I J C T A, 9(41) 2016, pp. 617-627 ISSN: 0974-5572

A NOVEL MODEL FOR FAST AND ROBUST RETRIEVAL OF3D BIO-IMAGES USING INTELLIGENT VISION ALGORITHM

Meenakshi Srivastava[*], Dr. S.K.Singh*, Dr. S.Q. Abbas[**] and Neelabh[***]

Abstract: Vast growth of multimedia information like images, audio and video on web has induced the scientific community for storage and retrieval of multimedia information more intuitively. Since one image can be interpreted in numerous ways, keyword based searching systems can retrieve the information to a limited extent. Moreover in few domains for example, as in the case of Bio Images, the content information is not even of textual nature. This limitation has led to the requirement for development of Content Based Image Retrieval Systems. Image Retrieval systems which can recognize the objects like the human eyes have been most popular. To construct a good intelligent vision system, high dimension feature vectors of images are required for pattern recognition. Though high dimension feature vectors provides good recall rate, they retard the search process speed. Search time duration increases further with increase in number of images in database. In the present manuscript authors have proposed a model which adequately addresses these constraints and provides a robust and fast retrieval mechanism of Bio Images.

Key Words: ***CBIR System, Protein Image Retrieval, modified fuzzy C-mean Clustering (MFCM), High Order Local Autocorrelation***

1. INTRODUCTION

Vast growth of multimedia information on web has induced the scientific community to develop system for storage and retrieval of multimedia information more intuitively. The research in last decade shows that retrieval of multimedia information system is hindered by the "semantic gap", residing due to use of metadata/ descriptors for explaining the existing contents in any multimedia object. Interpretation of an image can be done through various ways, there is no standardized terminology to represent the content of an image through keywords. Another way of representing any image is by computing the internal features either by automated technique or by

[*] *Amity Institute Of information Technology,Amity University, Lucknow, Uttar Pradesh, India*
msrivastava@lko.amity.edu, sksingh1@.amity.edu
[**] *Ambalika Institute Of Information Technology and Management, AITM, Lucknow, Uttar Pradesh, India*
sqabbas@yahoo.com

[***] *Department of Zoology (MMV), Banaras Hindu University, Varanasi, Uttar Pradesh, India*
srivastava.neelabh@gmail.com

semi-automated technique. In many fields of image retrieval automated analysis calculates some statistics which can further be used in the order to figure the associated content features [1]. Though such algorithms give required results and a satisfaction level in many fields has been achieved, yet a single technique that fits best in all sorts of user's requirements is still to be developed. Therefore, the doors are still open to invent new methodologies according to the requirements of image retrieval applications [1][2]. The requirement of user mainly depends on the context and the application scenario rather than the matching keywords only. So, it is very much required that the, multimedia processing systems understand, the content embedded in the multimedia data and recover the semantic structures as required by the end user. The conventional text based searching systems are able to facilitate retrieval in a limited sense. Due to this limitation, in past decades more attention has been directed towards techniques which can automatically recover the semantically meaningful structures from multimedia data like images. More emphasis has been put on retrieval algorithms which can recognize the objects like the human eyes. But as we live in 3D environment and our brain is trained to recognize the 3D objects by fetching the meaningful features from the object through our eyes. The development of modeling, digitizing and visualizing techniques for 3D shapes has additionally prompted to an expanding sum of 3D models, on web based domain-specific databases. This development has motivated the researchers to design and develop the searching algorithms which can query and find the required information from these domain specific databases. Various object recognition algorithms have been developed in recent years which follow the feature extraction and calculation of distance metrics in extracted feature to compute the similarity. The high-dimension of feature vector is another factor which decreases the speed of image clustering and image retrieval [32][29]. Generally, the representation of feature vectors which leads to efficient and fast execution of algorithms are chosen against the more complex and computationally time-consuming approaches.

Objective of present study is to develop an efficient system which will recognize the 3D image as human eyes does. In the present work, authors have focused on development of a new mechanisms which will aim at ‘‘querying’’ the databases of complex data sets such as bimolecular images by their content in terms of visual similarity, rather than by their textual annotations only. The prototype has been implemented as a ‘‘query-by example” system. The selected matching criteria are the visual features concerning the 3D images themselves. Authors approach complements the traditional approach of querying the database by textual key words only.

