

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

Adopting Biological Sequence Alignment Tools in 2D Shape Recognition

Mohammed Naseeb VP¹, Muzameel Ahmed²

P.G. Student, Department of Information Science Engineering, DSCE, Bangalore, India¹

Research Scholar, JAIN University, Bangalore, India²

Associate Professor, Department of Information Science Engineering, DSCE, Bangalore, India²

ABSTRACT: Recognition of two dimensional shapes is without doubts an significant and still open research part in computer vision and pattern recognition .Adopting biological alignment tools into this problem, sequence alignment in biological field is already investigated by different techniques. Human vision reveals that the shape feature is a significant perception to differentiate different objects. This stimulates researchers to find consistent and efficient shape features for discerning objects by machine vision. Many methods have been suggested in the past , many of them based one on features extracted from the boundary: actually, shape contours have shown to be very sensitive in many perspectives. In the proposed method, we encrypt shapes as biological sequences, using standard and well recognized sequence alignment tools to invent a similarity score, finally used in a nearest neighbour situation.

KEYWORDS: Shape recognition,binarisation, Feature extraction, chain codes,and Alignment tools.

I. INTRODUCTION

Image Recognizing shape is very important in pattern and computer vision, many different techniques are already proposed in image processing. Most of the techniques are based on boundary features only image contours will give more expressive features for an shape.

In this paper the detection of shapes are depends on the contours uses the techniques that are used in biological field efficiently .In biological field the pattern recognition is already become successful the interaction between bioinformatics and shape recognition are unidirectional here we are using a different way for the interaction , by deploying advanced bioinformatics solution into pattern recognition to solve the shape recognition problem bioinformatics application like sequence modelling, phylogeny, database searches are already successfully evaluated so this type of bioinformatics solution we can deploy for shape recognition problem in image processing system

There are already has many interesting solutions used for different field like where internet videos were encoded as “video DNA sequences” and analysed with phylogenetic related tools, this paper utilizes big amount of work successfully evaluated in bioinformatics field

Bioinformatics uses a important technique called sequence alignment we utilizes this technique for recognizing 2d shapes it is already well researched in biological field in case of amino acids sequence, bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of likeness that may be a concern of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that duplicate or like characters are aligned in consecutive columns. Sequence alignments are also used for non-biological sequences, such as those present in natural language or in financial.

In this project the first step is to change the given shape into sequence form by extracting the features from the boundary ,we are using freeman chain code algorithm for converting image contours into sequence form then we can use any altered sequence alignment tools for receiving the best alignment from the given sequences the popular sequence alignment approaches are smith water man’s and Needleman Wunsch algorithm, here the smith water man is

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

largely used for finding the similar parts of two objects. Needle man Wunsch algorithm used to find out the finest alignment between two sequences

The Needleman–Wunsch algorithm is an algorithm in bioinformatics to align protein or nucleotide sequences it is very basic pair ways sequence alignment tool. It was one of the first applications of dynamic programming to relate biological sequences. This algorithm is established by Christian D. Wunsch and Saul B. Needleman. The algorithm mainly divides a big problem (e.g. the whole sequence) into a sequence of smaller problems and uses the answers to the smaller problems to reconstruct a solution to the bigger problem. It is also sometimes referred to as the optimum matching algorithm and the global alignment method.

Applications of offline Two dimensional shape recognition are

- Quality control and assembly in industrial plants.
- Recognize faces and voices
- Monitoring and surveillance
- Identify DNA profiles
- Recognize handwritten characters

II. BIOLOGICAL SEQUENCE ALIGNMENT

In Biological sequence alignment, a sequence alignment is a way of positioning the sequences of DNA or protein to recognize regions of similarity that may be a result of functional, structural, or relationships between the sequences. Well aligned sequences of nucleotide and amino acid residues are normally express as rows in a matrix. Other gaps are inserted among the sequences so that similar symbols are aligned in consecutive columns. It is are also used in other than biological sequences, those present in natural language.

Sequence alignment is a typical method in bioinformatics for expressing the relationships between sequences in a collection of structurally related proteins .Here a set of sequences of proteins are compared, after alignment displays the sequences for each protein symbols on one line, having gaps (“-”) introduced such that similar sequence appear in the similar column.

