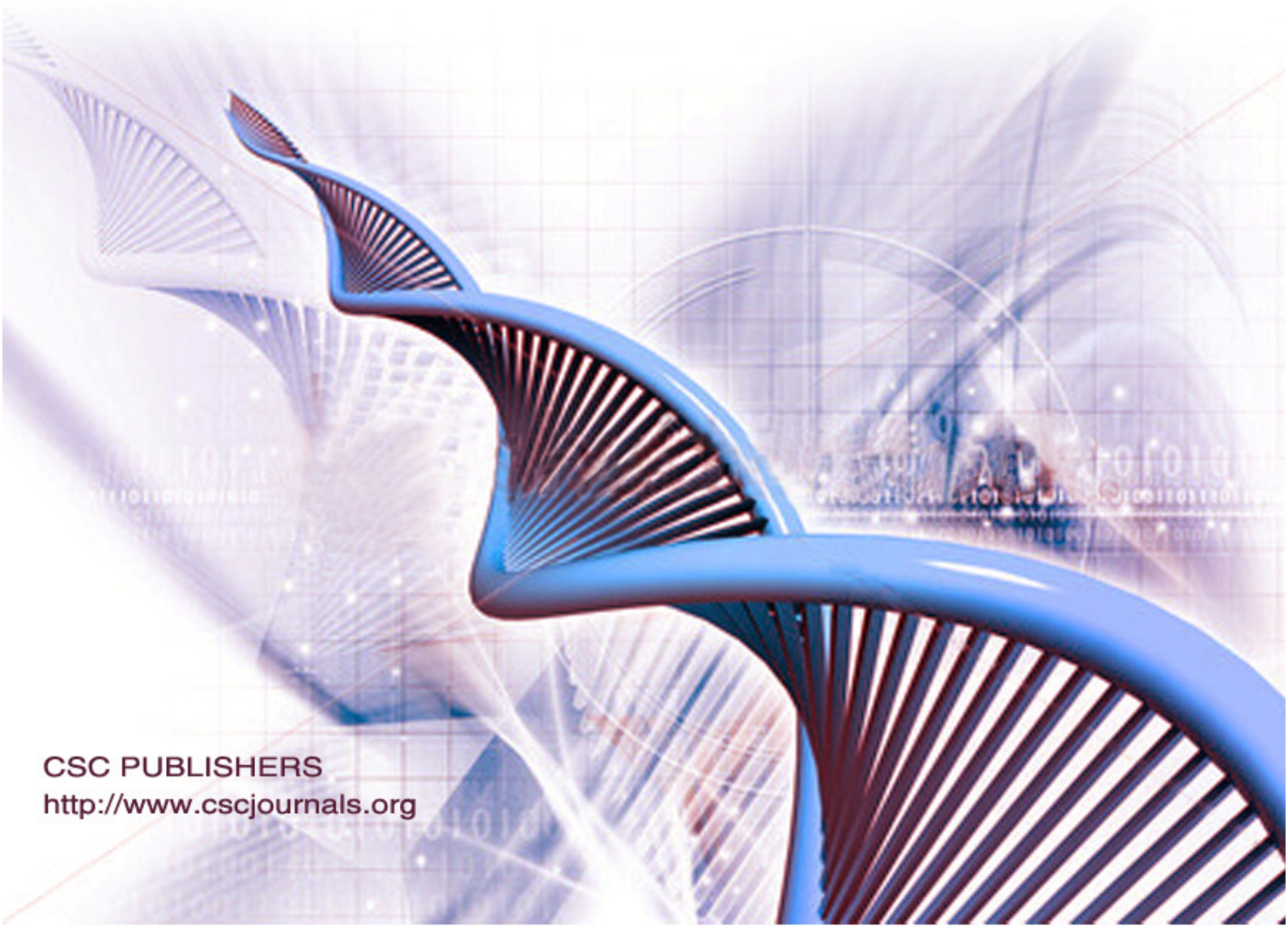
Volume 5 ▪ Issue 4 ▪ October 2011

INTERNATIONAL JOURNAL OF

BIOMETRICS AND BIOINFORMATICS (IJBB)

ISSN : 1985-2347
Publication Frequency: 6 Issues / Year



CSC PUBLISHERS
http://www.cscjournals.org

INTERNATIONAL JOURNAL OF BIOMETRICS AND BIOINFORMATICS (IJBB)

VOLUME 5, ISSUE 4, 2011

EDITED BY
DR. NABEEL TAHIR

ISSN (Online): 1985-2347

International Journal of Biometrics and Bioinformatics (IJBB) is published both in traditional paper form and in Internet. This journal is published at the website http://www.cscjournals.org, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJBB Journal is a part of CSC Publishers

Computer Science Journals

http://www.cscjournals.org

INTERNATIONAL JOURNAL OF BIOMETRICS AND BIOINFORMATICS (IJBB)

Book: Volume 5, Issue 4, October 2011
Publishing Date: 05-10-2011
ISSN (Online): 1985-2347

IJBB Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers, 2011

EDITORIAL PREFACE

This is the third issue of volume five of International Journal of Biometric and Bioinformatics (IJBB). The Journal is published bi-monthly, with papers being peer reviewed to high international standards. The International Journal of Biometric and Bioinformatics is not limited to a specific aspect of Biology but it is devoted to the publication of high quality papers on all division of Bio in general. IJBB intends to disseminate knowledge in the various disciplines of the Biometric field from theoretical, practical and analytical research to physical implications and theoretical or quantitative discussion intended for academic and industrial progress. In order to position IJBB as one of the good journal on Bio-sciences, a group of highly valuable scholars are serving on the editorial board. The International Editorial Board ensures that significant developments in Biometrics from around the world are reflected in the Journal. Some important topics covers by journal are Bio-grid, biomedical image processing (fusion), Computational structural biology, Molecular sequence analysis, Genetic algorithms etc.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 5, 2011, IJBB appears in more focused issues. Besides normal publications, IJBB intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

The coverage of the journal includes all new theoretical and experimental findings in the fields of Biometrics which enhance the knowledge of scientist, industrials, researchers and all those persons who are coupled with Bioscience field. IJBB objective is to publish articles that are not only technically proficient but also contains information and ideas of fresh interest for International readership. IJBB aims to handle submissions courteously and promptly. IJBB objectives are to promote and extend the use of all methods in the principal disciplines of Bioscience.

IJBB editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJBB provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Biometric and Bioinformatics (IJBB)

EDITORIAL BOARD

Dr Antonio Marco
University of Manchester
United Kingdom

Dr Peng Jiang
University of Iowa
United States of America

TABLE OF CONTENTS

Volume 5, Issue 4, October 2011

Pages

Indexing for Large DNA Database Sequences

Samer Mahmoud Wohoush *samer_wh@yahoo.com*
Faculty/Department/Division
Palestine Polytechnic University
Hebron, PO.Box 198, Palestine

Mahmoud Hasan Saheb *alsaheb@ppu.edu*
Palestine Polytechnic University
Hebron, PO.Box 198, Palestine

Abstract

Bioinformatics data consists of a huge amount of information due to the large number of sequences, the very high sequences lengths and the daily new additions. This data need to be efficiently accessed for many needs. What makes one DNA data item distinct from another is its DNA sequence. DNA sequence consists of a combination of four characters which are A, C, G, T and have different lengths. Use a suitable representation of DNA sequences, and a suitable index structure to hold this representation at main memory will lead to have efficient processing by accessing the DNA sequences through indexing, and will reduce number of disk I/O accesses. I/O operations needed at the end, to avoid false hits, we reduce the number of candidate DNA sequences that need to be checked by pruning, so no need to search the whole database. We need to have a suitable index for searching DNA sequences efficiently, with suitable index size and searching time. The suitable selection of relation fields, where index is build upon has a big effect on index size and search time. Our experiments use the n-gram wavelet transformation upon one field and multi-fields index structure under the relational DBMS environment. Results show the need to consider index size and search time while using indexing carefully. Increasing window size decreases the amount of I/O reference. The use of a single field and multiple fields indexing is highly affected by window size value. Increasing window size value lead to better searching time with special type index using single filed indexing. While the search time is almost good and the same with most index types when using multiple field indexing. Storage space needed for RDMS indexing types are almost the same or greater than the actual data.

Keywords: Large Database, DNA Sequence, Index Structure, Sequence Transformation, Wavelet Transformation, RDMS Indexing.

1. INTRODUCTION

Dealing with string of characters for large database is not easy in term of space and access time. Genome databases as NCBI have a huge size because of the daily addition of new data. Electronic books and biological data are good examples for large databases that include text and sequences. For genome database we can consider DNA sequence as a key value that distinguishes a sequence from another.

Most of the work on genome database tries to find small size, efficient digit value that can represents DNA sequence. The problem of large number of I/O operation when accessing large size database is very costly in term of space and performance. Accessing the database need to be in minimum amount and at last stage after filtrations to reduce the number of records need to be accessed.

We will consider the DNA sequence as the key value for genome database. Transforming this key to a digit is required to increase efficiency. Wavelet transformation technique [1] for DNA sequences is a suitable choice for our needs to do transformation as it gives us two advantages.

Firstly, it saves sequence order while considering amount of overlapping carefully. Secondly, transforming characters to digits depends on frequencies of characters.

Little amount of storage is needed as after finding first level wavelet coefficients[1], the second level can be calculated depending on the first level instead of referring to the original sequence again. Our evaluation use substring searching for matching identical pattern by sliding window, this will reduce candidate list of sequences need to be checked, and at the last stage refer to database disk for validations after pruning, in other words, optimization for the number of I/O operations.

Relational database provide different types of indexing like BTree, RTree[2], and hashing. Using these types of indexing to store Wavelet transformation will be discussed later.
The rest of the paper is organized as follows: section 2 reviews of related works. Section 3 presents data samples and methodology. Our results will be presented in section 4. Section 5 discusses the results. Conclusions will be discussed in section 6.

2. RELATED WORK

Different methods had been used to transform and index huge database systems. Dynamic programming [3, 4] has time and space complexity of O(nm) for two strings S and Q of lengths n and m, for database comparisons it will needs matrix of size nxm. Hence for long sequence and large database this method will be not practical in term of both space and time. It finds the difference between S and Q using a heavy computation method; the edit distance.
The use of r Binary masks [3] of size n, $M_1,M_2,\ldots,M_r$ to move through S, of size m, by word size of w has complexity of O(nmr/w). For large value of m, this complexity will be very close to dynamic programming.

Dictionary based indexing [3] for a database of sequences S_i (i:1,2,…,n), creates index structure of size n corresponding to database size, predefining query lower bound length (L) to be equal to *log*(n) assumed. Query with larger length will be partitioned into smaller parts. All substrings of length L mapped to integers using hashing function and for queries larger than L split it into sub-queries, then search each sub-query alone and combine the results. This method indexes all possible strings of a pre-specified length L. Dictionary based index size is larger than the database.

BLAST technique [5] used to find local similarity [6] and not global similarity. It is a string matching tool that has two phases: search all database sequences for a fixed substring length w (between 3 and 11) for exact matching (at i). And using a threshold 't', continue searching after the exact match at both direction, left and right, for distance more than 'i' and before 'i-w' till exceed 't'. It stores pointer for location 'i'. So, space needed is more than the database size.

Suffix array [7] scans database strings using a window (window size w, overlapping amount Δ) and count repetition of all possible k-tuples. It stores result at vector of size σ^k (σ referred to alphabet chars A,C,G,T). Then it indexes those vectors at hierarchical binary tree and to compare new query with those vectors it uses Edit distance method. It runs 25 to 50 times faster than BLAST. Disadvantage of this method is the allowing of false drops and index size increase linearly with k value.

The Multi-Resolution index Structure (MRS) [5, 8] uses a sliding window of size w. MRS seeks the result set in different resolution levels. However, the authors only focus on the cost of MRS, and do not evaluate the filtrations efficiency of their proposed technique.

SST [9, 10] scans the database by window w and map results to vector of size 4^w. Then hierarchical clusters, non overlapping, built using k-means algorithm, as any new query need to be processed against the database, using cluster mean and neglect clusters that are far away from the new query. Disadvantages of SST are the complexity of calculations, and false clustering.

The use of blocked inverted index [11] consists of index file (distinct terms) and a set of inverted lists with large-scale full-text system. This method solves two problems: The high storage overhead and considering posting list structure with differentiation between short and long lists. Through this work [11], blocked inverted index where used with skipping approach and propose the random access blocked inverted index (RABI) which enhance space and storage efficiency. This approach divide index into blocks and do compression to different parts of the block using encoding method. For compression it uses Binary interpolative coding (BIC). Access is done at both levels block level and inner block level.

Build self-index [10] for data records using stuffing of delimiters, and give an upper bound, limited by a number of bits, of permanent space in worst case. Analysis's done for space and time efficiency. Storage experiments compare the effects of using stuffing and performance examines three process construct, recover, and retrieval. Results show the effectiveness of FM-index in space and performance. This paper shows the advantages of using FM-index with the addition of adding delimiters.

Handle structural mathematical text and mathematical operation [12] by index real-world scientific documents containing mathematical notation based on full text searching. Mathematical indexing address the following issues, extraction and storing of mathematical notation, and ranking function.

Full text search [13, 14] can be done by different ways. One way is by using N-gram which means we take N characters each time we do processing. N characters processed for searching, by start with 2-gram index then supplement with higher-gram index. Frequently used search terms selected for the incremental index for that this approach have two functions, search engine and index creation engine. For long sequence number of AND operation is large which cause low performance for search, incremental indexing should solve this problem by carefully selecting search terms using search intensive approach. Experimental results show the effectiveness of incremental index even for further stored terms. This method build the incremental index upon subset of terms not for all terms to save space and provide efficiency for most search terms, while in our case we need to have index structure to be ready before searching.

Building suffix array needs time O(n) and space O(n) for constant size text and Suffix tree needs O(n*log*n) time and O(n) space. Suffix array and tree are suitable for pattern searching. Another method uses Compressed Suffix Array [15] and the output array of Burrows-Wheeler Transformation (BWT). In this approach a new algorithm has been developed using terminators at the end of each word.

The use of signature of files for documents retrieval for large database systems allows the use of parallel hardware architecture [16, 17] for full text searching. Controlling false drop and using suitable hashing function with buffer and good storage overhead. The parallel process can be applied for process a document and between documents too. It provide a way for don't care characters.

An ontology kit for full text searching [18] focuses on finding words related to a certain concept (using relevancy ranking function) from a set of concepts. This kit consists of three layers: Full text layer for full text indexing (a Word-based index) and full text searching, Ontology layer for concept definition and ontology maintenance, and User layer for the programming issues.
This kit made use of Apache Lucence and Jena development kits [http://lucene.apache.org/]. They work to get a relevancy ranking between documents that meet the query, which may be met by large number of documents. This kit is a sample of development kits used for evaluating index structure, this kit process depend on relevancy ranking rather than accurate matching.

Most of mentioned works try to build lower bound (D) for the edit distance (ED). Edit Distance is time and space consuming (O(|n|*|n|) time & space) for the whole string. In general we can see that for three strings s_1,s_3 and s_3, if $D(s_1,s_2) > D(s_1,s_3)$, then $ED(s_1,s_2) > ED(s_1,s_3)$).

3. DATA AND METHODOLOGY

Data used in our experiments shown in table 1, we have picked different types (Species kingdom) including Archaea, Eukaryotic, and Bacteria. The Sample file contains DNA sequences only and is of different sizes as show in table 1.

Species kingdom	Accretion No	Length	Size
Archaea	NC_010315.fasta	1,051	2KB
	NC_008318.fasta	15,717	16KB
	EU881703.1.fasta	28,643	29KB
	NC_006389.fasta	33,927	34KB
	AY596291.1.fasta	33,446	34KB
	NC_013966.fasta	63,034	630KB
	NC_011766.fasta	1,365,223	1353KB
	NS_000190.fasta	2,082,083	2000KB
Eukaryotic	AP009202.1.fasta	16,240	17KB
	NC_009684.fasta	16,604	17KB
	NC_010093.fasta	153,819	153KB
	NC_013009.fasta	879,977	872KB
Bacteria	NC_011841.fasta	30,652	31KB
	NC_009471.fasta	37,155	37KB
	NC_013210.fasta	191,799	190KB
	NC_009926.fasta	374,161	371KB
	NC_013009.fasta	879,977	872KB

TABLE 1: Data samples used by our experiments

3.1 Solution Approach

Our approach start by converting DNA sequence into eight columns vectors corresponding to the two wavelet transformation factors; A and B. Calculation like data center, four k-means, four vectors of size v with fixed size of eight-columns instead of a sequence of size n which is variable and long is less in storage size. This transformation has been used for computing second coefficient wavelet transformation with six different windows 'w' of sizes 8, 16, 32, 64, 128, and 256 chars Figure 1 shows the conversion steps.

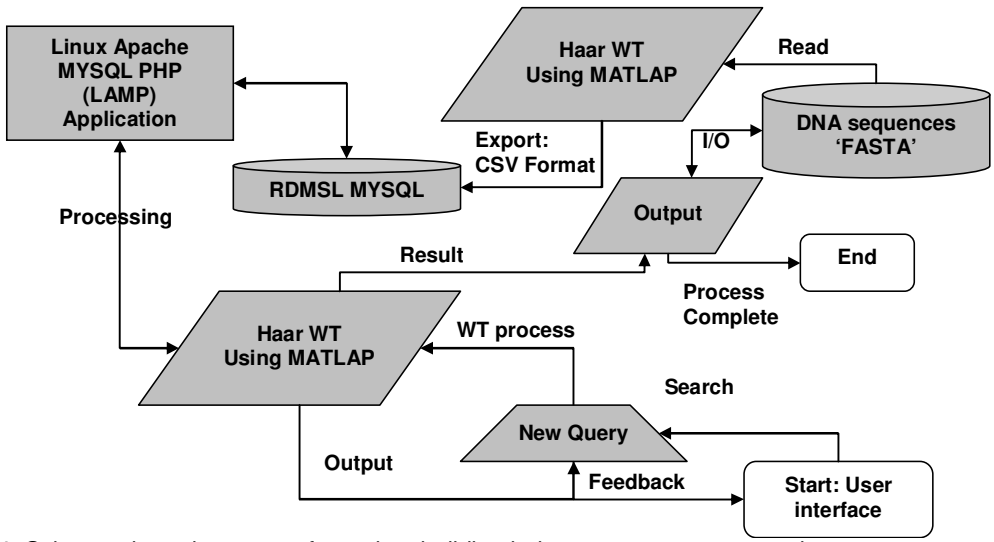


FIGURE 1: Schema chart shows transformation, building index structure, preprocessing new query, and comparison

After transformation we build an index which will be used for searching. Transforming data sequences to numerical representation (NR) will be accomplished.