The rest of the paper is organized as follows. In section II has discussed various 3D object retrieval models and the present retrieval methodology being used for bio molecules. Section III explains the architecture of the proposed retrieval system and the algorithms used for feature extraction and object recognition. Experimental details are presented in section IV. Result Analysis is done in section V, conclusion is discussed in section VI.

2. STATE OF ART

The application of 3D (Three-dimensional) images has emerged in various domains like bioinformatics, medicines and drug designing, archeology, cultural heritage, computer-assisted design (CAD), 3D face recognition etc[33]. Many domains have their own 3D repositories such as the national design repository for CAD models [1], Protein Data Bank for biological macromolecules [2], CAESAR for Anthropometry [3], the AIM@SHAPE shape repository [4] and the INRIA-GAMMA 3D Database [5] for research purposes. The experimental search engines for 3D shapes, i.e. the 3D model search engine has been developed at Princeton University [6]. Later on the 3D model retrieval system at the National Taiwan University [7], the Ogden IV system at the National Institute of Multimedia Education, Japan [8,9], the 3D retrieval engine at Utrecht University [4, 10], and the 3D model similarity search engine at the University of Konstanz [3, 12]

were developed. Though the retrieval of 3D images has been successful in many domain still the field of biology, retrieval of bio molecule images was deprived of the desired attention from the computer scientist. In biology, proteins are one of the most complex bio-molecules one can think of [13]. Studying the spatial arrangement of each atom inside the protein and also understanding and analyzing the bonding between the atoms can be a very tedious job. Computational Biologists and Bio- Bionformaticians have in their arsenal some useful bioinformatics tools which make this tedious job a bit easy. But having said that does not mean that we are at the zenith of the field of proteomics, the goal still being farfetched [14].In the field of bioinformatics protein structure has many fold effects starting from tools for predicting the 3D structure of protein and visualizing a protein through various possible ways to matching the protein structure for finding the similar proteins. Searching the protein structure similarity is very complex and time-consuming job because the construction of proteins is complex on its own. Existing methods in the area, like VAST [15], DALI [16], CE [17], LOCK2 [18], PFSC [19], FATCAT [20], FAST [21] and others, usually represents the complex protein structure in simpler form and then seek similarities using a pair-wise comparison of the given molecule to the subject molecule from a database[22].Another significant problem is a rising number of protein structures in world-wide repositories, like the Protein Data Bank which slows down the process of similarity searching since every time there are more molecules to compare in the database[22].Even though this problem is of keen interest to researchers the authors could not find any relevant portal which can search and retrieve the PDB files according to their visual similarity in efficient manner.

3. PROPOSED MODEL

One of the problems that is more frequent in the field of Bioinformatics is, answering the questions like “Which protein entities have the same structure as the given one?” This challenge still exists due to the reason discussed in earlier section, the conventional protein searching algorithms depend upon structure alignment based pair-wise comparison methods. These methods follow the two step process that is (i) find the best alignment between the two structures (ii) Compute the Root Mean Square Deviation (RMSD) between the core atomic positions, e.g., alpha carbon coordinates, of the aligned proteins. However, most of methods based on structural alignment cannot be used for protein structure search against large database, since it is computationally expensive to compute their similarities [24]].Our research aims to develop a Content-based image retrieval (CBIR) model for bio images for fast retrieval of “*look alike”* protein images. Various visualization of protein images like backbone, ribbons, rockets, etc can be obtained by using 3D molecular design software. Since proteins have a complex structure, relying on a single representation may be inadequate, so this algorithm works on computing the similarity on all these three visualization methods.

3.1 System Architecture

The proposed model takes into consideration the fact that Bio-Bionformaticians can easily recognize two similar proteins by viewing the 3D structure of protein from various angels/ views. The prototype of the proposed model has been implemented and tested on PDB (Protein Data Bank)[22] files. Our work focuses on retrieval of similar protein images by automatic feature extraction method. Authors have structured model (Fig 1) into five blocks (i) Query Interface block (ii) Geometrical Feature Extraction block (iii) Object Recognition block (iv) Image Database block and (v) Image Retrieval and Visualization block.

