vol. 44(6–7), pp. 533–549, 2013.
55. Rahman, M. M., Antani, S. K., and Thoma, G. R., Biomedical image retrieval in a fuzzy feature space with affine region detection and vector quantization of a scale-invariant descriptor. In: Proceedings of the 6th international conference on Advances in visual computing. pp. 261–270, 2010.
56. Ramsak, F., Markl, V., Fenk, R., Zirkel, M., Elhardt, K., and Bayer, R., Integrating the UB-tree Into a Database System Kernel. In: Proceedings of the 26th international conference on very large databases, cairo, Egypt vol. 2000, pp. 263–272, 2000.
57. Robertson, S., Walker, S., Beaulieu, M. M., and Gatford, M., Okapi at TREC-4. In: Proceedings of the Fourth Text Retrieval Conference. pp. 73–96, 1995.
58. Schaefer, G., Compressed domain image retrieval by comparing vector quantization codebooks. In: Proceedings of the visual communications and image processing 2002. vol. 4671, pp. 959–966, 2002.
59. Sculley, D., and Brodley, C.E., Compression and machine learning: A new perspective on feature space vectors. In: Proceedings of the Data Compression Conference. pp. 332–332, 2006.
60. Seward, J.: Bzip2 compression algorithm, http://www.bzip.org/, 2010.
61. Shannon, C.E., Coding theorems for a discrete source with a fidelity criterion. In: IRE Nat. Conv. Rec. vol. 4, pp. 142–163, 1959.

62. Sharma, N.S., Rawat, P.S., and Singh, J.S., Efficient CBIR using color histogram processing. *Signal & Image Processing: An International Journal* 2(1):94–112, 2011.
63. Snášel, V., Fuzzy Signatures for Multimedia Databases. In: Proceedings of the First International Conference on Advances in Information Systems. pp. 257–264, 2000.
64. Snášel, V., Horak, Z., Kudelka, M., and Abraham, A., Fuzzy signatures organized using S-Tree. In: Proceedings of the Systems, Man, and Cybernetics (SMC), 2011 IEEE. pp. 63–67, 2011.
65. Swain, M., and Ballard, D., Color indexing. In: International journal of computer vision. vol. 7(1), pp. 11–32, 1991. doi:10.1007/BF00130487.
66. Tanaka, T., and Yamashita, Y., Image coding using vector-embedded karhunen-loève transform. In: Proceedings of international conference on the image processing. vol. 1, pp. 482–486, 1999. doi:10.1109/ICIP.1999.821656.
67. Teng, S.W., and Lu, G., Image indexing and retrieval based on vector quantization. In: Pattern recognition. vol. 40(11), pp. 3299–3316, 2007.
68. Tousidou, E., Nanopoulos, A., and Manolopoulos, Y., Improved methods for signature-tree construction. In: The computer journal. vol. 43(4), pp. 301–314, 2000.
69. Vamos, T., Koczy, L.T., and Biro, G., Fuzzy signatures in datamining. In: Proceedings of the joint 9th IFSA world congress and 20th NAFIPS international conference, vancouver, BC, Canada. vol. 5, pp. 2842–2846, 2001.
70. Vavra, P., Nowaková, J., Ostruszka, P., Hasal, M., Jurcikova, J., Martinek, L., Penhaker, M., Ihnat, P., Habib, N., and Zonca, P., Colorectal cancer liver metastases: laparoscopic and open radiofrequency-assisted surgery. In: Videosurgery miniinv vol. 10(2), pp. 205–212, 2016.
71. Vitányi, P.M.B., Universal similarity. In: Proceedings of the IEEE Information Theory Workshop. pp. 238–243, 2005.
72. Vitányi, P.M.B., Balbach, F. J., Cilibrasi, R., and Li, M., Normalized Information Distance. In: Information theory and statistical learning, Springer US. pp. 45–82, 2008.
73. Wong, K.W., Gedeon, T.D., and Koczy, L. T., Construction of fuzzy signature from data: an example of SARS pre-clinical diagnosis system. In: Proceedings of the IEEE international conference on fuzzy systems (FUZZ-IEEE 2004), Budapest, Hungary pp. 1649–1654, 2004.
74. Yasmin, M., Mohsin, S., and Sharif, M., Intelligent image retrieval techniques: a survey. In: Journal of applied research and technology. vol. 12(1), pp. 87–103, 2014.
75. Zadeh, L.A., Fuzzy sets. In: Information and control. vol. 8(3), pp. 338–353, 1965. doi:10.1016/S0019-9958(65)90241-X.
76. Zezula, P., and Tiberio, P., Storage and retrieval: Signature file access. In: Encyclopedia of microcomputers. vol. 16. Marcel Dekker, Inc.,New York. pp. 377–403, 1995.
77. Zhu, W., Zeng, N., and Wang, N.: Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS®implementation, 2010.