The aim of using different window sizes 'Wx' is to have different resolution levels of representation of a sequence; we aim, through using different window sizes, to find the values of the window sizes where index structure remain stable, in other word we need the window size where space and search time is optimal. We assume the windows sizes 'Wx' to be 2x. By this assumption after finding the first order wavelet transformation by scanning the database by window Wx1, we can find the wavelet transformation for window values Wxi for i>1 depending on previous window value (Wx1) and no need to scan the database again.

3.1 RDMS indexing evaluation algorithm:

Input:
Database of n sequences, n is a large value.
Each sequence will be donated by Si, i Є [1, .., n] with length Li.
Preprocessing:
We have different window sizes (Wx), x =1, …,6
Transform each Si into number format using Wavelet Transformation (WT), Haar wavelet transformation for Wi+1 can be calculated from Wi as:
(A1,B1),(A2,B2) -> (A1+A2,A1-A2)

```
Initialize i=0
For each Wx value from (Wx min,…, Wx max){
      Slide window Wx over Si
      Calculate Wavelet Transformation
      Wx' = Wx' +1 and i++
Output: set of subsequences (SSi,j), j Є[1,…, m],m=Li/Wx, for each Si
}
Output: two values
1. Transformed subsequence
2. Sequence pointer
Loop through all sequences(i){
      For each pair value of (A,B) for a sequence(i)
            Remove duplicated (A,B) values
}
```

Store pair values at database table.
Build index:
Select index type from RDMS index types, and build data structure upon this index type.
Search by a query sequence:
Search for a new sequence NQ of length |NQ|.
Convert NQ to WT to produce |NQ|-Wx subsequences (NQi)
after moving Wx window over NQ.
Search the database for matching between NQi and SSi,j.

3.2 Constructing Wavelet Coefficients

Each NR row corresponds to a DNA sequence, consists of 8 columns vector. The columns is the wavelet second order coefficients (A,B), A is a 4 columns represent frequency of chars (A,C,G,T) second part B is the difference. Example bellow describes how wavelet works [3]:

$$v_{k,i} = \left(A_{k,i}, B_{k,i}\right),$$

$$A_{k,i} = \begin{cases} f(c_i) & k = 0 \\ A_{k-1,2i} + A_{k-1,2i+1} & 0 < k \le \log_2 n, \end{cases} \quad B_{k,i} = \begin{cases} 0 & k = 0 \\ A_{k-1,2i} - A_{k-1,2i+1} & 0 < k \le \log_2 n, \end{cases}$$

For a sequence u: [ACTC TAGC], consider frequency is done by the order (A, C, G, T) =(2, 3, 1, 2), divide u in two equal parts and recalculate frequency again then do subtraction, you get [1201, 1111] $\rightarrow$(2 3 1 2 , 0 1 –1 0).

The sequences will be represented using six window sizes, w_i for i={1, 2, 3, 4, 5, 6}, Each window size 'w_i' representation will correspond to a final matrix for each sequence, this means we will have six matrices corresponds to w_i value.

Second step: uploading the data on a rational database system (RDMS). We used RDMS to get advantage of RDMS indexing systems like BTree, and Hash. Before uploading data to database tables, all repeated rows had been eliminated to make index size less since there is no need for the repeated data rows. Table 3 shows results of transformation and percentage of repeated data.

Seven types of indexing have been used for evaluation. The types are index on primary key, normal index, primary index, full-text, unique, Hash, and BTree. Our experiments done using two PC's, one with 1GB memory 2GHz, second one is 2 x1GHz CPUs 4GB memory. RDMS used is MYSQL v5.0.1 [19, 20] to store index in, webserver is Apache and language script used is PHP v5 for testing index reach, and Matlab version 7 used for wavelet transformation.

3.3 Index Types

Full-text index allow search for natural language text, some features are: Excludes partial words and words less than x characters in length (3 or less), words that appear in more than half the rows, Hyphenated words are treated as two words, Rows are returned in order of relevance, descending, words in the stopword list (common words) are also excluded from the search results. Full-text had been used to achieve high performance indexing for XML[21].

"Normal" Indexes are the basic index type used by RDMS and require data field to be ordered, Normal Index have no restraints such as uniqueness. Unique Indexes are the same as "Normal" indexes with one difference: all values of the indexed column(s) must only occur once. Primary index are unique indexes for primary keys.

BTree index ,for n keys values, constructed by build a tree with height (h) and a degree (t). Where the degree (t) is greater than or equal to 2. The worst case of BTree is O(*log*n) comparisons. Number of branches for BTree index is larger than the number of branches of other balanced tree structures. Number of branches for a tree controls the logarithm base of complexity (*Log*n of base x where x equal the number of branches). So the base of logarithm tends to be large than required by other tree structures. And what this mean, it means that if we have n key values and we want to build a tree of base x, x branches, as we increase x number of nodes visited during search tends to be smaller. BTree tend to have smaller heights than other trees with the same number of key values. Path to leaf node not exceeding $Log_{n/2}K$ while a binary tree is Log_2K, where Search k-key values are K_1, K_2, … ,K_{n-1}.

BTree make all nodes full at least to a minimum percentage to save space and reducing number of disk references. Space complexity of BTree is O(L/B), where L: length of the sequence and B: block size[22].

In Hash index, bucket reached by key using a hashing function. Records with different key values may map to same bucket; thus entire bucket has to be searched sequentially to locate record.
Bucket Overflows caused by insufficient buckets and distribution of records (Overflow chaining) Collision handling with O(1) complexity, for worst cases performance may deteriorates to O(n). An ideal hash function is uniform/Random and worst map to one bucket. Space complexity for formal Hash function is O(n*log*n), where n: number of keys [22, 13]. Hashing functions divided into two types, Uniform distribution: all buckets have the same number of search-key values.

Random distribution: on average, at any given time, each bucket will have the same number of search-key values, regardless of the current set of values.

Primary index and Unique index both can consists of one or more fields, and both can be clustered/non clustered indexes. The difference is that Primary cannot be Null while Unique can be, there can be only one Primary index on a table but you can have more than one Unique index.

4. EXPERIMENTAL RESULTS

We used two approaches for index evaluation. In the first approach, we created the index on one field representing the coefficients of WT and investigate the effect of changing the type of the index on the response time, which is measured in millisecond. Table 2-a shows the response time. For the second approach, search field is spitted into two parts mainly which are the wavelet transformation coefficients (A, B). Each part consists of four columns. Table 2-b shows the results of this approach.

Index type	W1	W2	W3	W4	W5	W6
DEFAULT	0.002494	0.029923	0.233727	0.677167	0.9619	1.1533
Normal Index	0.01046	0.1679	1.3807	4.05	5.8911	6.3502
PRIMARY	0.093462	0.190009	1.4703	4.233	6.0642	6.4283
Fulltext	0.0028 0.002973 0.003	0.0261 0.028 0.0291	0.2165 0.2273 0.2282	0.259 0.2576 0.2278	1.0997 1.1553 0.993	1.4781 1.4783 1.5482
UNIQUE	0.009996	0.1726	1.406	4.0779	5.931	6.357
Hash	0.0104 0.0106 0.0111	0.1647 0.167 0.1706	1.3639 1.3764 1.3781	4.0591 4.13 4.2522	5.9206 6.0105 6.117	6.2829 6.3638 6.386
BTree	0.010 0.0104 0.010	0.166 0.1664 0.1668	1.3669 1.3791 2.3838	4.0927 4.054 4.249	5.9147 6.0438 6.0883	6.266 6.336 6.410

TABLE 2-a: Evaluation of sample data (under six resolutions Wx) using different index types.

Index type	W1	W2	W3	W4	W5	W6
DEFAULT ON pk	0.018718	0.010034	0.067594	0.209389	0.28135	0.27995
Normal Index	0.001046	0.001547	0.001531	0.001733	0.001544	0.001677
Hash	0.001588	0.001561	0.001646	0.001609	0.001599	0.001651
BTree	0.001605	0.001663	0.001614	0.001688	0.001605	0.001522

TABLE 2-b: Applying indexes for eight columns search fields.

For all index types we do search for the worst case if applicable or randomly. “Default on pk” ordered by order of entry, worst case is the last entry. The same is true for normal index, primary index, and unique index.

Through all experiments, the searching process is applied using the same value, while changing index type, so we can results correctly. For Hashing index type, most database engine uses random hash function, we do the experiment by randomly picking values then the average access time is calculated. BTree index, which is the most popular index over database systems, depends mainly on sorting the data.

Table 3 shows percentage number of returned references to the whole database size while changing window size

$$Ratio = \frac{sequences\ retrieved}{total\ size\ of\ dataset}$$

5. DISCUSSION

We have applied indexes in two different ways, one field index and multiple fields' index. When using one field index, the best performance achieved was using default index and Full-text. When we used BTree or primary index we get the worse performance over all for one field indexing. Almost all other types of indexes give performance close to BTree index.

W	with duplication	no duplication	References%
w1	7069	1250	0.81
w2	73562	23283	0.65
w3	325701	192058	0.41
w4	662746	553933	0.15
w5	813987	796371	0.02
w6	836376	835106	0.002

TABLE 3: Error amount at each resolution used corresponding to amount of reduction.

On the other hand when we used multiple fields index, we have got much better results as shown in table 2-b. Hash, BTree, and normal index on the eight fields give better results when compared with a single field index by table 2-a.

Multiple field index cause overhead for calculation and index address updates in case that amount of updates is high, and overhead for write operations and disk referring. But compared amount of overhead with multiple indexes (merge index) case, this overhead is less. For DNA database, update operations is much less than insert operations and can be neglected. To reduce size of WTR, singular value decomposition (SVD) [23] as a preprocessing step before building index structure for the genome database can be used.

Window size (resolution) affects mainly needed I/O references. When we increase window size the I/O reference operation decreases as shown in table 3. Changing the resolution of the wavelet transformation resolution from low value to high value (from 8 char to 256 char) leads to increase in size of the number of wavelet coefficients and the time of scanning the database. From table 3 when using w1 we get 81% of overall database reference while for w4 this percentage goes down to 15%. Changing window size affect I/O reference percentage directly so as this percentage can be used as a threshold according to application needs.

6. Index Space Complexity

Table3.a shows the space complexity of using one column index with the following index types: full-text, primary index, 8 column index of type's unique, primary, and normal index.

W value	UNIQUE	Primary	Index
w1	2048–2085 B	46250–55296B	92500–122880B
w2	861471–975872 B	861471–975872B	861471–975872B
w3	6940–7850 KB	6940–7850	6940–7850KB
w4	20015–22638 KB	20015–22638 KB	20015–22638KB
w5	28775–32545 KB	28775–32545KB	28775–32545KB
w6	30175–34128 KB	30175–34128KB	30175–34128KB

TABLE 3.A: Space complexity table (B: bytes, KB: Kbytes), range values stands for space used for data and for index.

Figure 2 shows that, depending on input data properties, while increasing sliding window size the size of index is the same even if we don't remove duplicated values.

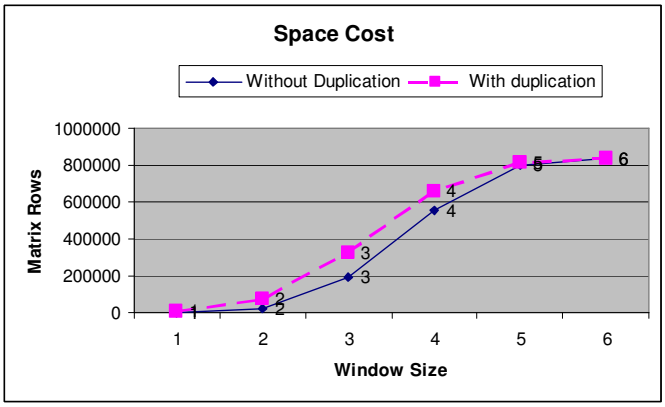
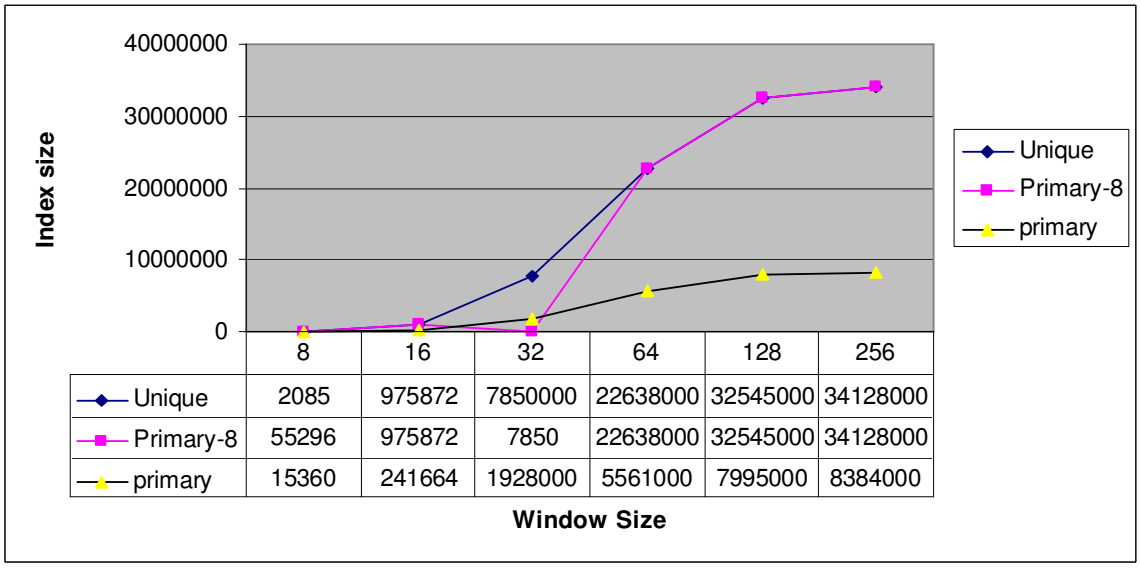
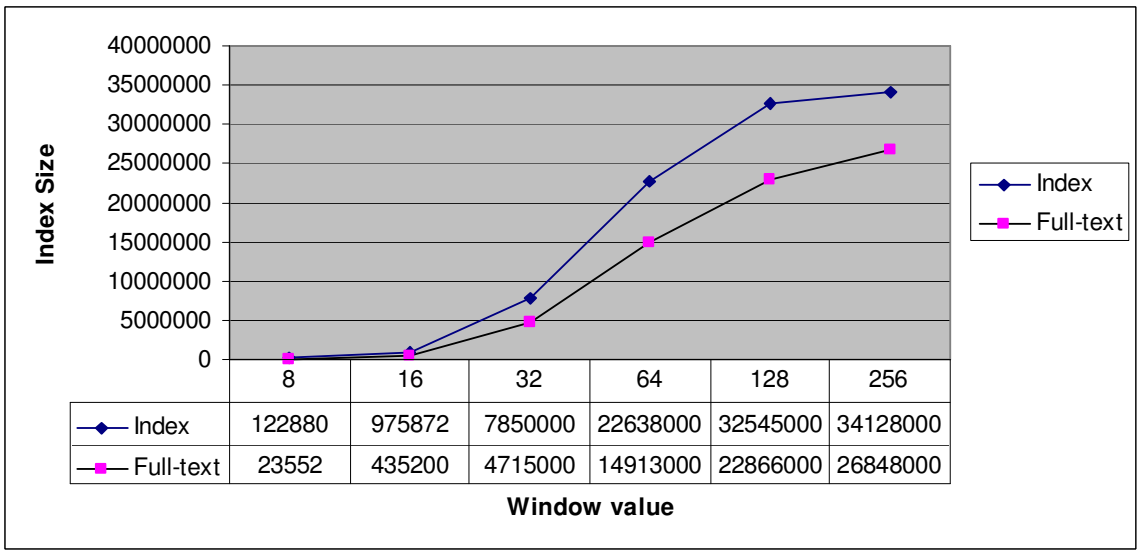


FIGURE 2: Space cost



	8	16	32	64	128	256
Unique	2085	975872	7850000	22638000	32545000	34128000
Primary-8	55296	975872	7850	22638000	32545000	34128000
primary	15360	241664	1928000	5561000	7995000	8384000

FIGURE 3-a: Space cost for all one field (primary, Full-text) and 8-column (index, primary, unique).



	8	16	32	64	128	256
Index	122880	975872	7850000	22638000	32545000	34128000
Full-text	23552	435200	4715000	14913000	22866000	26848000

FIGURE 3-b: Space cost for all one field (primary, Full-text) and 8-column (index, primary, unique).

Tables 3.a and table 3.b show that space complexity variation while changing window size for different index types, it needs to be considered carefully. We can see that index size is larger than data size for all types, as seen from table 3.b where index size is low compared with data size.

W value	UNIQUE	Primary	Index
w1	2048-2085 B	46250-55296B	92500-122880B
w2	861471-975872 B	861471-975872B	861471-975872B
w3	6940-7850 KB	6940-7850	6940-7850KB
w4	20015-22638 KB	20015-22638 KB	20015-22638KB
w5	28775-32545 KB	28775-32545KB	28775-32545KB
w6	30175-34128 KB	30175-34128KB	30175-34128KB

Table 3.a: Space complexity table (B: bytes, KB: Kbytes), range values stands for space used for data and for index.