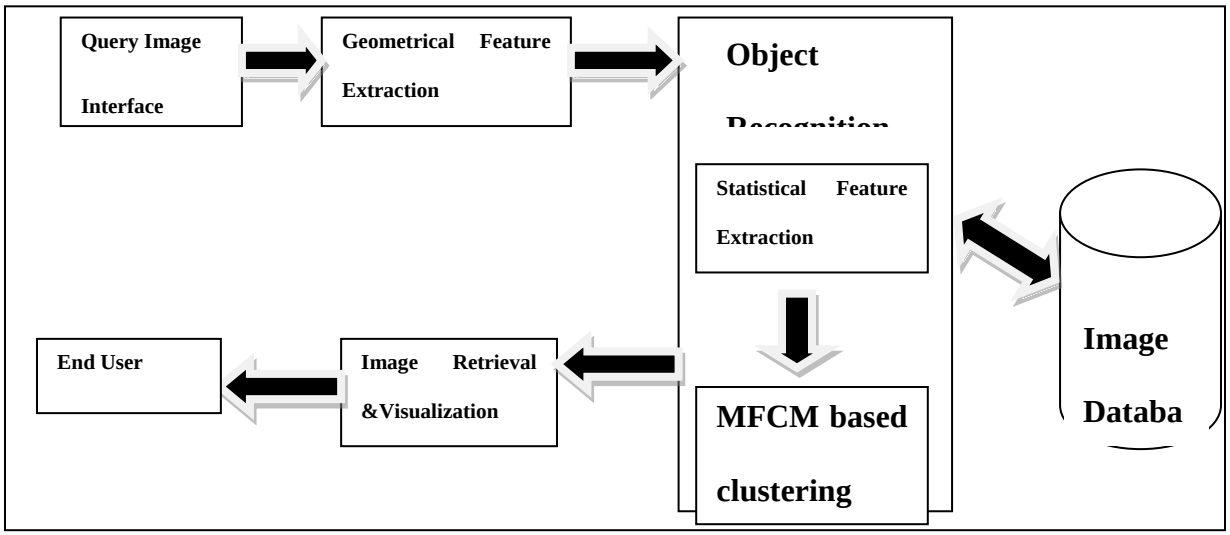
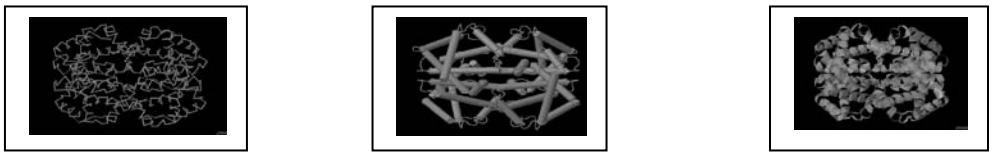


Fig 1. The system architecture, containing the five blocks: Query Image Interface block, Geometrical Feature Extraction block, Object Recognition Block, Image Database block, Image retrieval and visualization block.

3.2 Query Image Interface Block:

Query Image interface has been implemented as Query by Example interface. Depending upon the requirement of the user input query image can be represented by using only one visualization method, or combination of any two visualization method as well as by using the combination all the three visualization method. Present paper has discussed the query image representation and retrieval based on combination of three visualization method. In Fig. 2 various visualization of the query image (PDB ID 4HHB) have been shown.



(a) Backbone (b) Rocket (c) Ribbon

Fig. 2 Various Visualizations for Example Query Image (PDB ID- 4HHB)

3.3 Geometrical Feature Extraction Block:

The size of protein image has been fixed to 256 x 256 pixels using JMOL software [27].For geometrical feature extraction an intelligent vision algorithm proposed in [25][26]] has been deployed. Geometrical Feature Extraction concerns the extraction of features which are invariant under some transformation group acting on pattern [26]. The primitive features for an intelligent vision must be-

a. **Shift Invariant –** If z |f| denote a feature of image f(r) which is extracted over plane P. The shift invariant feature x|f| will be represented as,

$$x[T(a)\, f] = x[f| \text{ for } \forall a;\ \text{Supp}\,(T(a)\, f) \subset P. \quad (1)$$

The autocorrelation function is shift Invariant, and its N order with N displacement (a1,a2...., aN) is defined as follows:

$$x(a1, .., aN) = \sum I(r)I(r + a1)...I(r + aN) \quad (2)$$

Where r is the image coordinate vector. The order N is limited to the second order ($N \in \{0, 1, 2\}$).

b. **Additive -** z[f] is additive, is represented by

$$X[f1=f2]=x[f1]=x[f2] \text{ for Supp } (f1) \cap \text{Supp}(f2)=\emptyset. \quad (3)$$

Each supplied query image is rotated randomly around its three principal viewing axes and multiple-views of 2D images are stored. 2D HLAC (High Order Local Autocorrelation) features [28] are extracted from the query images. Duplicate configurations of r, r+a1, ..., r+a N are removed and local mask patterns are reduced to 35. Fig. 3 shows images generated through rotating the protein image on X,Y and Z axis using JMOL [27] in multiple visualization scheme. The combined HLAC features produce a 105-dimensional HLAC feature vector.