In each situation, an alignment delivers a bird’s eye vision of the underlying evolution, structural, and functional restraints characterizing family of a protein in a concise.

Importance of sequence alignment

The first point of biological sequence examination, in bimolecular sequences (DNA, RNA,)high sequence relationship usually indicates important functional or structural similarity.Fig.1 shows the structural representation for DNA it indicates that sequences present in the DNA are aligned in a particular order.

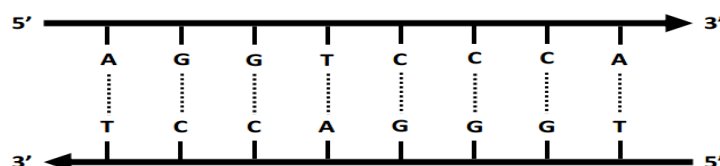


Fig.1 DNA format

Duplication with the modification, the huge majority of existing proteins are the result of a constant series of genetic repetitions and successive modifications. As a outcome, redundancy is a in-built characteristic of protein sequences or amino acid sequences, should not be surprised that numerous new sequences look like already known sequences everything in biology is based on huge redundancy.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

III. LITERATURE SURVEY

M. Bicego, A. Martins proposed a different method for contour-based 2D shape recognition is suggested, by a class of information theoretic kernels lately presented. This type of kernels, centered on a non-extensive generalization of the traditional Shannon information theory, is explained on probability measures. These kernels were used in text classification tasks, by presence exactly applied to different types of multinomial diagrams of the texts: relative term frequencies (also known as bags of words) and statistics (relative frequencies of subsequences of n symbols). It verified different information theoretic kernels, alternating from the classic Jensen-Shannon divergence kernel, to different versions of the recently introduced Jensen-Tallis kernels. These kernels are used in support vector machine and the nearest neighbour classifiers.

L. Chen, R. Feris, and M. Turk [2] proposed an efficient technique to find only the similar portions of two shapes and calculate the likeness of the shapes using only similar portions. Here the main contribution is to suggest a local shape identifying technique based on the Smith-Waterman algorithm [14] to efficiently find common portions of two shapes without needing exhaustive search for all likely parts. It explains that the partial shape matching method, which is based simply on common portions of two shapes, is robust to shape alterations, with rotation, scaling, and distortion and increase the recognition rate meaningfully.

R. Huang, V. Pavlovic [7] Propose a novel shape representation outline based on Problem Hidden Markov Models (PHMMs). PHMMs are powerfully linear, left-right HMMs, thus can model a shape extra specially than normal HMMs. This latest architecture comprises *insert* and *delete* states, in addition to the regular similar states, resulting in robustness to significant shape contour perturbations, containing rigid and non-rigid deformations, occlusions and omitted contour parts. At the similar time, the adopted framework points to a computationally effective set of algorithms for detecting shape, classification and segmentation.

S. Madeira and A. Oliveira [11] propose a huge number of clustering approaches for the analysis of gene appearance data acquired from microarray experiments. Though, the results of the application of typical clustering methods to genes are restricted. These limited results are compulsory by the presence of a number of experimental situations where the movement of genes is uncorrelated. A common limitation exists when collecting of conditions is performed. For this reason, algorithms in many numbers that perform concurrent clustering on the row and column measurements of the gene expression matrix has been planned to date. This concurrent clustering, usually chosen by biclustering, pursues to find sub-matrices, that is subclasses of genes and subgroups of columns, wherever the genes exhibit highly correlated actions for each condition.

S. Needleman and C. Wunsch [13] propose computer flexible method for discovering comparisons in the amino acid sequences of proteins are developed. From these results it is potential to determine whether any significant homology occurring between the proteins. This evidence is used to trace their probable evolutionary development. The extreme match is totally dependent upon the match of the sequences. One of its meanings is the biggest number of amino acids of one protein that can be compared with those of a second protein letting for all possible interruptions in whichever of the sequences. While the disruptions give increase to a very large number of differences, the method efficiently rejects from attention of those comparisons that cannot donate to the extreme match.