W value	Full-text	Primary
w1	25000-23552B	25000-15360B
w2	465660-435200 B	475812-241664B
w3	3751-4715 KB	4334-1928KB
w4	10974-14913 KB	13137-5561KB
w5	16253-22866 KB	19364-7995KB
w6	17931-26848 KB	21193-8384KB

TABLE 3.b: Space complexity table (B: bytes, KB: Kbytes), range values stands for space used for data and for index.

Figure 4 and figure 5 shows the DNA data size changing while increasing Wx value for different index types.

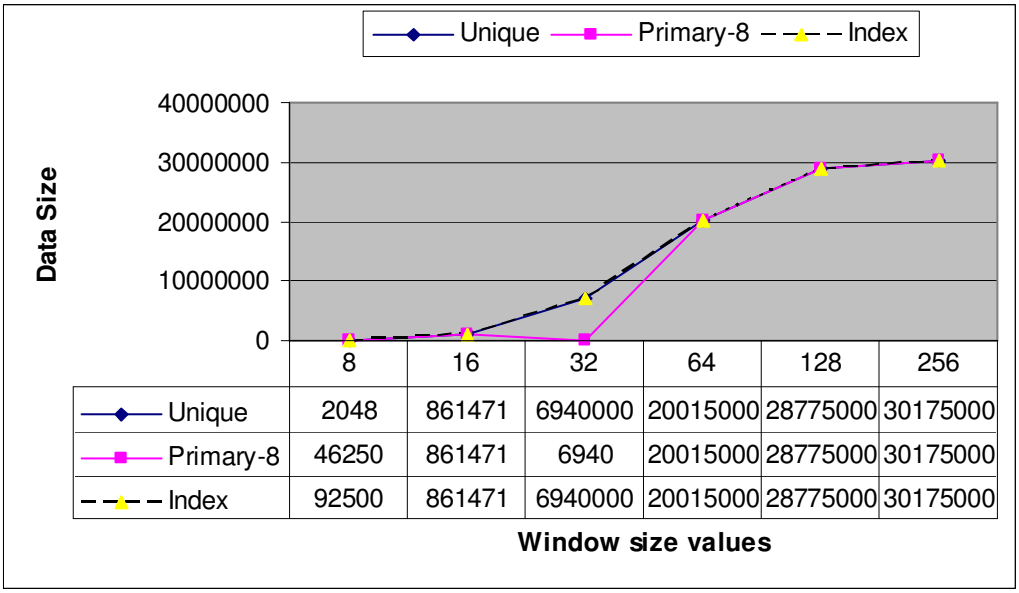
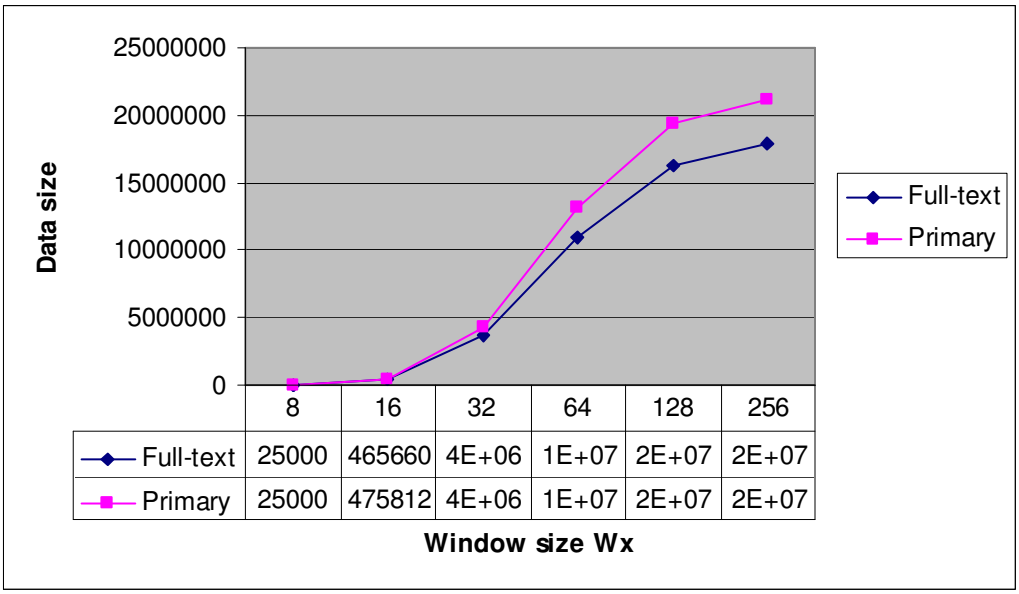


FIGURE 4: Data size while changing W_x value for Unique, Primary 8 columns and index.

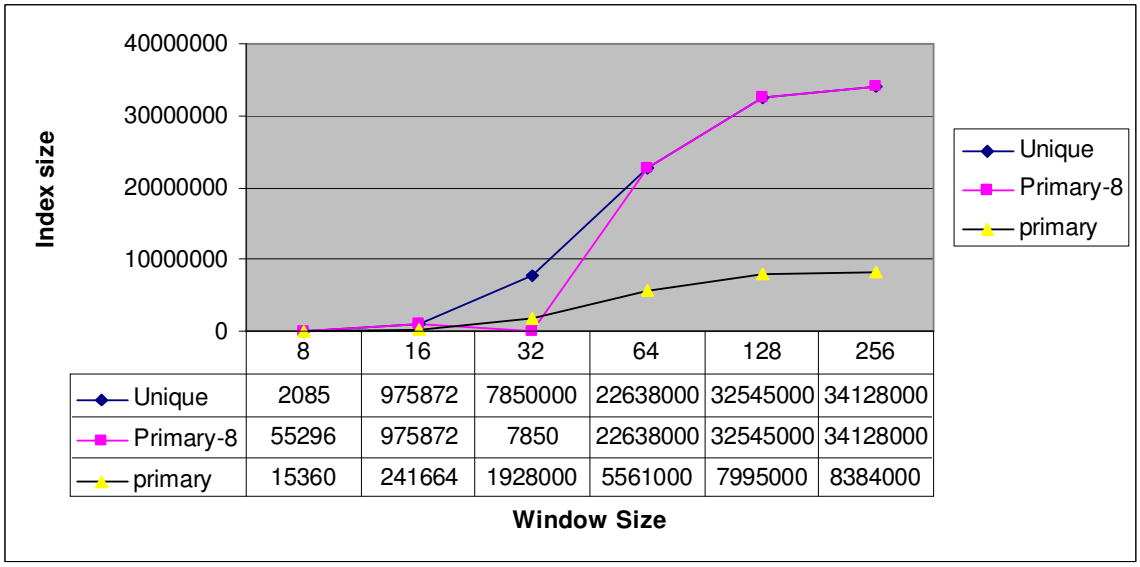


	8	16	32	64	128	256
Full-text	25000	465660	4E+06	1E+07	2E+07	2E+07
Primary	25000	475812	4E+06	1E+07	2E+07	2E+07

FIGURE 5: Data size while changing Wx value for Full-text and Primary index.

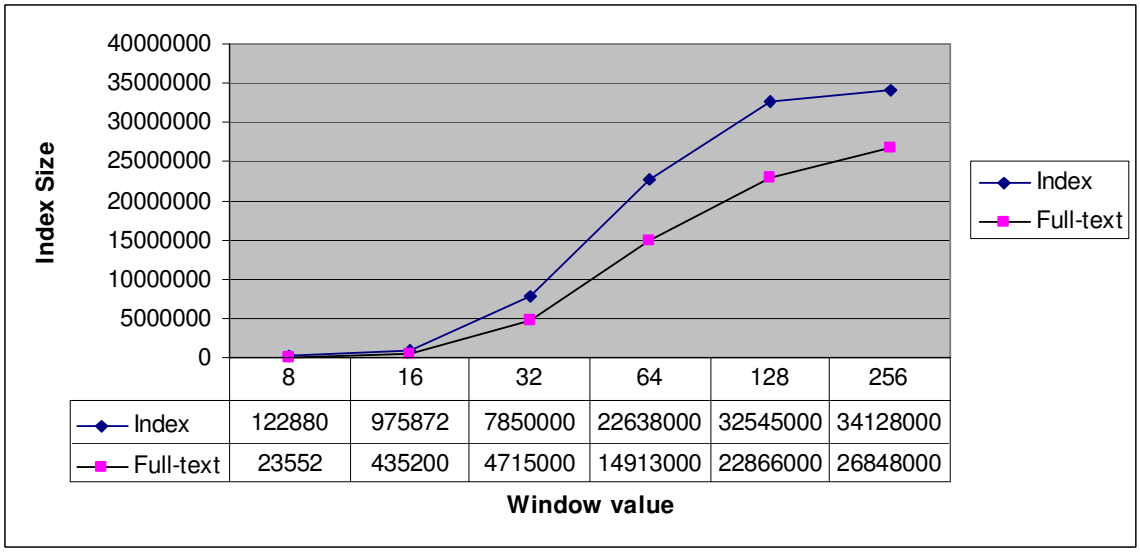
Data size: is highest when using 8-columns index structure, low value when using one field index.

Index size: when using 8-columns almost data and index size are the same. And when using one field index, data and index sizes are relatively the same too.



	8	16	32	64	128	256
Unique	2085	975872	7850000	22638000	32545000	34128000
Primary-8	55296	975872	7850	22638000	32545000	34128000
primary	15360	241664	1928000	5561000	7995000	8384000

FIGURE 6: Index size while changing Wx value for Unique, Primary 8 columns and Primary index.



	8	16	32	64	128	256
Index	122880	975872	7850000	22638000	32545000	34128000
Full-text	23552	435200	4715000	14913000	22866000	26848000

FIGURE 7: Index size while changing W_x value for Full-text and index.

Figures 6 and 7 display the increase at index size for five different index types while changing W_x values.

When comparing time with size of one field index, we found that best time performance achieved by DEFAULT ON pk and full-text but full-text had high space requirement. For 8-column index structure best time achieved by normal, hash, and BTree index.

Access time for 8-column is better than that of one field index but index size equal or more than data size which is a large value. Lowest index size is primary and Full-text as shown by figure 3.

7. CONCLUSION

Our study shows that using multi-fields index improve performance over all types of indexing in spite of the type of index we used. First experiment shows that using specialized index type like full-text or primary index in integer fields give the best performance over using BTree or Hash indexing.

Different window sizes provide multi-resolution index structure. This property gives user a threshold value to determine his needs, and support queries of different sizes. Through our work, we see that no need, when doing query search, to scan the whole database. Instead of scanning the whole database a subset of sequences, which we call candidate sequences, will be referenced from the database after the filtration step. By this way we have minimized the number of disk pages that will be visited at the final stage.

Space and time complexity shows that using special type of index (like Full-text) or using the primary index, of one field, leads to decrease index size, like the full text index when using w6 compared with unique index for the same window size as shown by table 3.a and 3.b. And a higher access time compared to eight fields index type, which lead to larger index size but better access time. This is true, as the Full-text get advantage of its properties as a special index for the search field and the primary index is on integer field, which is less in size than the 8 columns (64.303 compared with 29.577 about one half). This means that a good representation of search field must occupy less space. Small size index, which can be fit in memory, allow the use of in-memory searching mechanisms which gives fast searching time.

From the discussed results, we can see that we need to try to find a less size index structure. Index size is larger than database size, when building index upon eight columns search field. Building the primary index upon a small size relation field is efficient in time and space. Sequence

transformation to numerical format (compact form), good performance index structure (size and time and the use of multi-field index type), and early pruning of false sequences hits leads to build the desired structure.

8. REFERENCES

[1] Effective Indexing and Filtering for Similarity Search in Large Biosequence Databases. Ozgur Ozturk Hakan Ferhatosmanoglu bibe, pp.359, Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03), 2003.

[2] An efficient similarity search based on indexing in large DNA databases, In-Seon Jeong, Kyoung-Wook Park, Seung-Ho Kang, Hyeong-Seok Lim, 2010.

[3] An Efficient Index Structure for String Databases. Tamer Kahveci Ambuj K. Singh Department of Computer Science, University of California Santa Barbara, CA 93106 {amer,ambuj}cs.ucsb.edu, 2001.

[4] Fast Dynamic Programming Based Sequence Alignment Algorithm. Nur'Aini Abdul Rashid', Rosni Abdullah, Abdullah Zawawi Haji Talib, Zalila Ali, IEEE, 2006**.**

[5] MAP: Searching Large Genome Databases. T. Kahveci, A. Singh Pacific Symposium on Biocomputing 8:303-314(2003).

[6] Indexing and retrieval for genomic database.Hugh E. Williams, Member, IEEE, and Justin Zobel, Member, IEEE Computer Society, IEEE, 2002.

[7] S. Muthukrishnan and S. C. Sahinalp. Approximate nearest neighbor and sequence comparison with block operations, 2000.

[8] CoMRI: A Compressed Multi-Resolution Index Structure for Sequence Similarity Queries. Hong Sun1, Ozgur Ozturk1, Hakan Ferhatosmano glu, IEEE, 2003.

[9] E. Giladi et al., SST: An Algorithm for Finding Near-Exact Sequence Matches in Time Proportional to the Logarithm of the Database Size. Bioinformatics 18, 873–877, 2002.

[10] An Efficient Approach for Building Compressed Full-text Index for structured Data: Jun Liang, Lin Xiao, Di Zhang IEEE, 2009.

[11] Efficient Maintenance Schema of Inverted Index for Large-scale Full-Text Retrieval, Xiaozhu Liu, State Key Lab of Software Engineering Wuhan University Wuhan 430072, China , School of Automation Wuhan University of Technology IEEE, 2010.

[12] Mathematical Extension of Full Text Search Engine, Jozef Misutka, Leo Galambos, Department of Software Engineering, Charles University in Prague, Ke Karlovu 3, 121 16 Prague, Czech Republic, 2008.

[13] Experimental Simulation on Incremental Three-gram Index for Two-gram Full-Text Search Systems, Hiroshi Yamamoto Seishiro Ohmi Hiroshi Tsuji IEEE, 2003.

[14] A Compact Memory Space of Dynamic Full-Text Search using Bi-Gram Index, El-Sayed Atlam, El-Marhomy Ghada, Masao Fuketa, Kazuhiro Morita and Jun-ichi Aoe, Department of Information Science and Intelligent Systems, University of Tokushima Tokushima,770-8506, Japan 2004.

[15] Breaking a Time-and-Space Barrier in Constructing Full-Text Indices, Wing-Kai Hon, Kunihiko Sadakane_ Wing-Kin Sung IEEE, 2003.

[16] Parallel Selection Query Processing Involving Index in Parallel Database Systems. J. Wenny Rahayu David Taniar, IEEE, 2002.

[17] An Architecture for Parallel Search of Large, Full-text Databases, Nassrin Tavakoli and Hassan Modaress-Razavi, Department of Computer Science, The University of North Carolina at Charlotte, Charlotte, NC 28223 IEEE, 1990.

[18] An Ontology Enhanced Development Kit for Full Text Search, Su Jian, Weng Wenyong, Wang Zebing, Lab of Digital City & Electronic Service, Zhejiang University City College, Hangzhou 310015, China IEEE, 2009.

[19] Alexander Rubin, Senior Consultant, MySQL AB, Full Text Search in MySQL 5.1 New Features and HowTo, http://www.mysqlfulltextsearch.com/full_text.pdf, 2006.

[20] Moshe Shadmon, The ScaleDB Storage Engine, http://www.scaledb.com/pdfs/ScaleDB_MySQL_Preso2009.ppt, 2009.

[21] A Hybird Method for Efficient Indexing of XML Documents. Sun Wei, Da-xin Lui, IEEE, 2005.

[22] The SBCTree: An Index for RunLength Compressed Sequences, Mohamed Y. Eltabakh ,Wing-Kai Hon, Rahul Shah, Walid G. Aref, Jeffrey S. Vitter Purdue University, 2008, 2008

[23] Efficient Filtration of Sequence Similarity Search Through Singular Value Decomposition. S. Alireza Aghili Ozgur D. Sahin Divyakant Agrawal Amr El Abbadi, IEEE 2004.

Case Based Medical Diagnosis of Occupational Chronic Lung Diseases From Their Symptoms and Signs

Prem Pal Singh Tomar *singhppst@rediffmail.com*
Faculty of Engineering,
Dayalbagh Educational Institute,
Uttar Pradesh,
Agra, India

Dr. Ranjit Singh *rsingh_dei@yahoo.com*
Faculty of Engineering,
Dayalbagh Educational Institute,
Uttar Pradesh,
Agra, India

Dr. P K Saxena *premkumarsaxena@gmail.com*
Faculty of Engineering,
Dayalbagh Educational Institute,
Uttar Pradesh,
Agra, India

Jeetu Sharma *jitusharma_19@rediffmail.com*
Electronics & Electrical engineering
Anand Engg. College
Agra

Abstract

The clinical decision support system using the **c**ase based reasoning (CBR) methodology of Artificial Intelligence (AI) presents a foundation for a new technology of building intelligent computer aided diagnoses systems. This Technology directly addresses the problems found in the traditional Artificial Intelligence (AI) techniques, e.g. the problems of knowledge acquisition, remembering, robust and maintenance. In this paper, we have used the Case Based Reasoning methodology to develop a clinical decision support system prototype for supporting diagnosis of occupational lung diseases. 127 cases were collected for 14 occupational chronic lung diseases, which contains 26 symptoms. After removing the duplicated cases from the database, the system has trained set of 47 cases for Indian Lung patients. Statistical analysis has been done to determine the importance values of the case features. The retrieval strategy using nearest-neighbor approaches is investigated. The results indicate that the nearest neighbor approach has shown the encouraging outcome, used as retrieval strategy. A Consultant Pathologist's interpretation was used to evaluate the system. Results for Sensitivity, Specificity, Positive Prediction Value and the Negative Prediction Value are 95.3%, 92.7%, 98.6% and 81.2% respectively. Thus, the result showed that the system is capable of assisting an inexperience pathologist in making accurate, consistent and timely diagnoses, also in the study of diagnostic protocol, education, self-assessment, and quality control. In this paper, clinical decision support system prototype is developed for supporting diagnosis of occupational lung diseases from their symptoms and signs through employing Microsoft Visual Basic .NET 2005 along with Microsoft SQL server 2005 environment with the advantage of Object Oriented Programming technology

Key words: Clinical Support System, Artificial Intelligence, Case-Based Reasoning, Pathologist

1. INTRODUCTION

The use of artificial Intelligence (AI) technique i.e case-based reasoning (CBR), in the the development of Clinical Support System has a relatively young history, arose out of the research in cognitive science. The earliest contributions in this area were from Roger Schank and his colleagues at Yale University [1],[2]. During the period 1977–1993, CBR research was highly regarded as a plausible high-level model for cognitive processing. It was focused on problems such as how people learn a new skill and how humans generate hypotheses about new situations e cognitive-based researches were to construct decision based on their past experiences. Many prototype of decision support system based on CBR technique were built during this period: for example, Cyrus [3],[4], Mediator [5], Persuader [6], Chef [7], Julia [8], Casey, and Protos [9]. Three CBR workshops were organized in 1988, 1989, and 1991 by the U.S. Defense Advanced Research Projects Agency (DARPA). These formally marked the birth of the discipline of Decision Support System using case-based reasoning.