3.4 Object Recognition Block:

Object recognition is the next step after geometrical feature extraction, and it deals statistical feature extraction [29]. After the extraction of features, the next problem is how to use combine those primitive features for pattern recognition, adaptability and trainability. This leads to use of multivariate data analysis methods. In multivariate data analysis methods, new features are given by linear combinations of the primitive features.. Since a good pattern recognition algorithm uses large training set so that it could generate good classification result 12,000 images for each protein were sampled. The high-dimension feature vector leads to decrease the speed of image clustering and image retrieval. To reduce the dimension of feature vector Principal Component Analysis is performed on the HLAC feature vector. PCA performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. While computing the PCA first the covariance/ correlation matrix of the feature vector is constructed and then Eigen vectors on covariance matrix are computed. The principal components are those Eigen vectors that correspond to the largest Eigen values. Practically it has been seen that the first few Eigen vectors can be used to represent the system. The clustering is performed through modified fuzzy C-means (MFCM) clustering index algorithm [29].

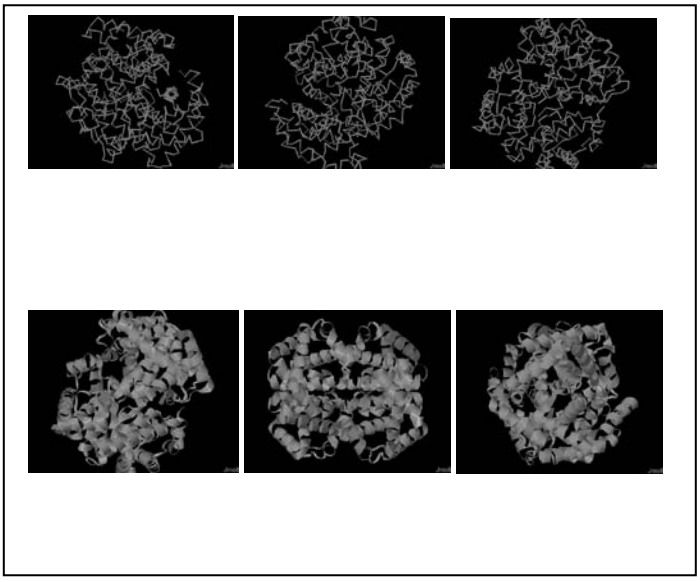


Fig. 3 Multi Views of the Protein image used in the Example

3.5 Image Database Block:

100 proteins were randomly collected from the RCSB PDB [[3]] using JMOL [27].MFCM [29] algorithm has been used to cluster the high dimensional data in the database. Suppose if there are Q *protein images* in the database, described by Np which is a high-dimension feature vector. By applying MFCM database can be classified into *K* clusters and each cluster will be represented by a clustering center. Each image will be assigned to a cluster where the distance between image and clustering center is the closest of all [32]. After classification, the similar images will be assigned to the same cluster. The most important step during the whole progress of MFCM algorithm is selection of reasonable initial center. The number of initial centers is selected as the formula defined in [30][32].

$$K = \max(\sqrt{X}, N) \quad (4)$$

If

$$\sum_{i=1}^{k} D(Xi, Xk+1) = \max \left\{ \sum_{i=0}^{k} D(Xi, Xj), j = 1,2, \ldots, Q-k \right\} \quad (5)$$

Then

$$C_{k+1} = X_{k+1}$$

Here, *k C* represents the k^{th} initial center, *i X* is the *i*th protein's image feature vector, D(Xi, Xj) is the distance between the two protein image feature vectors *Xi* and Xj [32].

3.6 Image Retrieval& Visualization Block:

Retrieval speed is an important issue in image retrieval system, while image information is abundantly, and the data of image database is large [29][32]. In order, searching of image database always results in enormous computations which are very time consuming. By classifying all images in database according this indexing structure, image searching scope has been reduced greatly which has fastened the retrieval rate. One more advantage of clustering index scheme is that the increasing number of feature dimensions and the size of image database time of image retrieval won't increase in linear proportion.