Computerized evidence-based guidelines and Clinical decision support systems (CDSS) have been promoted as the key to improving effectiveness and efficiency of clinical decisions [14]. Although the use of decision support systems (DSS) in the field of medicine has accelerated in recent years [15].
Many researchers are working on Clinical Support System using CBR with many diverse applications ranging from psychiatry and epidemiology to clinical diagnosis Most of them aim for a successful implementation of CBR methods to enhance the work of health experts to improve the efficiency and quality of health care Researchers who have contributed substantially to CBR in medicine include Gierl Schmidt and their colleagues who focused on a range of applications including children dysmorphic syndromes, antibiotics therapy advising for intensive care and monitoring emerging diseases (Gierl, 1993 Schmidt & Gierl, 2001) Notable is their ICONS system Gierl, 1993),first applied to the determination of antibiotic therapy treatment for intensive care then to the prognosis of kidney function defects, For this latter application ICONS learned prototypes associated with graded levels of severity through temporal abstraction Gierl, 1993),and matched new cases with these prototypes to predict the severity of a renal disease [16].

Some real Clinical Support Systems based on CBR technique are: CASEY that gives a diagnosis for the heart disorders [10], GS.52 which is a diagnostic support system for dysmorphic syndromes, NIMON is a renal function monitoring system, COSYL that gives a consultation for a liver transplanted patient [11] and ICONS that presents a suitable calculated antibiotics therapy advise for intensive care patients [12].
Computerized evidence-based guidelines and Clinical decision support systems (CDSS) have been promoted as the key to improving effectiveness and efficiency of clinical decisions [17]. Although the use of decision support systems (DSS) in the field of medicine has accelerated in recent years [18].

However, none of the aforementioned studies presented results that showed evidence of first, the inclusion of all the 14 occupational lung diseases perspective; and secondly a system capable of assisting a Pathologist who is not specialized in the aspect of occupational lung diseases diagnosis. Thus, this system proposes a medical decision support system for diagnosis of occupational lung diseases as an improvement of earlier works.

2. JUSTIFICATION FOR STUDY

In clinical practice, making decision involves a careful analysis of harms and benefits associated with different treatment options. These decisions, often associated with high stake and important long term consequences, are frequently made in presence of limited resources and information and an incomplete clinical picture. Under such circumstances, a rigorous and objective analysis of outcomes and probabilities is essential to achieve the best possible decision given a specific clinical situation.

Therefore, Pathologist is required to be fully conversant with the diversity of possible patterns, recognize and diagnose them, timely and accurately. Hence, a Pathologist who is not a specialist in the pathology of the occupational lung diseases has to refer to textbooks and study past diagnosis before concrete diagnosis can be made and conclusion reached. Hence, there is the need for a system, which can assist the Pathologist to reach timely and accurate decision.

3. KNOWLEDGE ENGINEERING TASKS IN DEVELOPING CBR BASED SYSTEM

The problem-solving life cycle in a CBR system consists essentially of the following :

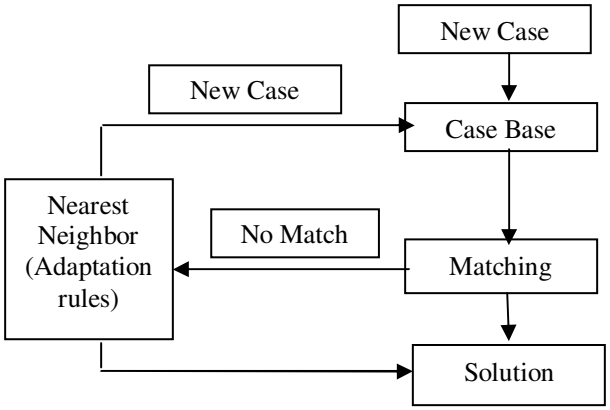


FIGURE 1: Case Base Reasoning Technique

The figure shows that when a new case comes to the system, the system carries out the work of matching. Upon getting exact match same result is displayed while in case no exact match is found the nearest neighbor is looked for whose result is adjusted according to the new case using the adaptation rules and the result is displayed. Such a new case is also saved in the case base for future assistance. Accordingly the methodology of developing Clinical Support System, CBR-based systems in specific domain can be summarized in the following steps:

1. Retrieving :The system will search its Case-Memory for an existing case that matches the input problem specification.

2. Reusing :If we are lucky (our luck increases as we add new cases to the system), we will find a case that
Exactly matches the input problem and goes directly to a solution.

3. If we are not lucky, we will retrieve a case that is similar to our input situation but not entirely appropriate to provide as a completed solution.

4. Revising: The system must find and modify small portions of the retrieved case that do not meet the input specification. This process is called "case-adaptation".

5. Retaining :The result of case adaptation process is (a) Completed solution, and (b) generates a new case that can be automatically added to the system's case- memory for future use

4. KNOWLEDGE ACQUISITION

Knowledge acquisition is a process of acquiring, organizing and studying knowledge for the lung diseases. The data and knowledge of Clinical Support System based on Case-Base technique are collected from different sources. The first primary source is, acquired from a physician (Domain Expert). The second source is from specialized databases like lung disease diagnostic laboratory at Agra, Uttar Pradesh, India, books and a few electronic websites. This knowledge can be divided by important fact into 26 facts, which are shown on table 1.

No.	Chief symptoms	No.	Chief symptoms
1.	Cough	14	Alcohol use
2.	Dyspnea	15	Heart rhythm problem
3.	Chest Discomfort	16	Abdominal pain
4.	Malaise	17	Shoulder pain
5.	Fever	18	difficulty in swallowing
6.	Wheezing	19	Pain under rib cage
7.	Hemoptysis	20	Chemicals exposure
8.	Persistent cough	21	Fungi exposure
9.	Fever with chill	22	humidifiers exposure
10	Night sweat	23	Coke oven emissions
11.	Asbestosis exposure	24	Silica exposure
12.	Excessive sweating	25	Coal dust
13.	Smoking	26	Cotton Dust

TABLE 1: Fact List of Symptoms

5. ASSIGNING IMPORTANCE VALUE TO CASE SYMPTOMS

Features weights for most problem domains are context dependent. The weight assigned to each feature of the case tells how much attention to pay to matches and mismatches in the field when computing the distance measure of a case. Those that are good predictors are then assigned higher importance for matching [10].

The importance of the feature depends upon its prevalence among the diseases. If a feature is common among all diseases like Cough, then it will have the least importance in leading to a diagnosis.

6. CASE INDEXING AND RETRIEVAL

Here, we focus our discussion on case indexing and retrieval strategy. Case indexing and retrieval are two separate but closely related processes. Since a case memory may contain thousands of cases, case indices organize their key features to expedite the search process. Case retrieval searches the case base to find candidate cases that share significant features with the new case. Existing literature in case-based reasoning has proposed several mechanisms for case indexing and retrieval. A good review of early literature can be found in [13].

7. RETRIEVAL USING NEAREST-NEIGHBOR TECHNIQUE

If however an exact match is not found, which can be the case many times, nearest neighbor technique (ref. table 2) and adaptation rules have to be used. Let T is new case and C1 and C2 are old cases then Nearest neighbor formula = sum of (weight*similarity)/sum of weight

(T, C1)=72/97=0.74

(T, C2)=65/97=0.67

So, C1 is the nearest neighbor.

Then the presence of the symptoms in the new and the old case is listed in the next two columns. Local similarity is given in Clinical Decision support system. The total of all the weights is calculated by adding them which is 97. Then the
sum of weight*similarity is calculated by adding all the products of weight*similarity. In the first comparison the sum is 72 while in the second comparison it is 65. The sum of weights in the first comparison is: 97
The nearest neighbor value is: 72/97= 0.74
In the second comparison:
The sum of weights is: 65
The sum of weights in the first comparison is: 67
The nearest neighbor value is: 38/67 = 0.57

Therefore, the first comparison, which is case C1, is the nearest neighbor for the new case T.

The system will use the result of the nearest match found and use adaptation rules to 'revise' this result according to the demands of the novel situation. The system uses the Nearest-neighbor algorithm that finds the closest matches of the cases already stored in the database to the new case using a distance calculation, which determines how similar two cases are by comparing their features, the pseudo code of this algorithm [10] can be written as follows:
For each feature in the input case:
Find the corresponding feature in the stored case
Compare the two values to each other and compute the degree of match
Multiply by a coefficient representing the importance of the feature to the match
Add the results to derive an average match score
This number represents the degree of match of the old case to the input.
A case can be chosen by choosing the item with the largest score.
Nearest-neighbor techniques applied to the retrieval phase of a CBR system (i.e., measuring similarity among cases). The equation

$$\text{Similarity}(T, S) = \Sigma f(Ti\ Si) \times Wi / \Sigma Wi$$

represents a typical nearest-neighbor technique that describes a situation for which T(Target case) and S (Source case) are two cases compared for similarity, n is the number of attributes in each case, i is an individual attribute from 1 to n, and Wi is the feature weight of attribute. Similarities are usually normalized to fall within the range 0 to 1, where 1 means a perfect match and 0 indicates a total mismatch.

8. DEVELOPMENT AND RESULTS

Development of clinical decision support system prototype is through employing Microsoft Visual Basic .NET 2005 environment with the advantage of Object Oriented Programming technology. The Microsoft SQL server 2005 was used to develop the database module.

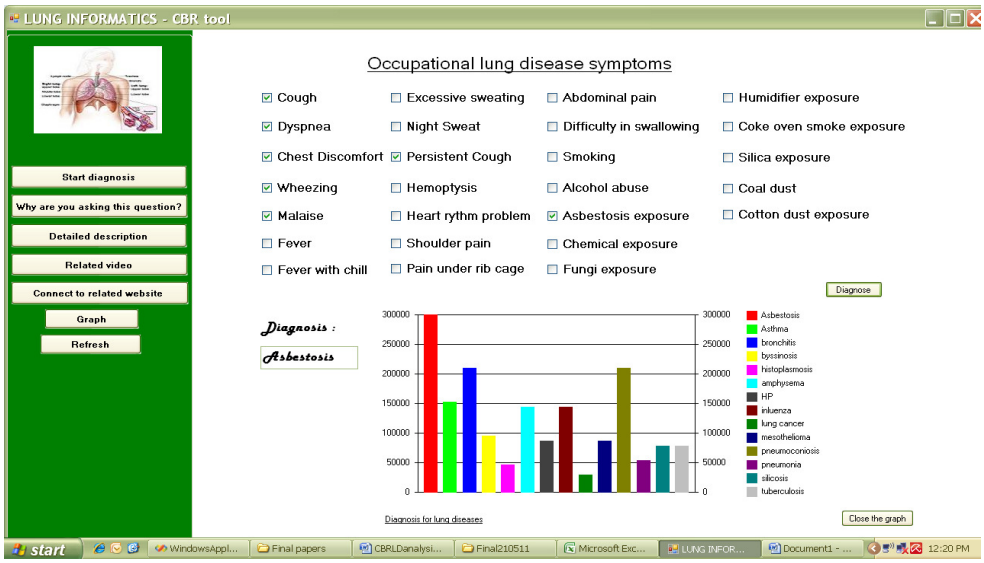


FIGURE 2: Diagnostic window

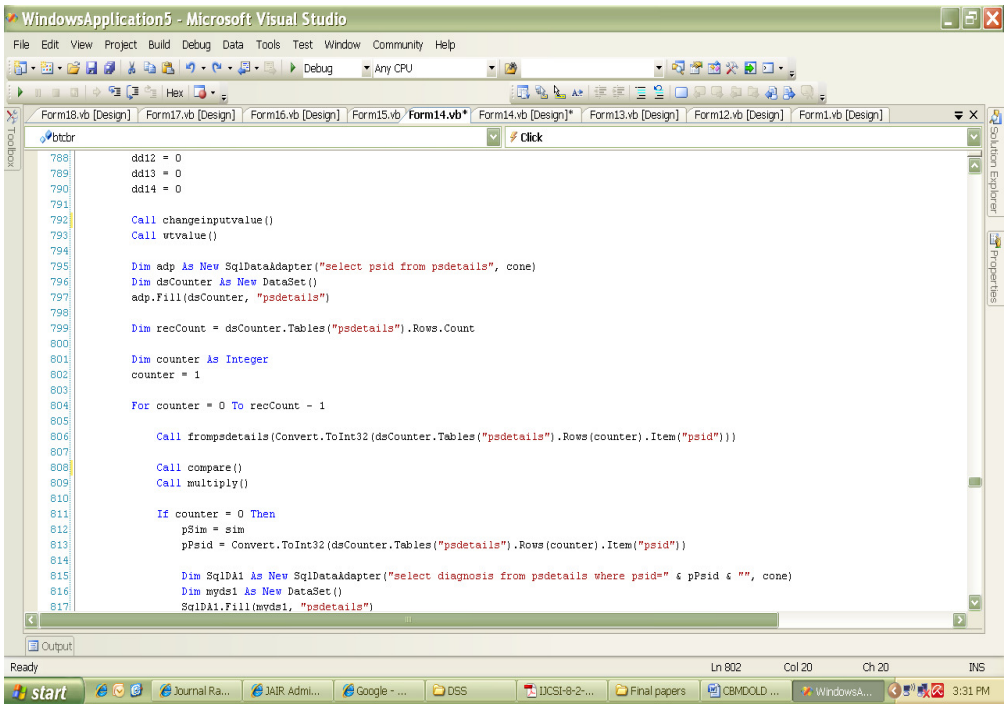


FIGURE 3 : Code window

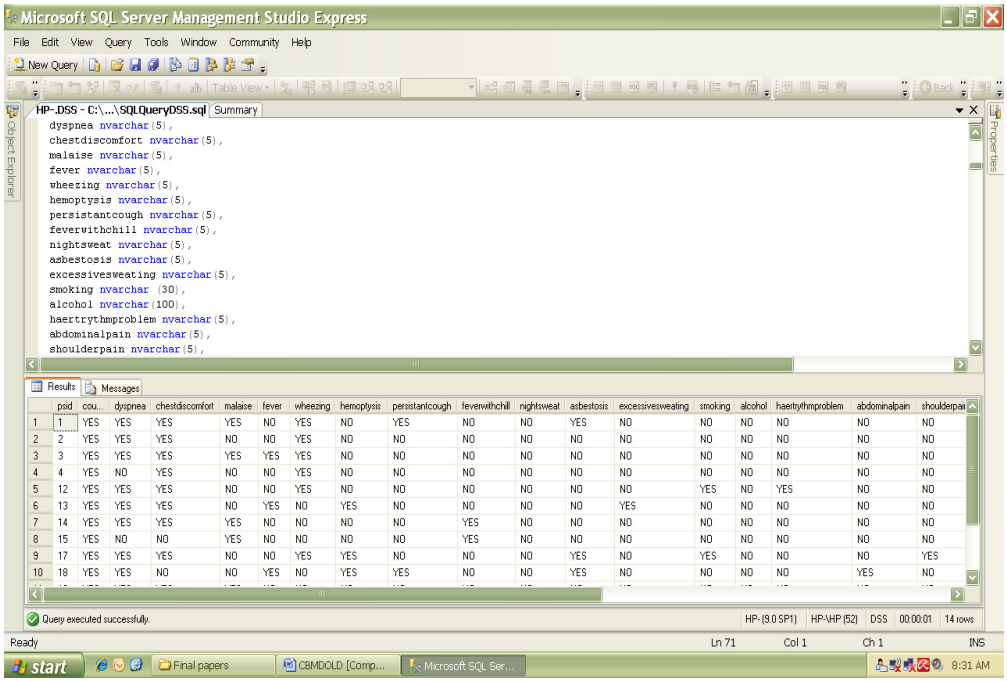


FIGURE 4: SQL server window

In this paper, the architecture, and the implementation of a prototype of a Clinical Support System using case-based technique that supports diagnosis of occupational lung diseases was developed. Knowledge structure was represented via a formalism of cases. The system used nearest-neighbor techniques technique for the retrieval process.

Using a Consultant Pathologist's interpretation as a "gold standard" (reference test), the system's parameters for diagnosing occupational lung diseases were calculated.

(i) True positive (TP):
The diagnostic system yields positive test result for the sample and thus the sample actually has the disease;
(ii) False positive (FP):
The diagnostic system yields positive test result for the sample but the sample does not actually have the disease;
(iii) True negative (TN):
The diagnostic system yields negative test result for the sample and the sample does not actually have the disease; and
(iv) False negative (FN):
The diagnostic system yields negative test result for the sample but the sample actually has the disease.

The formulas for used for calculating Sensitivity,
Specificity, PPV and NPV are:

Sensitivity = [TP/(TP+FN)] x 100% …… (1)

Specificity = [TN/TN+FP] x 100%……. (2)

PPV = [TP/ (TP+FP)] x 100% …… (3)

NPV = [TN/ (TN+FN)] x100%……..…(4)

Using equations (1), (2), (3) and (4), respectively, the Sensitivity, Specificity, Positive
Prediction Value (PPV) and the Negative Prediction Value of the system are:
Sensitivity = 95.3%;
Specificity = 92.7%;
PPV = 98.6%
NPV = 81.2%.

8. SUMMARY AND CONCLUSION

In this paper we presented a clinical support system, which could be used by stakeholders for arriving at very vital decisions regarding the diagnosis of 14 occupational chronic lung diseases. The focus was on the development of a clinical support system that can assist Pathologist, especially those who may not be specialist in the area of occupational chronic lung diseases treatment. Thus, the system attempts to improve the effectiveness of diagnosis (in relation to accuracy, timeliness and quality) that is performed by a human pathologist, rather than improve their efficiency with respect to decision making. Therefore, the diagnoses made by the system are at least as good as those made by a human expert.
From the development and analysis of Clinical Support System, it is evident that CBR technique of Artificial Intelligence (AI) is appropriate methodology for all medical domains and tasks for the following reasons: cognitive adequateness, explicit experience and subjective knowledge, automatic acquisition of subjective knowledge, and system integration. CBR technique presents an essential technology of building intelligent Clinical Support System for medical diagnoses that can aid significantly in improving the decision making of the physicians. the
proposed method gives an Sensitivity = 95.3%; and PPV = 98.6% which is better than the existing methods. Future research involves more intensive testing using a larger occupational chronic lung disease database to get more accurate results.

The use of Microsoft Visual Basic .NET 2005 along with Microsoft SQL server 2005 as database is found to be very effective in producing the system under windows environment. For future work, more cases will be added to the case memory.

REFERENCES

[1] R. Schank and R. Abelson (eds.), Scripts, Plans, Goals and Understanding, Lawrence Erlbaum Associates, Hillsdale, NJ, 1977

[2] R. Schank, Dynamic Memory: A Theory of Reminding and Learning in Computers and People, Cambridge University Press, New York, 1982

[3] J. L. Kolodner, Maintaining organization in a dynamic long term memory, Cognitive Science, vol. 7, pp. 243–280, 1983.

[4] J. L. Kolodner, Reconstructive memory, a computermodel, Cognitive Science, vol. 7,pp. 281–328, 1983.

[5] R. L. Simpson, A computer model of case-based reasoning in problem solving: an investigation in the domain of dispute mediation, Ph.D. dissertation, School of Information and Computer Science, Georgia Institute of Technology, Atlanta, GA, 1985.

[6] K. Sycara, Using case-based reasoning for plan adaptation and repair, in Proceedings of the Case-Based Reasoning Workshop, DARPA, Clearwater Beach, FL, Morgan Kaufmann,San Francisco, pp. 425–434, 1988.

[7] K. J. Hammond, Case-Based Planning, Academic Press, San Diego, CA, 1989.

[8] T. R. Hinrihs, Problem Solving in OpenWorlds, Lawrence Erlbaum Associates, Hillsdale,NJ, 1992.

[9] R. Bareiss, Exemplar-Based Knowledge Acquisition: A Unified Approach to ConceptRepresentation, Classification, and Learning, Academic Press, San Diego, CA, 1989.

[10] J. L. Kolodner (Ed.),Case-Based Reasoning,Morgan Kaufmann Publishers: California, 1993

[11] M. Lenz, S. Wess, H. Burkhard and B. Bartsch, Case-Based Reasoning Technology:From Foundations to applications, Springer, 1998.

[12] B. Heindl. Et al,: A Case-Based Consiliarius for TherapyRecommendation (ICONS): Computer-Based Advice for calculated Antibiotic Therapy in Intensive Care Medicine. Computer Methods and Programms in Biomedicine 52 , pp. 117-127, 1997.

[13] E. L. Rissland, J. Kolodner, & D. Waltz, Case-based reasoning from DARBA: “Machine learning program plan”. Proceedings of the case-Based ReasoningWorkshop. San Mateo, CA: Morgan Kaufmann, pp. 1-13,1998.

[14] Fieschi M, Dufour J-C, Staccini P, et al. Medical decision support systems: old dilemmas and new paradigms? Tracks for successful integration and implementation. Methods Inform Med 2003; 42:190–198.8

[15] Decision support systems for antibiotic prescribing, Vitali Sintchenkoa,b,c, Enrico Coierac and Gwendolyn L. Gilberta,b , Current Opinion in Infectious Diseases 2008,21:573–579

[16] Medical applications in case based reasoning , ALEC HOLT, ISABELLE BICHINDARITZ, RAINER SCHMIDT, and PETRA PERNER , The Knowledge Enineering review, Vol.20:3, 289-292 – 2006, Cambridge University Press

[17] Graham B, Detsky AS, the application of decision analysis to the surgical treatment of early osteoarthritis of wrist. J Bone Joint Surg Br 2001:83:650-4

[18] Detsky AS, Nalie G, Krahn MD, Naimark D, Redelmeier DA. Primer on medical decision analysis : Part 1- Getting started. Med Decision Making 1997 : 17:123-5

Application of Microarray Technology and Softcomputing in Cancer Biology : A Review

P.K.Vaishali *vaishali5599@gmail.com*
Department of Computer Science & Information Technology,
Jyothishmathi Institute of Technology & Science,
JNTU, Hyderabad, AP, INDIA.

Dr.A.Vinayababu *dravinayababu@yahoo.com*
Professor of CSE, Director of Admissions
JNTUH University,
Hyderabad, AP, INDIA.

Abstract

DNA microarray technology has emerged as a boon to the scientific community in understanding the growth and development of life as well as in widening their knowledge in exploring the genetic causes of anomalies occurring in the working of the human body. microarray technology makes biologists be capable of monitoring expression of thousands of genes in a single experiment on a small chip. Extracting useful knowledge and info from these microarray has attracted the attention of many biologists and computer scientists. Knowledge engineering has revolutionalized the way in which the medical data is being looked at. Soft computing is a branch of computer science capable of analyzing complex medical data. Advances in the area of microarray –based expression analysis have led to the promise of cancer diagnosis using new molecular based approaches. Many studies and methodologies have come up which analyszes the gene espression data by using the techniques in data mining such as feature selection, classification, clustering etc. emboiding the soft computing methods for more accuracy. This review is an attempt to look at the recent advances in cancer research with DNA microarray technology , data mining and soft computing techniques.

Keywords: DNA Microarray, Classification, Data Mining ,Soft Computing ,Gene Expression.

1. INTRODUCTION

Deoxyribonucleic acid (DNA) micro array technology provides tools for studying the expression levels of a large number of distinct genes simultaneously [9]. Micro array technology allows biologists to simultaneously measure the expressions of thousands of genes in a single experiment [8] [10] [11].

Gene expression data is widely used in disease analysis and cancer diagnosis [5]. Gene expression data from DNA micro arrays are characterized by many measured variables (genes) on only a few observations (experiments) although both the number of experiments and genes per experiment are growing rapidly [4] [6]. Gene expression data from DNA micro array can be characterized by many variables (genes), but with only a few observations (experiments). Prediction, classification, and clustering techniques are being used for analysis and interpretation of the data [1]. An important application of gene expression micro array data is classification of biological samples or prediction of clinical and other outcomes [2]. Micro array technology is to classify the tissue samples using their gene expression profiles as one of the several types (or subtypes) of cancer. Compared with the standard histopathological tests, the gene expression profiles measured through micro array technology provide accurate, reliable and objective cancer classification. The DNA micro array data for cancer classification consists of large number of genes (dimensions) compared to the number of samples or feature vectors [3] [7]. Classification analysis of micro array gene expression data has been widely used to uncover biological features and to distinguish closely related cell types that often appear in the diagnosis of cancer [37].Many researchers have developed and demonstrated different classification techniques for cancer classification based on micro array gene expression data. Feature selection techniques [12],[13] have been suggested before classification, which finds the top features that discriminate various classes. Kernel based techniques [14],[15] like SVM have already been used

for binary disease classification problems. Gene selection[16] and neural networks[17] based classifications were also reported in microarray data analysis. soft computing has been successively used in bioinformatics thereby providing low cost, low ,better approximation and indeed good and more accurate solutions.

2. DNA MICROARRAY TECHNOLOGY

Although all of the cells in the human body contain the same genetic material, the same genes are not active in all of those cells. Studying which genes are active and which are inactive in different kinds of cells helps scientists understand more about how these cells function and about what happens when the genes in a cell don't function properly. In the past scientists have only been able to conduct such genetic analyses on a few genes at once. With the development of DNA microarray technology, however, scientists can now examine thousands of genes at the same time, an advance that will help them determine the complex relationships between individual genes. The mountain of information that is the draft sequence of the human genome may be impressive, but without interpretation that is all it remains — a mass of data. Gene function is one of the key elements researchers want to extract from the sequence, and the DNA microarray is one of the most important tools at their disposal.Microarray technology will help researchers learn more about many different diseases—heart disease, mental illness, and infectious disease, to name only a few. One intense area of microarray research at the NIH is the study of cancer.In the past, scientists have classified different types of cancer based on the organs in which the tumors develop. With the help of microarray technology, however, they will be able to further classify these types of cancer based on the patterns of gene activity in the tumor cells and will then be able to design treatment strategies targeted directly to each specific type of cancer. Additionally, by examining the differences in gene activity between untreated and treated—radiated or oxygen-starved, for example—tumor cells, scientists can better understand how different types of cancer therapies affect tumors and can develop more effective treatments.
In short the usefulness of dna technology can be listed as
1.Can follow the activity of MANY genes at the same time.
2.Can get a lot of results fast
3.Can COMPARE the activity of many genes in diseased and healthy cells
4.Can categorize diseases into subgroups

Microarray Technology: Application in medical

Gene discovery: DNA Microarray technology helps in the identification of new genes, know about their functioning and expression levels under different conditions.

Disease Diagnosis: DNA Microarray technology helps researchers learn more about different diseases such as heart diseases, mental illness, infectious disease and especially the study of cancer. Until recently, different types of cancer have been classified on thedevelop. Now, with the evolution of microarray technology, it will be possible for the researchers to further classify the types of cancer on the basis of the patterns of gene activity in the tumor cells. This will tremendously help the pharmaceutical community to develop more effective drugs as the treatment strategies will be targeted directly to the specific type of cancer.

Drug Discovery: Microarray technology has extensive application in *Pharmacogenomics.* Pharmacogenomics is the study of correlations between therapeutic responses to drugs and the genetic profiles of the patients. Comparative analysis of the genes from a diseased and a normal cell will help the identification of the biochemical constitution of the proteins synthesized by the diseased genes. The researchers can use this information to synthesize drugs which combat with these proteins and reduce their effect.

Toxicological Research: Microarray technology provides a robust platform for the research of the impact of toxins on the cells and their passing on to the progeny. Toxicogenomics establishes correlation between responses to toxicants and the changes in the genetic profiles of the cells exposed to such toxicants

3. DATAMINING AND SOFT COMPUTING PARADIGM IN THE AREA OF GENE EXPRESSION IN CANCER DATA - RECENT RELATED RESEARCH : A REVIEW

There exists considerable literature on the application of different soft computing paradigm in the area of gene expression cancer data sets:One of the first landmark studies using microarray data to analyze tumor samples was done by Golub *et al.* [18] . This study on human acute leukemia showed that it was possible to use microarray data to distinguish between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without any previous knowledge. For the first time the potential of microarray data was shown by illustrating its use in discovering new classes using the previously introduced class discovery methods (i.e., unsupervised analysis) and, second, by using microarray data to assign tumors to known classes (i.e., supervised analysis).

Perou *et al.* [19] did a similar analysis using hierarchical clustering on breast cancer and found different groups of breast tumors.

Alon *et al.* [20] adopted a two-way clustering method whereby both genes and tumors were clustered. They showed that colon tumors and normal colon tissues were separated based on the microarray data. Also, they showed that co-regulated families of genes also clustered together.

Tibshirani *et al*. [21] built further on the development of supervised microarray data analysis methods by developing the nearest shrunken centroid method, also known as PAM. This technique not only allows predicting of classes, but also tries to limit the number of genes necessary to make the prediction. By limiting the number of genes, it is possible to develop cheaper methods to make a diagnostic test, such as smaller microarrays or quantitative PCR. After these studies research groups focused more on class.

Ahmad M. Sarhan [22] has presented a stomach cancer detection system based on Artificial Neural Network (ANN), and the Discrete Cosine Transform (DCT). The proposed system has extracted classification features from stomach micro arrays using the DCT. The features extracted from the DCT coefficients have been then applied to an ANN for classification (tumor or non-tumor). The micro array images used in this study have been obtained from the Stanford Medical Database (SMD). Simulation results have shown that the proposed system produces a very high success rate.

Bharathi *et al*. [23] have attempted to find the smallest set of genes that can ensure highly accurate classification of cancer from micro array data by using supervised machine learning algorithms. The significance of finding the minimum gene subset has been three fold:1) It has greatly reduced the computational burden and noise arising from irrelevant genes.2) It has simplified gene expression tests to include only a very small number of genes rather than thousands of genes, which could significantly bring down the cost for cancer testing. 3) It has called for further investigation into the possible biological relationship between these small numbers of genes and cancer development and treatment. Their simple yet very effective method has involved two steps. In the first step, they have chosen some important genes using a 2 way Analysis of Variance (ANOVA) ranking scheme. In the second step, they have tested the classification capability of all simple combinations of those important genes using a good classifier such as Support Vector Machines. Their approach has obtained very high accuracy with only two genes.

Bo Li *et al*. [24] have discussed that the gene expression data collected from DNA micro array are characterized by a large amount of variables (genes), but with only a small amount of observations (experiments). They have proposed a manifold learning method to map the gene expression data to a low dimensional space, and then explore the intrinsic structure of the features so as to classify the micro array data more accurately. The proposed algorithm could project the gene expression data into a subspace with high intra-class compactness and inter-class separability. Experimental results on six DNA micro array datasets have demonstrated that their method is efficient for discriminant feature extraction and gene expression data classification. Their work is a meaningful attempt to analyze micro array data using manifold learning method; there should be much room for the application of manifold learning to bioinformatics due to its performance.

Xiaosheng Wang *et al*. [25] have discussed that a gene selection is of vital importance in molecular classification of cancer using high-dimensional gene expression data. Because of the distinct characteristics inherent to specific cancerous gene expression profiles, developing flexible and robust feature selection methods has been extremely crucial. They have investigated the properties of one feature selection approach proposed in their previous work, which has been the generalization of the feature selection method based on the depended degree of attribute in rough sets. They have compared the feature selection method with the established methods with respect to the depended degree, chi-square, information gain, Relief-F and symmetric uncertainty, and analyzed its properties through a series of classification experiments. The results have revealed that their method is superior to the canonical depended degree of attribute based method in robustness and applicability. Moreover, their method has been comparable with the other four commonly used methods. More importantly, their method could exhibit the inherent classification difficulty with respect to different gene expression datasets, indicating the inherent biology of specific cancers.

Mallika et al. [26] have presented a novel method for improving classification performance for cancer classification with very few micro array Gene expression data. The method employs classification with individual gene ranking and gene subset ranking. For selection and classification, the proposed method has used the same classifier. The method has been applied to three publicly available cancer gene expression datasets from Lymphoma, Liver and Leukaemia datasets. Three different classifiers namely Support vector machines-one against all (SVM-OAA), K nearest neighbour (KNN) and Linear Discriminant analysis (LDA) have been tested and the results have indicated improvement in performance of SVM-OAA classifier with satisfactory results on all the three datasets when compared to the other two classifiers.

Chhanda Ray [27] has discussed DNA micro array gene expression patterns of several model organisms and provided a fascinating opportunity to explore important abnormal biological phenomena. The development of cancer has been a multi-step process in which several genes and other environmental and hormonal factors play an important role. They have proposed a new algorithm to analyze DNA micro array gene expression patterns efficiently for huge amount of DNA micro array data. For better visibility and understanding, experimental results of DNA micro array gene pattern analysis have been represented graphically. The shape of each graph corresponding to a DNA micro array gene expression pattern has been determined by using an eight-directional chain code sequence, which has been invariant to translation, scaling, and rotation. The cancer development has been identified based on the variations of DNA micro array gene expression patterns of the same organism by simultaneously monitoring the expressions of thousands of genes. At the end, classification of cancer genes has also been focused based on the distribution probability of codes of the eight-directional chain code sequences representing DNA micro array gene expression patterns and the experimental result has been provided.

While Chu *et al.* [31] [36]has used a five-genes set for 100% correct classification on the lymphoma data in the fuzzy NF framework, A dynamic fuzzy neural network, involving self-generation,parameter optimization, and rulebase simplification, is used [31][36] for the classification of cancer data such as lymphoma,liver cancer.

Banerjee *et al.* [35] [36]obtained a misclassification for just two samples from the test data using a two-genes set. In case of the leukemia data, a two-genes set is selected, whereas the colon data results in an eight-genes reduct size. An evolutionary rough feature selection algorithm [35][36] has been used for classifying microarray gene expression patterns. The effectiveness of the algorithm is demonstrated on three cancer datasets, viz., colon, lymphoma, and leukemia

S.Mitra et al. Has discussed An evolutionary rough *c*-means clustering algorithm applied to microarray gene expression data [28][36]. RSes are used to model the clusters in terms of upper and lower approximation.