The retrieved image set has been visualized and shown to the end user in decreasing order of similarity index.

$D = min\ (D\ (Xq, Ck)k = 1,2,\ldots,K$

$Xq\ clust\ [k] \qquad q = 1,2,\ldots.,Q \quad (6)$

Here,

Cluster No = $clust[k]$

Size of the *k* th cluster center = $count[k]$

4. EXPERIMENTAL DETAILS

The effectiveness of the proposed method has been evaluated on bio molecular images collected randomly from the RCSB PDB[15]. 200 proteins were collected, Jmol[27] was used to synthesize the protein, each protein image has been rotated on its prime axis and multiple views of protein images have been stored. 4000 images for each visualization type have been synthesized total 12000 images at 256 x 256 pixels for each protein have been acquired. The color scheme has been set to gray-scaled.

HLAC features have been extracted from the multi view images. In the Nth order displacement the value of N has been fixed to 2. Combined High Order Local Autocorrelation feature vector of 105 dimensions has been generated. On image feature matrix x[M x N], eigenvector corresponding to the absolute maximal eigen value of the covariance of matrix $x[M \square\square N]$ has been calculated. The dimensions of the reference Subspace (database) and the Input subspace (Query Image) were varied .MFCM clustering index has been applied to the high dimensional feature vector.

Prototype of the model is implemented in MATLAB [31].

5. RESULT ANALYSIS

In order to test the validity and efficiency of the proposed algorithm MFCM method is compared with other retrieval methods, including c-means clustering and MSMbased clustering.

Table 1. Comparison of Precision and Retrieval time

	C-means Clustering	MSM based k-means clustering	MFCM Clustering
Precision	95.21	98.78	96.19
Recall	96.34	97.23	98.23
Retrieval time (ms)	208.4	176.32	90.73

Table1 shows that compared with in order retrieval, MFCM has the fast retrieval speed than C-means clustering and

Mutual Sub Space based k-means clustering.98 % recall rate and 96 % precision rate at a very fast speed was achieved.

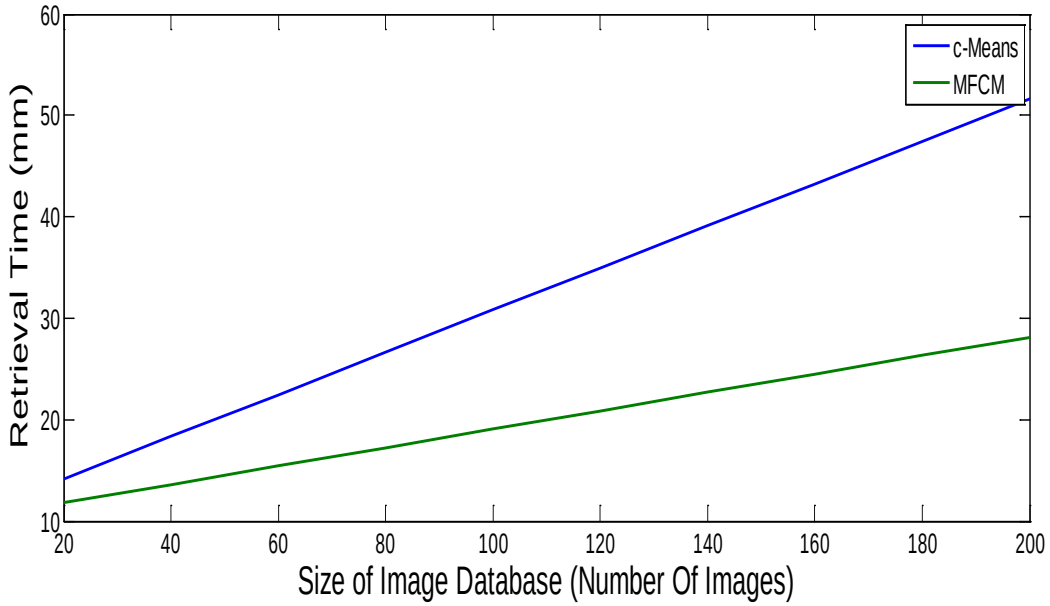
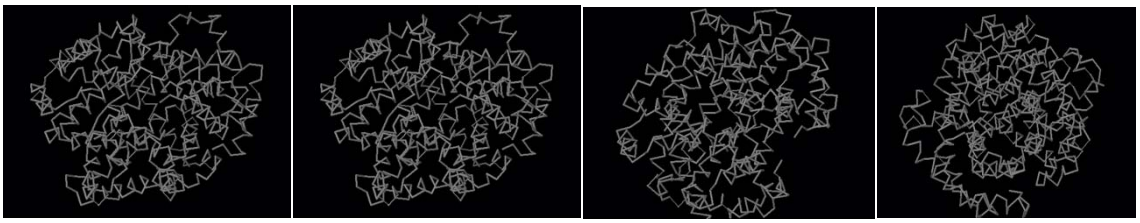