S.Bicciato et.al. discussed An autoassociative neural network for *simultaneous* pattern identification, feature extraction, and classification of gene expression data [30] Results are demonstrated on leukemia and colon cancer20 datasets. The identification of gene subsets for classifying two-class

Modelling/data mining tasks	Result obtained	Reference

disease samples has been modeled as a multiobjective evolutionary optimization problem by K.Deb et.al. [32], involving minimization of gene subset size to achieve reliable and accurate classification based on their expression levels. Classification of acute leukemia, having highly similar appearance in gene expression data, has been made by combining a pair of classifiers trained with mutually exclusive features [29].

RSes have been applied mainly tomicroarray gene expression data, in mining tasks like classification [38], [33], H.Midelfart .et.al usedClassification rules (in *if–then* form) for extracting data from microarray data [38], using RSes with supervised learning. Gastric tumor classification in microarray data is made using rough set-based learning [33].

4. ASSEMBLANCE OF SOFT COMPUTING TECHNIQUES WITH MICROARRAY TECHNOLOGY IN CANCER BIOLOGY

The esemblence of soft computing techniques applied to Microarray Technology in Cancer Biology is described in table1. The table also gives the information of different datamining tasks involved.

5. CONCLUSION

Microarray technology technology has suppressed the conventional cancer diagnostic methods based on the morphological appearance of the cancerous cell which quiet often were misdiagnosed.
For more precision and effective results the emboiding of the different soft computing approaches is really recommendable.Based on the numerous publications investigating the use of DNA microarray technology ,data mining and soft computing to predict outcome in different cancer sites, this technology seems to be the most mature technology from all the omics.

Class Discovery methods.	Assigning tumors to known classes.	[34]
2-way Clustering	Both genes & tumors were clustered	[38]
Hirarchical clustering	Found different groups of Breast cancer.	[35]
Nearest Shrunken centroid method (PAM)	Limit on the number of genes necessary to prediction.	[39]
ANN & DCT	Very high success rate for classification of tumor and non tumors	[22]
Supervised machine learning	High accuracy with only two genes	[23]
Manifold learning method	Efficient discriminant feature extraction and gene expression data classification.	[24]
Rough sets ,feature selection	Superior in applicability and robustness.	[25]
Gene ranking and gene subset ranking	Improved classification performance	[26]
ANN,classification	Simultaneous pattern extraction,Leukemia classification	[29][30]
GA,classification	reliable and accurate classification based on their expression levels,minimization of gene subset size	[32]
NF,feature selection	Feature selection	[32]
Fuzzy NN(dynamic structure growing),feature selection. ANN,classifiers	Colon classification,Classification of acute leukemia, having highly similar appearance in gene expression data	[30][32]
RS+GA,clustering	effectiveness of the algorithm is demonstrated on three cancer datasets, viz., colon, lymphoma, and leukemia.	[28]
NF,self-generation, parameter optimization, and rulebase simplification,classification,feature selection.	Lymphoma classification.	[31]
NF,rule base simplification.	Classification of Small round blue cell tumor	[31]
NF,classification	Liver cancer100% correct classification on the lymphoma data	[31]
RS+GA,classification	Gastric tumor classification	[33]
GA(multi objective approach)	Lung cancer & mixed lineage leukemia more efficient resuls in gene selection as compared to single objective	[37]
GA ,SVM	Multiclass cancer categarization	[38]

TABLE 1: Use of Microarray Technology with Softcomputing in Cancer Research

REFERENCES

[1] Nguyen and Rocke, Classification of Acute Leukemia based on DNA Micro array Gene Expressions using Partial Least Squares, Kluwer Academic, Dordrecht, 2001

[2] Jian J. Dai, Linh Lieu, and David Rocke, "Dimension Reduction for Classification with Gene Expression Micro array Data", Statistical Applications in Genetics and Molecular Biology: Vol. 5, No. 1, 2006

[3] Alok Sharma and Kuldip K. Paliwal, "Cancer classification by gradient LDA technique using micro array gene expression data", Data & Knowledge Engineering, Vol. 66, pp. 338-347, 2008

[4] Chun-Hou Zheng, Bo Li, Lei Zhang and Hong-Qiang Wang, "Locally Linear Discriminant Embedding for Tumor Classification", In Proceedings of ICIC, pp.1093-1100, 2008

[5] Cheng-San Yang, Li-Yeh Chuang, Chao-Hsuan Ke and Cheng-Hong Yang, "A hybrid Feature Selection Method for Micro array Classification", International Journal of Computer Science, Vol. 35, No. 3, 2008

[6] Danh V. Nguyen, David M. Rocke, "Tumor Classification by Partial Least Squares Using Micro array Gene Expression Data", Bioinformatics, Vol. 18, No. 1, pp. 39-50, 2002

[7] Pengyi Yang and Zili Zhang, "An Embedded Two-Layer Feature Selection Approach for Microarray Data Analysis", IEEE Intelligent Informatics Bulletin, Vol.10, No.1, pp. 24-32, 2009

[8] Yuh-Jye Lee and Chia-Huang Chao, "A Data Mining Application to Leukemia Micro array Gene Expression Data Analysis", International Conference on Informatics, Cybernetics and Systems (ICICS), Kaohsiung, Taiwan, 2003

[9] James J. Chen and Chun-Houh Chen, "Micro array Gene Expression", Encyclopedia of Biopharmaceutical Statistics, 2nd Edition, Marcel Dekker, Inc., pp. 599-613, 2003

[10] Seeja and Shweta, "Microarray Data Classification Using Support Vector Machine", International Journal of Biometrics and Bioinformatics (IJBB), Vol. 5, No. 1, pp. 10-15, 2011

[11] Yee Hwa Yang and Natalie P. Thorne, "Normalization for Two-color cDNA Microarray Data", Science and Statistics: A Festschrift for Terry Speed, Vol. 40, pp. 403-418, 2003

[12] Fei Pana, Baoying Wanga, Xin Hub and William Perrizoa, "Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis", Journal of Biomedical Informatics, Vol. 37, pp. 240–248, 2004. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring”, Science, 286(15):531–537, 1999.

[13] Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Michèl Schummer, and David Haussler, “Support vector machine classification and validation of cancer tissue samples using microarray expression data “, Bioinformatics6(10): 906-914 , 2000

[14] Zhang, X. and Ke, H.,” ALL/AML cancer classification by gene expression data using SVM and CSVM approach”, Genome Informatics, Universal Academy Press, pp. 237-239, 2000.

[15] Xin Zhao, Leo Wang-Kit Cheung, “Kernel-imbedded Gaussian processes for disease classification using microarray gene expression data”, BMC Bioinformatics.,8:67,2007.

[16] Wenlong Xu, Minghui Wang, Xianghua Zhang, Lirong Wang, Huanqing Feng," SDED: Anovel filter method for cancer-related gene selection", Bioinformation 2(7): 301-303,2008.

[17] D.P. Berrar, C.S. Downes, W. Dubitzky, "Multiclass Cancer Classification Using Gene Expression Profiling and Probabilistic Neural Networks", Pacific Symposium on Biocomputing 8:5-16, 2003.

[18] Golub TR, Slonim DK, Tamayo P, et al. Molecular classifi cation of cancer: class discovery and class prediction by gene expression monitoring. Science 1999 ; 286 : 531 -7

[19] Perou CM, SÃ¸rlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000 406 : 747 -52

[20] Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999 ; 96 : 6745 -50

[21]. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002 ; 99 : 6567 -72

[22] Ahmad M. Sarhan, "Cancer Classification Based on Micro array Gene Expression Data Using DCT and ANN", Journal of Theoretical and Applied Information Technology, Vol. 6, No. 2, pp. 208-216, 2009

[23] Bharathi and Natarajan, "Cancer Classification of Bioinformatics data using ANOVA", International Journal of Computer Theory and Engineering, Vol. 2, No. 3, pp. 369-373, June 2010

[24] Bo Li, Chun-Hou Zheng, De-Shuang Huang, Lei Zhang and Kyungsook Han, ""Gene expression data classification using locally linear discriminant embedding", Computers in Biology and Medicine, Vol. 40, pp. 802–810, 2010

[25] Xiaosheng Wang and Osamu Gotoh, "A Robust Gene Selection Method for Micro array-based Cancer Classification", Journal of Cancer Informatics, Vol. 9, pp. 15-30, 2010

[26] Mallika and Saravanan, "An SVM based Classification Method for Cancer Data using Minimum Micro array Gene Expressions", World Academy of Science, Engineering and Technology, Vol. 62, No. 99, pp. 543-547, 2010

[27] Chhanda Ray, "Cancer Identification and Gene Classification using DNA Micro array Gene Expression Patterns", International Journal of Computer Science Issues, Vol. 8, Issue 2, pp. 155-160, March 2011.

[28] S. Mitra, "An evolutionary rough partitive clustering," *Pattern Recognit. Lett*., vol. 25, pp. 1439–1449, 2004

[29] S. B. Cho and J. Ryu, "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," *Proc. IEEE*, vol. 90, no. 11, pp. 1744–1753, Nov. 2002.

[30] S. Bicciato,M. Pandin, G. Didon`e, andC.DiBello, "Pattern identification and classification in gene expression data using an autoassociative neural network model," *Biotechnol. Bioeng.*, vol. 81, pp. 594–606, 2003.

[31] F. Chu, W. Xie, and L. Wang, "Gene selection and cancer classification using a fuzzy neural network," in *Proc. 2004 Annu. Meet. North Amer. Fuzzy Information Processing Soc. (NAFIPS)*, vol. 2, pp. 555–559.

[32] K. Deb and A. Raji Reddy, “Reliable classification of two-class cancer data using evolutionary algorithms,” *BioSystems*, vol. 72, pp. 111–129, 2003.

[33] H. Midelfart, J. Komorowski, K. Nørsett, F. Yadetie, A. K. Sandvik,and A. Lægreid, “Learning rough set classifiers from gene expression and clinical data,” *Fundamenta Inf.*, vol. 53, pp. 155–183, 2002.

[34] M. E. Futschik, A. Reeve, and N. Kasabov, “Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue,” *Artif. Intell. Med.*, vol. 28, pp. 165–189, 2003.

[35] M. Banerjee, S. Mitra, and H. Banka, “Evolutionary-rough feature selection in gene expression data,” *IEEE Trans. Syst., Man, Cybern. C, Appl.*

[36] S.Mitra, YHaya.shi,Bioinformatics with softcomputing” IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 36, NO. 5, SEPTEMBER 2006.

[37] Fei Pana, Baoying Wanga, Xin Hub and William Perrizoa, "Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis", Journal of Biomedical Informatics, Vol. 37, pp. 240–248, 2004

[38] H. Midelfart, A. Lægreid, and J. Komorowski, *Classification of Gene Expression Data in an Ontology,* vol. 2199. Lecture Notes in Computer Science, Berlin, Germany: Springer-Verlag, 2001, pp. 186–194

Offline Handwritten Signature Identification and Verification Using Multi-Resolution Gabor Wavelet

Mohamad Hoseyn Sigari *hoseyn_sigari@ieee.org*
Machine Vision Res. Lab, Computer Eng. Department, Ferdowsi University of Mashhad, Mashhad, Iran

Muhammad Reza Pourshahabi *reza.pourshahabi@gmail.com*
Machine Vision Res. Lab, Computer Eng. Department, Ferdowsi University of Mashhad, Mashhad, Iran

Hamid Reza Pourreza *hpourreza@um.ac.ir*
Machine Vision Res. Lab, Computer Eng. Department, Ferdowsi University of Mashhad, Mashhad, Iran

Abstract

In this paper, we are proposing a new method for *offline* (*static*) handwritten signature identification and verification based on *Gabor wavelet* transform. The whole idea is offering a simple and robust method for extracting features based on Gabor Wavelet which the dependency of the method to the nationality of signer has been reduced to its minimal. After pre-processing stage, that contains noise reduction and signature image normalisation by size and rotation, a *virtual grid* is placed on the signature image. Gabor wavelet coefficients with different frequencies and directions are computed on each points of this grid and then fed into a classifier. The *shortest weighted distance* has been used as the classifier. The weight that is used as the coefficient for computing the shortest distance is based on the distribution of instances in each of signature classes.

As it was pointed out earlier, one of the advantages of this system is its capability of signature identification and verification of *different nationalities*; thus it has been tested on four signature dataset with different nationalities including *Iranian*, *Turkish*, *South African* and *Spanish* signatures. Experimental results and the comparison of the proposed system with other systems are consistent with desirable outcomes. Despite the use of the simplest method of classification i.e. the nearest neighbour, the proposed algorithm in comparison with other algorithms has very good capabilities. Comparing the results of our system with the accuracy of human's identification and verification, it shows that human identification is more accurate but our proposed system has a lower error rate in verification.

Keywords: Signature Identification; Signature Verification; Multi-Resolution Analysis; Gabor Wavelet; Nearest Neighbour.

1 INTRODUCTION

Nowadays, person identification (recognition) and verification is very important in security and resource access control. For this purpose, the first and simple way is to use Personal Identification Number (PIN). But, PIN code may be forgotten or may be misused. Now, an interesting method for identification and verification is biometric approach. Biometric is a measure of identification or verification that is unique for each person. Always biometric is carried along with person and cannot be forgotten. In addition, biometrics usually cannot be misused. Handwritten signature is one of the oldest biometrics.

Handwritten signature identification or verification is simple, fairly secure, inexpensive, non-intrusive and acceptable in society. Nevertheless, it has some drawbacks: lower identification rate with respect to other biometrics, non-linear changes with size changing and dependency to time and emotion. Another problem of processing the handwritten signature is the differences between signatures from different nationalities. For example, European signature is the same as his/her name written in a special style but Persian signature contains some curves and symbols [1, 2, 3].
Signature processing can be used for two different purposes: (1) *identification* (*recognition*) and (2) *verification* (*authentication*); signature verification is more useful than signature identification in both practical systems and researches. In signature identification, the input is an unknown signature and system must identify the owner of that. But the goal of signature verification is examination of an input signature to determine whether it is genuine or forgery. So, in the verification system the major problem is the presence of signature forgery. There are three types of forgery:

(1) *Random forgery*: this type of forgery is not intentional. If the forger uses the name of a person in his/her own style to create a forgery, it is known as the random forgery. In fraudulent cases, the majority of them are random forgeries, and they could be easily detected.

(2) *Simple or casual forgery*: the forger does not have any prior experience and imitates the signature in amateur style. This imitation is done by observing the signature just in a matter of time.

(3) *Expert or skilled or simulated forgery*: the most difficult forgeries are created by expert forger who has experience in copying the signatures. The forgery signatures that are created in this way will be almost a genuine replica.

There are two types of signature identification and verification: (1) *static* or *offline* and (2) *dynamic* or *online*. In the offline type, input of the system is a 2-dimentional image of the signature. In contrast, in the online type, the input is the signature trace in time domain. In the online type, a person signs on an electronic tablet by an electronic pen and his/her signature is sampled. Each sample has 3 attributes: x and y in 2-dimentions coordinates and t as the time of sampling. Therefore, in the online type, the time attribute of each sample help us to extract useful information such as start and stop points, velocity and acceleration of signature stroke. Some electronic tablets in addition of time sampling, can digitize the pressure. This additional information existing in the online type will increase the identification rate in comparison with the offline type. Although the accuracy rate in the online type is higher than the offline type, but the online type has a major disadvantage; it is online. So, it cannot be used for some important applications that the signer cannot be presented in the singing place.

In this paper, we propose an offline signature identification and verification system, which emphasizes on feature extraction using Gabor wavelet. Extracting suitable and robust features are more important than selecting a classifier. So, in our proposed system, we used a simple classifier known as nearest neighbor.

Remain of this paper is organized as follow: in section 2, some previous works are reviewed. Section 3 is a brief description of our proposed system. Section 4, 5 and 6 are about pre-processing, feature extraction and classification of the proposed system respectively. Section 7 shows experimental results on four different signature databases. The last section of this paper is about the conclusions and future works.

2 RELATED WORKS

In this section a short review on offline handwritten signature identification and verification systems is presented. Major of these researches are about signature verification, however some of them are about signature identification.

Frias-Martinez et al [4] proposed an offline handwritten signature identification system using Support Vector Machines (SVM) and compared this system with another system which used Multi-Layer Perceptrons (MLP) as classifier. Both of these systems have been tested with two different feature extraction approaches: (1) extracting some global and moment-based features, (2) using raw bitmap data of signature image as feature vector. Their proposed system used just one signature per class as training data similar to the practical systems. Experimental results show that SVM is better than MLP for classification in both approaches of feature extraction.

Ozgunduz et al [5] described an offline handwritten signature identification and verification system using the global, directional and grid features of signatures. Before extracting features, all signature images were pre-processed by background elimination, noise reduction, width normalization and thinning the stroke. SVM is used to identify or verify the signatures. Experimental results show that the performance of SVM is higher than MLP.

Kalera et al [6] presented a quasi multi resolution approach for offline signature identification and verification. First, all signature were normalized by rotation. Then GSC (Gradient, Structural and Concavity) features are extracted and fed into a Bayesian classifier. Gradient features are local; and structural and concavity features are global. So feature extraction acts like a multi-resolution processing.

Deng et al [7] proposed a wavelet-based offline signature verification system. This system extracts robust features that exist within different signatures of the same class and verify whether a signature is a forgery or not. After pre-processing stage, the system starts with a closed contour tracing algorithm to extract closed contour of signature. The curvature data of the closed contours are decomposed to low and high frequency bands using wavelet transform. Then the zero crossings information corresponding to the curvature data are extracted as features. Classification stage in this system is very simple and performed by applying a threshold. The threshold value used for verifying an input signature is calculated automatically based on the distribution of features in each class. Experimental results were done on two different signature databases: English and Chinese; these results show that nationality had no impact on the nature of the system.

Herbst et al [8] designed a signature verification system using Discrete Radon Transform and Dynamic Programming. First, all signatures are normalized by Translation, Rotation and Scaling. Then Radon Transform has been applied to extract features. A grid relation between features of input signature and features of reference signatures has been created using Dynamic Programming. Afterward, matching analysis was done to accept or reject the input signature.

Coetzer et al [9] have used Radon Transform and Hidden Markov Model (HMM) for offline signature verification. Features are extracted by Radon Transform and fed to a HMM classifier. The ring topology of HMM classifier has been used in this paper.

Ferrer et al [10] introduced some new geometric features for offline signature verification based on signature curvature and distribution of strokes in Cartesian and Polar coordination. These features were used by HMM, SVM and Nearest Neighbor (NN) classifiers to verify an input signature image. Experimental results shown that HMM is more accurate than SVM and NN classifiers.

Kiani et al [11] extracted appropriate features by using Local Radon Transform applied to signature curvature and then classified them using SVM classifier. Their proposed method is robust with respect to noise, translation and scaling. Experimental results were implemented on two signature databases: Persian (Iranian) and English (South African).

Pourshahabi et al [12] presented an offline signature identification and verification using Contourlet Transform. Contourlet is a two dimensional multi-resolution transform that extracts curves from an image with different thicknesses and curvatures. In this paper, after signature

normalization, features were extracted using Contourlet Transform and then classified by Euclidean Distance. This method was applied on two signature databases: Persian (Iranian) and English (South African).

3 PROPOSED SYSTEM

The proposed system is consisting of three stages: (1) *pre-processing* stage, (2) *feature extraction* stage and (3) *classification* stage. In pre-processing stage, noise elimination of the signature image is performed. Rotation and size normalization of the signature image are also achieved in this stage. Feature extraction stage is based on the computation of Gabor wavelet coefficients on specific points of the pre-processed signature image. Extracted features (wavelet coefficients) are then fed to a classifier. In the signature identification system, the identity of the signer is recognized in the classification stage whereas; in the signature verification system the forgery or genuine type of the signature is determined. The diagram of the proposed system is shown in Figure 1.

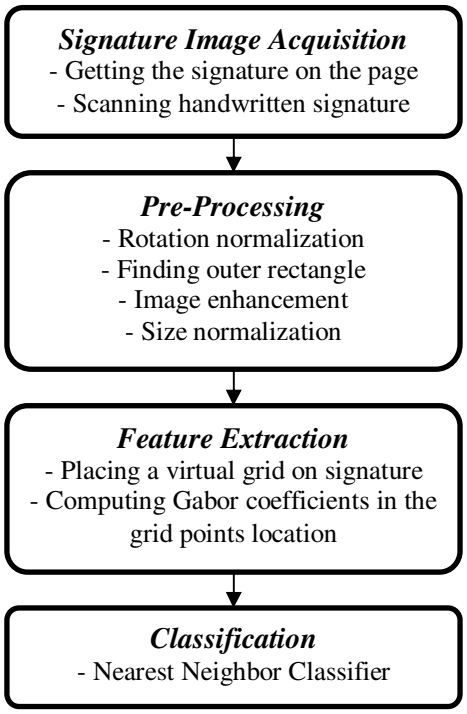


FIGURE 1: Flowchart of the proposed system.

4 PRE-PROCESSING

This is the first part of the proposed system, consisting of rotation normalization, determination of the outer rectangle of the signature, size normalization and finally image enhancement.

4.1 Rotation Normalization

In order to accomplish rotation normalization, the signature image contour is rotated in so far as the minimum inertia is located in the horizontal wise. This method has been presented by Kalera et al [6]. In this method the signature contour is indicated with C that comprises of N pixels.

$$C = \left\{ X(i) = \begin{bmatrix} u(i) \\ v(i) \end{bmatrix}, i = 1, \ldots, N \right\} \quad (1)$$

$X(i)$ = the vector comprising of x and y coordinates of the ith pixel of the signature contour

$u(i)$ = x coordinate of the ith pixel of the signature contour

$v(i)$ = y coordinate of the ith pixel of the signature contour

The $(\bar{u}, \bar{v})$ coordinates of the center of gravity of the signature contour are obtained according to (2) and (3).

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u(i) \tag{2}$$

$$\bar{v} = \frac{1}{N}\sum_{i=1}^{N} v(i) \tag{3}$$

The second order moment, $\overline{u^2}$ and $\overline{v^2}$ of the signature contour is then obtained according to (4) and (5).

$$\overline{u^2} = \frac{1}{N}\sum_{i=1}^{N} (u(i) - \bar{u})^2 \tag{4}$$

$$\overline{v^2} = \frac{1}{N}\sum_{i=1}^{N} (v(i) - \bar{v})^2 \tag{5}$$

The orientation of the minimum inertia axis is determined according to the orientation of the minimum eigenvector of the following matrix.

$$I = \begin{pmatrix} \overline{u^2} & \overline{uv} \\ \overline{uv} & \overline{v^2} \end{pmatrix} \tag{6}$$

4.2 Finding the Outer Rectangle

The outer rectangle is the smallest rectangle surrounding the signature contour. It is determined by applying a threshold on the horizontal projection and vertical projection of the binary image. Image binarization is done using Otsu [13] method. By finding the outer rectangle, signature image will be robust to displacement (shift).

4.3 Size Normalization

In so many signature images, the signature elongation is in horizontal or vertical direction. Considering this point, a method is presented for size normalization. In this method at first, the length and the width of the signature are computed and then the larger one is selected. A constant number is also chosen as normal size. Now the length and the width of the image will be changed in the case that the larger dimension will be equaled to this constant normal size. Therefore, in signature images with larger width than length, the normalization will be based on the width, and vice versa. The constant normal size in this paper is considered as 200 pixels.

4.4 Image Enhancement

The resulted binary image in previous section is employed for the image enhancement operation. The white signature contour is located in the black background in this binary image. At first, closing operation is applied on the complement of this binary image. Closing operation is one of the morphological operations including dilation and erosion. Unwanted gaps in the signature contour are removed by closing operation. Afterward all of the spot areas containing lower pixels than a specific number are omitted, whereby all of the probable noisy areas are deleted. This operation is achieved by detecting all of the white connected components in the binary image and counting their pixels.

In the enhanced gray-level image, the gray-levels corresponding to the white pixels in the binary image preserve their values and other pixels value are set to white gray-level.

5 FEATURE EXTRACTION

In the proposed system, Gabor wavelet is used as feature extractor. Initially, Gabor wavelet and its specifications is introduced and then the application of Gabor wavelet in the proposed system as feature extractor is explained.

5.1 Gabor Wavelet

Gabor wavelet is obtained by multiplying a sinusoid function with a Gaussian function in time domain. By convolving a signal with the Gabor wavelets, the frequency information of the signal nearer to the center of the wavelets is obtained. A one-dimensional Gabor wavelet is shown in (7):

$$W(x, x_0, \omega) = e^{-\sigma(x-x_0)^2} e^{-i\omega(x-x_0)} \tag{7}$$

In (7), x_0 is the center of wavelet, ω is the angular frequency ($\omega = 2\pi f$) and σ is the radius of Gaussian function.

Convolution of Gabor wavelet and a given function $g(x)$ is defined as follow:

$$C_{x_0,\omega}(g(x)) = \int_{-\infty}^{+\infty} g(x) W(x, x_0, \omega) dx \tag{8}$$

In general, the result of the convolution is a complex number that comprises of real and imaginary parts:

$$C_{x_0,\omega}(g(x)) = a_{real} + ia_{imag} \tag{9}$$

Gabor wavelet coefficients can be stated based on angle and magnitude or based on real and imaginary parts as follow:

$$a_{real} = |a| \cos \angle a \tag{10}$$

$$a_{imag} = |a| \sin \angle a \tag{11}$$

$$|a| = \sqrt{a_{real}^2 + a_{imag}^2} \tag{12}$$

$$\angle a = \arctan(a_{imag} / a_{real}) \tag{13}$$

In the above equations, a is the complex coefficient of Gabor wavelet, a_{real} and a_{imag} are real and imaginary parts of a ; $|a|$ and $\angle a$ are amplitude and angle of a .

In image processing, the two-dimensional Gabor wavelet transform is used. These wavelets are the result of the multiplication of a sinusoid function by the two dimensional Gaussian function. The sinusoid signal extracts frequency information corresponding to its frequency and the Gaussian function determines the region of effects of the sinusoid signal. Therefore, Gabor wavelet operates as like as a local edge detector. Larger wavelength of sinusoid will cause the wavelet to be more sensitive to the edges with larger width and vice versa. By increasing the length of the radius of the Gaussian function, frequency information related to the larger area of the image will be extracted. The two dimensional form of Gabor wavelet is as follow:

$$w(x, y, \theta, \lambda, \varphi, \sigma, \gamma) = e^{-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}} \cos(2\pi \frac{x'}{\lambda} + \varphi) \tag{14}$$

In which we have:

$$x' = x \cos\theta + y \sin\theta \tag{15}$$

$$y' = -x \sin\theta + y \cos\theta \tag{16}$$

By convolving the two dimensional Gabor wavelets with the image, wavelet coefficients can be computed. These coefficients will be in the form of a matrix, which each of its elements is a wavelet coefficient related to its corresponding pixel of the input image. The absolute value of the coefficients of pixels related to edges will be much greater. In the Gabor wavelet there are five control parameters: θ, λ, φ, σ and γ.

θ determines the orientation of the wavelet. This parameter rotates the wavelet around its center. The orientation of the wavelet specifies the angles of edges that the wavelet responds to them. In many cases, θ includes values between zero and π. As the symmetric property of the wavelet, θ values between π and 2π are redundant.

λ specifies the wavelength of cosine signal or in other words it specifies the frequency of the wavelet. Wavelets with larger wavelength are more sensitive to the gradual changes in the image and wavelets with smaller wavelength are more sensitive to the edges.

φ is the phase of the sinusoid. Generally, Gabor wavelets are based on the cosine or the sine waves. Here, cosine waves are real parts of the wavelet and sine waves are imaginary parts of it. In most of the researches, the phase is assumed to be zero or $\pi/2$. Thus, if the phase value is assumed to be zero and $\pi/2$, real and imaginary parts of the convolution are obtained, which are the parts of complex numbers.

σ denotes the Gaussian radius. The length of the Gaussian radius, determines the size of the region that should be affected by the convolution. This parameter is usually proportional to the wavelength, so we would have $\sigma = c\lambda$.

γ specifies the aspect ratio of the Gaussian. Generally, this parameter is set to 1.

As can be seen, the independent parameters of Gabor wavelet are the rotation angle (θ) and the wavelength (λ). Other parameters are usually set to their default values or determined based on independent parameters.

5.2 Gabor Wavelet Coefficients as Feature Vector

In the proposed system, features are extracted based on Gabor wavelet. As said before, each 2D Gabor wavelet can detect specific edges with respect to the direction of rotation angle and the wavelength of wavelet; therefore, Gabor wavelet is restricted by two factors:

- Direction of edge which is related to the rotation angle
- Width of edge which is related to the wavelength

In order to detect all of the edges in an image, many Gabor wavelets must be used with lots of rotation angles and wavelengths; but it is not practical. To overcome this issue, Gabor wavelet coefficients are only computed for limited number of rotation angles and wavelengths.

Selected rotation angles have to cover all of the degrees between 0 and 2π, uniformly. As the symmetric property of Gabor wavelet, Gabor wavelets with rotation angles between π and 2π are the same as their corresponding ones with rotation angles between 0 and π. For example, for given parameters, Gabor wavelet with rotation angle equal to $\pi/6$ is as the same as wavelet with rotation angle equal to $7\pi/6$. Accordingly the quantized rotation angles between 0 and π are sufficient to cover all of the angles. For example $0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{5\pi}{8}, \frac{3\pi}{4}, \frac{7\pi}{8}$ can be considered as quantized rotation angles. The more number of rotation angles, the more accuracy for edge detection and more computational complexity as a result.

The wavelengths are selected based on the application. The narrower edges can be detected by smaller wavelengths and vice versa. The number of wavelengths depends to the variety of edges in the image. The selection of different wavelengths results in multi-scale or multi-resolution processing.

The convolution of Gabor wavelet with all of the pixels of an image is computationally complex, especially when the number of rotation angles and wavelengths are increased. Furthermore due to too many computed coefficients, feature selection or dimensionality reduction methods might be needed. To overcome this problem, Gabor wavelets are applied on only a few points (pixels), instead of all pixels of the image. These points are selected uniformly over the image which they

are placed on a virtual grid. The distance between non-diagonal adjacent grid points are named as grid distance, which are equal to a constant value.
The main reason of uniform distribution of grid points is that we do not have any prior knowledge about the shape and structure of signatures. This will become significant when one of the purposes is developing such a system that deals with signatures from different nationalities.
As the Gabor wavelet is translation, rotation and scale variant, in pre-processing stage the signature image have to be normalized by translation, rotation and scale (size).
In the proposed method for extracting features, a virtual grid is placed on the signature image that is shown in Figure 2. Then Gabor coefficients are computed on cross points of virtual grid in given rotation angles and wavelengths. These Gabor coefficients form the feature vector. The cross point of virtual grid is named as feature point.

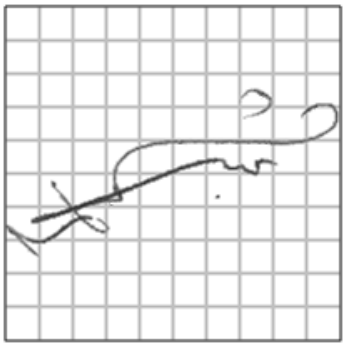


FIGURE 2: A virtual grid was placed on signature image.

Total number of features per signature image is pertaining to grid distance, number of rotation angles and wavelengths. For example, considering a signature image with 200x200 pixels and its grid distance equal to 20, the virtual grid consists of 9x9=81 feature points. By assuming 5 wavelengths and 8 rotation angles for Gabor wavelet, there are 5x8=40 coefficients per feature point, therefore the feature vector of each signature image comprises of 81x40=3240 coefficients totally.

The proposed method for feature extraction is independent of the nationality of signers. Unlike many other systems, the proposed system has high performance in identification and verification of signatures with different nationalities due to its independency of shape and structure of signature.

6 CLASSIFICATION

There are two different purposes in our system: (1) signature identification and (2) signature verification. Classifier type and classification procedure for each of these purposes are different from another, but both of them are based on distance.

In identification, classifier must determine the class of an input sample. In this case, the input of the system is a signature and the output is a class number that determines the class of input signature. In other word, the ultimate goal of identification is recognizing true class of an unknown input signature. To do this task, we used nearest neighbor classifier in our system. So, the class of input sample signature is same as class of the nearest training sample.

In verification, classifier must examine an input signature to determine whether it is genuine or not. Therefore, the input of a signature verification system has two parts: (1) a signature and (2) a claimed signer (class). Classifier must verify or reject claimed signer, whether is it genuine or forgery? In our proposed system for signature verification, classifier calculates distance of the input signature from all sample signatures of claimed class in the feature space. If minimum distance is less than a threshold, the input signature will be accepted; otherwise it will be known as a forgery signature and will be rejected.

There are many methods to calculate the distance between two points, for example: City Block (Manhattan) Distance, Euclidean Distance and Mahalanobis Distance. Euclidean Distance is a famous method that calculates the distance between two points $P_1=(x_1,x_2,x_3,\ldots,x_n)^T$ and $P_2=(y_1,y_2,y_3,\ldots,y_n)^T$ in n-dimension space by the formula:

$$D = (P_1 - P_2)^T (P_1 - P_2) \tag{17}$$

Mahalonobis Distance is a generalized form of Euclidean Distance that weighted each dimension of space by a matrix named *A*. *A* is a square and usual symmetric matrix.

$$D_A = (P_1 - P_2)^T A (P_1 - P_2) \tag{18}$$

Matrix *A* has two main effects on calculating this distance:
(1) Diagonal elements of matrix *A* change the weights of different dimensions, as weighty dimensions will have major effects. In other word, the points that have equal Euclidean Distance from an origin are on a hyper-sphere, whereas in Mahalanobis Distance, the points place on a hyper-ellipse. If matrix *A* is an identity matrix, hyper-ellipse will be converted to hyper-sphere and Mahalanobis Distance will be equal to Euclidean Distance. But if matrix *A* is a diagonal matrix, diameters of hyper-ellipse will be in parallel with the main axes of the space.

(2) In more general form, matrix *A* is a square matrix with non-zero values on non-diagonal elements. In this case, matrix *A* is same as the affine transform matrix. So, the points those have equal Mahalanobis Distances from an origin will be on a hyper-ellipse which can be rotated around some or all main axes.

In our proposed system, Mahalonobis distance is used for classification which *A* is a diagonal matrix. Matrix *A* must be computed for each class using training genuine samples, so for computing distance of an input sample from each class, we must use corresponding *A* matrix of that class. Because of applying diagonal condition on *A*'s, only the diagonal elements of matrix *A* must be computed and other elements are considered to be zero.

For a given class, if the samples in a given dimension are more distributed, these samples will have more variances in that dimension. Therefore, this dimension will not be a significant dimension and then, the corresponding diagonal element of matrix *A* will have small values. On the contrary, if the samples in a given dimension are more concentrated, this dimension will be an effective dimension and as a result, the corresponding diagonal element of matrix *A* will have bigger number.

7 EXPERIMENTAL RESULTS

In the signature identification, the system evaluation is determined by the correct classification rate (CCR). The accuracy of such a system is equal to the ratio of the number of correct identified signatures to the total number of signatures. More efficient systems results in closer value of CCR to one. Achieving the CCR=1 is difficult especially in a system with a large numbers of signers.
Unlike many biometric systems which is used for identification, that are evaluated by CCR, in the signature verification system, False Rejection Rate (FRR) and False Acceptance Rate (FAR) are two types of error rates and are used for evaluating the system. FRR and FAR are also named as Type 1 error and Type 2 error. FRR is related to the rejection of genuine signatures and FAR is pertaining to the acceptance of forgery signatures. In an ideal signature verification system, both of FRR and FAR must be approached to zero, but existing systems cannot achieve this purpose. Considering the application of verification system, a trade off should be determined between the FRR and FAR. Decreasing the FAR results in increasing FRR and vice versa.