Fig4. Retrieval comparison MFCM and C-means

As discussed earlier, one of the main problems associated with retrieval of protein images is the regular growth of new proteins in PDB database. Our results show that if the number of images is increased in the database then also the time of retrieval doesn't increase linearly with them. In Fig. 4results of retrieval time contrast has been shown.

Two case of result retrieval have been given in Fig. 5 and Fig 6.

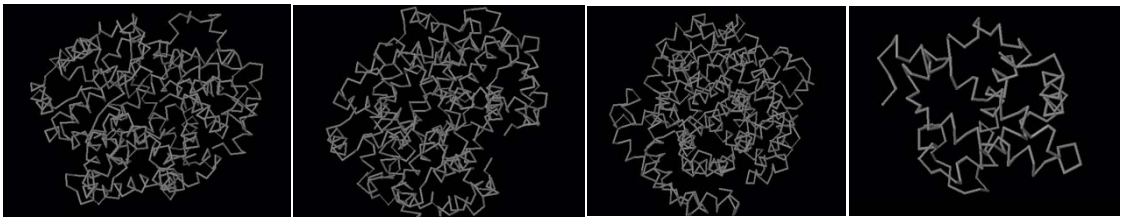
In first case the query Image 1A3N was already existing in the database, so it was retrieved at the first place. Fig 5 (a) is the query image and Fig5 (b) (c) (d) are the retrieved result.



(a) PDB ID 1A3N **(b) PDB ID 1A3N** **(c) PDB ID 1GZX** **(d) 1HBB**

Fig. 5 Retrieval Results (Query Image was Present in the database)

Now at second time the image 1A3n was intentionally removed, the retrieved results have validated our first phase retrieval.



(a) PDB ID 1A3N (c) PDB ID 1GZX (c) PDB ID 1HBB d) PDB ID1MBN

Fig. 6 Retrieval Results (Query Image was not Present in the database)

6. CONCLUSION AND FUTURE WORK

Model proposed can efficiently retrieve the bio molecular images. Prototype model has been implemented in MATLAB for protein images. Approach is based on retrieving the images which are visually similar. The conventional protein similarity algorithms suffers from the disadvantage that they highly depends on the backbone length of the protein structures, which makes it sensitive to local error due to non-optimal alignment. Adopted approach has overcome this drawback as has used intelligent vision algorithm which extract high level autocorrelation features from the images. Another problem associated with protein similarity search, is the regular growth of proteins in the protein databank, and as the growth increases the speed of retrieval decreases. This problem has been overcome by applying MFCM clustering index algorithm. Experimental results show that increase in number of images in data base does not linearly effect the retrieval time. Presently performed tests are on very small data set, in the future work more PDB data will be collected and clustering on them will be performed. A web based portal will be developed which can retrieve 3D protein images based on their visual similarity.

References

[1] Mussarat Yasmin, Sajjad Mohsin, Muhammad Sharif, Intelligent Image Retrieval Techniques: A Survey ,Journal of Applied research and technology, Vol 14 December 16.

[2]` `H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourrne. The Protein Data Bank model search engine *Nucleic Acids Res.*, 28:235–242, 2000.3D

[3] http://shape.cs.princeton.edu.

[4] V. P. Min, J. A. Halderman, M. Kazhdan, and T. A. Funkhouser. Early experiences with a 3D model search engine. In *Web3D Symposium*, pages 7–18, March 2003..

[5] https://www-roc.inria.fr/gamma/gamma/Logiciels/index.en.html

[6] Shilane P, Kazhdan M, Min P, Funkhouser T (2004) The princeton shape benchmark. In: Proc. shape modeling International Conference2004, pp 157–166.