In literatures another term is defined as the Equal Error Rate (EER). When system parameters are tuned in a way that the FRR is equal to FAR, this equal value is considered as EER. Usually EER is considered as the optimum state of the verification system.

7.1 Experiment Results on Iranian (Persian) Signature Database

To test our proposed system, we use a common Persian signature database, which contains 20 classes. There are 20 genuine and 10 expert forgery signatures per class. This database is available online at [14]. All of the signatures were signed by black pen on 10x10 cm white paper and scanned by MICROTEK ScanMarker 3630 at 300 DPI resolutions. The pre-processing stage was applied on all images. All of the algorithms were implemented in MATLAB 7.1 environment.
Except the λ parameter, other parameters were set to their default values as stated before. The λ as the main parameter is tested with different values. Five parameters of the Gabor wavelet were determined as follows.

- θ has to cover the angles between 0 and π degree. In the proposed system θ includes $0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{5\pi}{8}, \frac{3\pi}{4}, \frac{7\pi}{8}$.
- φ was set to 0 and $\pi/2$. 0 and $\pi/2$ refer to real and imaginary parts of the wavelet respectively.
- σ is usually proportional to the wavelength i.e. $\sigma = c\lambda$. In the proposed system *c* was set to 3.
- γ determines the aspect ratio of the mask, that was equal to 1 in order to form a square mask.

Two different value sets were investigated for λ parameter: $2\sqrt{2},4,4\sqrt{2},8$ and $4,4\sqrt{2},8,8\sqrt{2}$. By selecting these two sets, the effect of wavelength on the system performance can be examined. In addition, the grid distance is set to 10 and 20 in two different experiments.
Table 1 shows the CCR of the proposed system on the Persian signature database. In this test, there are 10 samples for training and 10 samples for testing.

TABLE 1: Signature identification results on Iranian (Persian) signature database

λ	Grid Distance	CCR (%)
$2\sqrt{2},4,4\sqrt{2},8$	10	97.5
	20	98.0
$4,4\sqrt{2},8,8\sqrt{2}$	10	99.5
	20	100

The EER values related to signature verification on the Persian signature database is illustrated in Table 2. In this experiment, 10 genuine signatures were used for training phase and 10 genuine and 10 expert forgery signatures were used for test phase. Experimental setup is similar to the real world conditions that there is no forgery samples for training phase and the system must be trained only by genuine samples.

TABLE 2: Signature verification results on Iranian (Persian) signature database

λ	Grid Distance	EER (%)
$2\sqrt{2},4,4\sqrt{2},8$	10	17.0
	20	19.5
$4,4\sqrt{2},8,8\sqrt{2}$	10	15.0
	20	17.0

According to the experimental results on the Persian signature database shown in Table 1 and 2, $4,4\sqrt{2},8,8\sqrt{2}$ were selected as λ values and the grid distance was set to 10.

7.2 Comparison With Other Methods

In this section, the proposed system is compared with some methods with different signature databases. These databases are from countries with different signature styles. The proposed system is also compared with the subjective method (human discrimination capability).

7.2.1 Persian Signature Database

The database that is used in this experiment is the same as the one that used in the previous section and contains 20 classes [14]. The CCR and EER of proposed system are 100% and 15% respectively.

The proposed system is compared with Mohamadi's [15] method, Kiani et al's [11] method, Pourshahabi et al's [12] method. Mohamadi [15], proposed a signature identification system based on PCA and MLP and achieved CCR equal to 91.5%. Kiani et al [11] presented a signature verification system which employed local Radon transform and SVM. In the average case FAR and FRR are 20% and 10.5% respectively. Pourshahabi et al [12] extracted features using Contourlet transform and classified them by Euclidian distance. CCR is equal to 100% in signature identification, FAR and FRR are 14.5% and 12.5% respectively in signature verification.

In Table 3**Error! Reference source not found.**, the comparison of the proposed system with the methods stated in [11], [12] and [15] are shown. The CCR of the proposed method is better than the CCR in [15] and is equal with the CCR in [12]. However the methods which presented in [11] and [12] have lower FRR compared to our proposed method, but the difference of EER between our proposed method and these methods is negligible.

TABLE 3: Comparison of proposed system with other systems on Iranian (Persian) signature database

Method	Signature Identification	Signature Verification		
	CCR (%)	FAR (%)	FRR (%)	EER (%)
Kiani et al [11]	-	20.0	10.5	15.25
Pourshabi et al [12]	100	14.5	12.5	13.5
Mohamadi [15]	91.5	-	-	-
The proposed system	100	15.0	15.0	15.0

7.2.2 South African Signature Database

This database contains 924 English signatures collected from South Africa which used in [9] in order to evaluate signature verification system. There are 22 classes in this database. There are 10 genuine signatures for training purpose, 20 genuine signatures, 6 simple forgery signatures and 6 expert forgery signatures in each class for test.

The proposed system achieved the EER rate equal to 6.3% and 16.8% for simple and expert forgery respectively. However the results show that the proposed system has higher error rate in simple forgery compared to the method presented in [9], but it is more reliable for expert forgery signatures.

Kiani et al [11] achieved the average FAR and average FRR equal to 0.5% and 42.7% respectively for simple forgery. In addition, the average FAR and average FRR of their system are equal to 12.1% and 42.7% respectively for expert forgery. Pourshahabi et al [12] reported 2.3% and 23.2% as the FAR and FRR respectively for simple forgery. In this system, FAR and FRR for expert forgery are 22.7% and 23.2% respectively. All of these results are summarized in Table 4. From Table 4, it is obvious that the proposed system is the most reliable system against expert forgery signature.

TABLE 4: Comparison of the proposed system with other systems on South African signature database

Method	Simple Forgery			Expert Forgery		
	FAR (%)	FRR (%)	EER (%)	FAR (%)	FRR (%)	EER (%)
Coetzer et al [9]	4.5	4.5	4.5	18.0	18.0	18.0
Kiani et al [11]	0.5	42.7	21.6	12.1	42.7	27.4
Pourshahabi et al [12]	2.3	23.2	12.75	22.7	23.2	22.95
The proposed system	6.3	6.3	6.3	16.8	16.8	16.8

7.2.3 Turkish Signature Database

This signature database is used by [5] and comprises 40 classes. There are 8 genuine signatures and 4 forgery signatures in each class. 30 different individuals other than genuine signers signed all of the forgery signatures.

Ozgunduz et al [5] presented a signature verification method by considering three types of features: (1) global features, (2) directional features, and (3) grid features. The FAR and FRR of this system are 11% and 2% respectively, while the proposed system achieved the FAR and FRR equal to 10% and 8% respectively. The results of the comparison are presented in Table 5.

TABLE 5: Comparison of the proposed system with the other system on Turkish signature database

Method	FAR (%)	FRR (%)	EER (%)
Ozgunduz et al [5]	11.0	2.0	6.5
The proposed method	10.0	8.0	9.0

7.2.4 Spanish Signature Database

Spanish signature database is collected by Frias-Martinez et al [4]. This database includes 228 signatures from 38 persons (6 signatures per class). They proposed a signature identification system based on SVM and compared that with similar system that used MLP. In the best case, their system can identify an input signature with CCR equal to 71.2%. In their experiment, only 1 training signature is used per class, therefore, it is very similar to the real world conditions. They concluded that the global features are better than raw bitmap features and SVM classifier has higher CCR compared to MLP. In the same experimental conditions, our proposed system could achieve higher CCR (77.3%) in comparison with the method presented in [4].

In another experiment, the proposed system is evaluated on this Spanish signature database by using more training samples. Table 6 shows the results of this experiment.

TABLE 6: Comparison of the proposed system with the other system on Spanish signature database

Number of training samples	CCR (%)	
	Frias-Martinez et al [4]	The proposed method
1	71.2	77.3
2	-	89.3
3	-	92.9

7.3 Comparison With Human Accuracy in Signature Identification and Verification

In this section, human accuracy in signature identification and verification is investigated. For this purpose, the Persian signature database, which was introduced in previous section, is used. As stated before, this database contains 20 classes and in each class, there are 20 genuine signatures and 10 expert forgery signatures. In the first experiment for signature identification, only genuine signatures are used: 10 signatures for training and 10 signatures for testing. In another experiment for signature verification, only 10 genuine signatures are used for training. 10 other genuine signatures and 10 forgery signatures are also used for testing phase. As you can see, in this experiment no forgery signatures were used for training similar to the real world conditions. For evaluating the human accuracy, 10 persons (25 to 36 years old) were invited to participate in these experiments.

In the first experiment for evaluating the human accuracy in signature identification, all of the training signatures were shown to each of the participant. Each participant could look at signatures without any time limitation. Moreover, during the testing phase, the participant could see the training signatures again. The participant had to identify the class of each test signature. There is not any specific order in displaying signatures to the participants and the signatures were selected randomly from different classes. All of the participants could identify signatures correctly, in other word the CCR of all 10 participants was 100%. Therefore, it can be concluded that the accuracy of the proposed system is the same as the accuracy of the humans.

In another experiment for investigating the human accuracy in signature verification, each participant can only look at 10 training genuine signatures pertaining to a specific class. Afterward genuine and forgery signatures corresponding to that class were randomly displayed to the participants in order to be accepted or rejected. This operation was repeated for all 20 classes. In Table 7 the FAR and FRR of each participant is shown.

As shown in Table 7, the average FAR and FRR of participants are 25.2 and 17.25 respectively. In the best case, the minimum FAR and FRR of participants are 22.5% and 15.5% that are related to person No. 3 and No. 8 respectively, while the EER of the proposed system is 15%. The average FAR and average FRR of the method presented in [11] are 20% and 10.5%. In addition, the FAR and FRR of Pourshahabi et al's [12] method are 14.5% and 12.5% respectively. These results show that the FAR of automatic systems and the humans are greater than FRR; therefore, it can be considered that the forgery signatures are expert type and are difficult to detect. In addition, with respect to the lower FAR of all automatic signature verification systems, it shows that these automatic verification systems are more accurate than the humans verification.

TABLE 7: Results of the human accuracy in signature identification and verification upon 10 subjects

Subject	FAR (%)	FRR (%)
Person 1	26.5	17.5
Person 2	28.5	18.0
Person 3	22.5	16.5
Person 4	23.0	19.5
Person 5	24.5	16.5
Person 6	27.0	16.0
Person 7	25.5	17.5
Person 8	24.0	15.5
Person 9	26.0	18.0
Person 10	24.5	17.5
Mean	25.2	17.25
Standard Deviation	1.86	1.16

8 CONCLUSION AND FUTURE WORKS

The algorithm, presented in this paper, is employing Gabor wavelet for feature extraction could achieve satisfying accuracy, although the simplest method i.e. nearest neighbor was used as the classification stage. As the pixel distribution of signature curvature is unknown overall image, unlike many current approaches, the proposed method is independent of the shape and the style of signature. The proposed system has higher performance in identification and verification of the signatures with different nationalities due to its independency of the shape and the structure of signatures. This is verified by testing the proposed system on 4 signature databases with different nationalities including Iranian (Persian), South African, Turkish and Spanish signatures. In addition, comparative experiments with 6 methods [4, 5, 9, 11, 12, 15] are presented. Even the system structure of the proposed method is simple; its accuracy is equal or even greater than the similar systems. According to another experiment, it was shown that the accuracy of our proposed system is equal to and greater than the human accuracy in signature identification and signature verification respectively.

Riesenhuber et al's [16] algorithm as a powerful method for object recognition is suggested for future works. This method is a hierarchical model inspired by the cortex structure of human's brain. The object recognition procedure in cortex employs a kind of hierarchical multi-resolution and -direction edge detection. This model is known as HMAX.

In the proposed system, the weighted distance was used in nearest neighbor classifier. It is suggested to use the other powerful statistical pattern recognition method such as SVM in classification stage.

9 REFERENCES

[1] Y. Gu, "*Approaching Real Time Dynamic Signature Verification from a Systems and Control Perspective*", M.Sc Thesis, University of the Witwatersrand, Johannesburg, 2003.

[2] Weiping Hou, Xiufen Ye, Kejun Wang, "A Survey of Off-Line Signature Verification", *International Conferenlce on intelligent Mechatronics and Automation*, Chengdu, China pp. 536-541, August, 2004.

[3] Edson J. R. Justino, Fla´vio Bortolozzi, Robert Sabourin, "A comparison of SVM and HMM classifiers in the off-line signature verification", *Elsevier Pattern Recognition Letters,* vol. 26, no. 9, pp. 1377-1385, 2004.

[4] E. Frias-Martinez, A. Sanchez, J. Velez, "Support Vector Machines versus Multi-Layer Perceptrons for Efficient Off-Line Signature Recognition", *Engineering Applications of Artificial Intelligence,* vol. 19, no. 6, pp. 693-704, September, 2006.

[5] Emre Ozgunduz, Tulin Senturk, M. Elif Karsligil, "Off-Line Signature Verification and Recognition by Support Vector Machine", *European Signal Processing Conference*, Antalya, Turkey, pp., September, 2005.

[6] Meenakshi K. Kalera, Sargur Sriharly, Alhua Xu, "Offline Signature Verification and Identification Using Distance Statistics", *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 18, no. 7, pp. 1339-1360, 2004.

[7] Peter Shaohua Deng, Hong–Yuan Mark Liao, Chin Wen Ho, Hsiao–Rong Tyan, "Wavelet–based Off–line Signature Verification", *Computer Vision and Image Understanding,* vol. 76, no. 3, pp. 173-190, 1997.

[8] Ben Herbst, Hanno Coetzer, "On An Offline Signature Verification System", *9th Annual South African Workshop on Pattern Recognition*, pp. 39-43, 1998.

[9] J. Coetzer, B.M. Herbst, J.A.Du Preez, "Offline Signature Verification Using the Discrete Radon Transform and a Hidden Markov Model", *Eurasip Journal on Applied Signal Processing,* vol. 4, pp. 559–571, 2004.

[10] Miguel A. Ferrer, Jesu´s B. Alonso, Carlos M. Travieso, "Offline Geometric Parameters for Automatic Signature Verification Using Fixed-Point Arithmetic", *IEEE Transactions On Pattern Analysis And Machine Intelligence,* vol. 27, no. 6, pp. 993-997, June, 2005.

[11] Vahid Kiani, Reza Pourreza, Hamid Reza Pourreza, "Offline Signature Verification Using Local Radon Transform and Support Vector Machines", *International Journal of Image Processing,* vol. 3, no. 5, pp. 184-194, 2009.

[12] Muhammad Reza Pourshahabi, Mohamad Hoseyn Sigari, Hamid Reza Pourreza, "Offline Handwritten Signature Identification and Verification Using Contourlet Transform", *International Conference of Soft Computing and Pattern Recognition*, Malacca, Malaysia, pp. 670-673, December, 2009.

[13] N. Otsu, "A Threshold Selection Method form Gray-Level Histograms", *IEEE Transaction on Systems, Man and Cybernetics,* vol. 9, no. 1, 1979.

[14] FUM-PHSDB: The FUM-Persian Handwritten Signature Database, Available on: *mvlab.um.ac.ir*, Last-Access: February 2011.

[15] Seyedeh Zahra Mohamadi, "*Static Persian Signature Recognition*", Bachelor of Science Thesis, Electrical Engineering Department, Ferdowsi University of Mashhad, Mashhad, 2006.

[16] Maximilian Riesenhuber, Tomaso Poggio, "Hierarchical Models of Object Recognition in Cortex", *Nature Neuroscience,* vol. 2, no. 11, pp. 1019-1025, 1999.

INSTRUCTIONS TO CONTRIBUTORS

The *International Journal of Biometric and Bioinformatics (IJBB)* brings together both of these aspects of biology and creates a platform for exploration and progress of these, relatively new disciplines by facilitating the exchange of information in the fields of computational molecular biology and post-genome bioinformatics and the role of statistics and mathematics in the biological sciences. Bioinformatics and Biometrics are expected to have a substantial impact on the scientific, engineering and economic development of the world. Together they are a comprehensive application of mathematics, statistics, science and computer science with an aim to understand living systems.

We invite specialists, researchers and scientists from the fields of biology, computer science, mathematics, statistics, physics and such related sciences to share their understanding and contributions towards scientific applications that set scientific or policy objectives, motivate method development and demonstrate the operation of new methods in the fields of Biometrics and Bioinformatics.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 5, 2011, IJBB appears in more focused issues. Besides normal publications, IJBB intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

LIST OF TOPICS

The realm of International Journal of Biometrics and Bioinformatics (IJBB) extends, but not limited, to the following:

- Bio-grid
- Bioinformatic databases
- Biomedical image processing (registration)
- Biomedical modelling and computer simulation
- Computational intelligence
- Computational structural biology
- DNA assembly, clustering, and mapping
- Fuzzy logic
- Gene identification and annotation
- Hidden Markov models
- Molecular evolution and phylogeny
- Molecular sequence analysis
- Bio-ontology and data mining
- Biomedical image processing (fusion)
- Biomedical image processing (segmentation)
- Computational genomics
- Computational proteomics
- Data visualisation
- E-health
- Gene expression and microarrays
- Genetic algorithms
- High performance computing
- Molecular modelling and simulation
- Neural networks

CALL FOR PAPERS

Volume: 6 - **Issue:** 1 - February 2012

i. Paper Submission: November 30, 2011 **ii. Author Notification:** January 01, 2012

iii. Issue Publication: January / February 2012

CONTACT INFORMATION

Computer Science Journals Sdn BhD
B-5-8 Plaza Mont Kiara, Mont Kiara
50480, Kuala Lumpur, MALAYSIA

Phone: 006 03 6207 1607
006 03 2782 6991

Fax: 006 03 6207 1697

Email: cscpress@cscjournals.org

CSC PUBLISHERS © 2011
COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA

PHONE: 006 03 6207 1607
006 03 2782 6991

FAX: 006 03 6207 1697
EMAIL: cscpress@cscjournals.org