[7] Shape modeling international 2004, pp 157–166P. Kumswat, Ki. Attakitmongcol and A. Striaew, "A New Approach for Shen Y-T, Chen D-Y, Tian X-P, Ouhyoung M (2003) 3D model search engine based on

[8] Shen Y-T, Chen D-Y, Tian X-P, Ouhyoung M (2003) 3D model search engine based on light field descriptors.In: Proc. eurographics 2003.

[9] Kazhdan M, Funkhouser T, Rusinkiewicz S (2004) Shape matching and anisotropy. In: Proc. SIGGRAPH 2004

[10] Körtgen M, Park G-J, Novotni M, Klein R (2003) 3D shape matching with 3D shape contexts. In: Proc. 7th central European seminar on computer graphics, April.

[11] Mamic G, Bennamoun M (2002) Representation and recognition of 3D free-form objects. Digit Signal Process 12:47–67

[12] Min P (2004) A 3D model search engine. PhD thesis, Princeton University.

[13] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). The shape and Structure of proteins.

[14] Blueggel, M., Chamrad, D., & Meyer, H. E. (2004). Bioinformatics in proteomics. *Current Pharmaceutical biotechnology*, *5*(1), 79-88.

[15] Gibrat J.F., Madej T., Bryant S.H.: Surprising similarities in structure comparison.Current Opin Struct Biology 6(3), pp. 377–385 (1996)

[16] Holm L., Sander C.: Protein structure comparison by alignment of distance matrices. J Mol Biol. 233(1), pp. 123–138 (1993)

[17] Shindyalov I.N., Bourne P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11(9), pp. 739–747 (1998)

[18] Shapiro J., Brutlag D.: FoldMiner and LOCK2: protein structure comparison andmotif discovery on the web. Nucleic Acids Res. 32, pp. 536–541 (2004)

[19] Yang J.: Comprehensive description of protein structures using protein folding shape code. Proteins 71(3), pp. 1497–1518 (2008)

[20] Friedberg I., et al.: Using an alignment of fragment strings for comparing proteinstructures. Bioinformatics 23(2), pp. 219–224 (2007)

[21] Zhu J.H., Weng Z.P.: FAST: A novel protein structure algorithm. Proteins 58, pp.618–627 (2005)

[22] Alina Momot, Dariusz Mrozek, Sylwia Górczyńska- Kosiorz, Michal Momot Improving Performance of Protein Structure Similarity Searching by DistributingComputations in Hierarchical Multi-Agent System. Conference Paper · January 2010.

[23] Berman H.M., et al.: The Protein Data Bank. Nucleic Acids Res. 28, pp. 235–242 (2000).

[24] Sungchul Kim,Postech ,York Korea, Fast Protein 3D Surface Search, *ICUIMC(IMCOM)'13*, January 17-19, 2013, Kota Kinabalu, Malaysia Copyright 2013 ACM 978-1-4503-1958-4.

[25] Motofumi T. Suzuki, Texture Image Classification using Extended 2D HLAC Features,KEER2014, LINKÖPING | JUNE 11-132014 International Conference On Kansai Engineering And Emotion Research Texture

[26] Nobuyuki OTSU and Takio Kurita, A New Scheme for Practical Flexible And Intelligent Vision Systems, lAPR Workshop on CV -Special Hardware and Industrial Applications OCT.12-14, 1988, Tokyo.

[27] A. Herr´aez. Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education*, 34(4):255–261, 2006.

[28] H. Sakano and S. T. Classifiers under continuous observations. *Lecture Notes in Computer Science*, Volume 2396/2002:798, 2002.,

[29] Liu Pengyu, Jia Kebin, Lv Zhuoyi, An Effective and Fast Retrieval Algorithm for Content-based Image Retrieval, 2008 Congress on Image and Signal Processing

[30] Kouassi R., Gouton P., Paindavoine M.: "Approximation of the Karhunen- Loeve transformation and its application to color images", Signal Processing: Image Communication, vol. 16, 2001, pp. 541-551.

[31] MATLAB and Statistics Toolbox Release 2014b, The MathWorks, Inc., Natick, Massachusetts, United States.

[32] Zhuoyi Lv. "An Effective and Fast Retrieval Algorithm for Content-Based Image Retrieval", 2008 Congress on Image and Signal Processing, 05/2008

[33] Meenakshi Srivastava, Dr. S.K. Singh, Dr. S.Q. Abbas, “Web Archiving: Past Present and Future of Evolving Multimedia Legacy”, International Advanced Research Journal in Science, Engineering and Technology Vol. 3, Issue 3.