IV. PROPOSED METHODOLOGY

Our proposed methodology for shape recognition goes through the following four steps.

- Data acquisition
- Pre-processing
- Feature extraction
- Sequence Generation
- Sequence Analysis

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

First step is to collect different Two dimensional shapes and resize the image, Once the image is resized next it has to undertake pre-processing stage, where the image is converted into gray scale image. For additional process the image has to be converted into binary design which contains of zeros and ones. The collected images were stored in .bmp format fig.2 show the some sample images



Fig.2 Sample dataset

In pre-processing the input image has to be pre-processed, so its quality gets increase, and features can be easily identified. Initially the image has to be converted into grey scale, from there to Binarization.

Object boundary detection lies in a very trivial concept that at the boundary of an object there is a sudden abrupt change in the intensity value as we just cross the boundary(Fig.3). We make use of this elegant concept. What we do is take any pixel and traverse to its corresponding neighbour pixel and see the change in the intensity values during such travel. If the change in intensity is too high, we assign the pixels from which we traverse as the boundary pixels Fig.4.



Fig. 3 Feature extraction

Here we need to convert the extracted feature (here boundary) to the sequence representation of the same, it is carried out by using chain code generation technique called Freeman chain code. There are two different type Freeman chain code 4-connectivity and 8-connectivity. In 4-connectivity we loss diagonal points So here we use 8-connectivity Freeman chain code.

A boundary code shaped as a sequence of such directional numbers or characters are referred to as a Freeman Chain Code. The sequence or chain code of a boundary depends on the starting point. Functioning with sequence numbers offers a unified way to analyse the shape of the boundary. Chain Code or sequence tracks the contour in boundary clockwise method and keeps track all the directions of a boundary as we go from first contour pixel to another fig.4.

Chain codes have been requested as one of the methods that are able to recognize many things like characters and digits effectively. This is since of several advantages controlled by this method. The first advantage completed the representation of a similar shapes is that the chain sequences are a solid representation of a binary shape. Second, the

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

chain codes are a conversion invariant representation of binary shapes. Due to that, it is calmer to compare objects using this method.

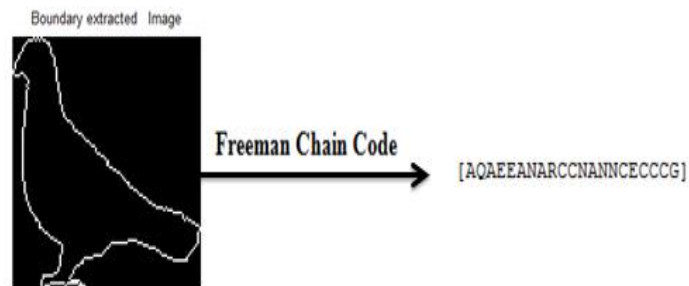


Fig.4 Sequence Generation

From a real-world fact of view, alignment is found by inserting blank spaces inside the sequences (it is called gaps) in order to increase the point to point matching between them. A big amount of methods have been proposed in the past to recover this problem with already effective methods old and in the seventies or early eighties. A detailed treatment we are not discussing. Here, since we are interested in examining the basic potentialities of our ideas, we chose two very basic pairwise sequence alignment tools (called the Needleman-Wunsch and the Smith-Waterman approaches).

In specific, the NeedlemanWunsch algorithm [13] is a dynamic programming method for finding the greatest global alignment between two aminoacid sequences – it represents the one and only first application of dynamic programming to biological sequence evaluation. The basic idea is to exploit the similarity among two sequences by

- i) Making use of a similarity matrix (known as Scoring Matrix) which defined as similarity among every pair sequence of symbols in the alphabet.
- ii) By considering into account penalty values for gap opening and extension. Different possible scoring matrices, which are normally built based on biological knowledge The algorithm consists of three basic steps:
 1. Matrix fill step
 2. Initialization step
 3. Trace back step

After trace back step we will get the best alignment between two sequences, given below. Similarity is based on how well aligned the sequences are.

```

G A A T T C A G T T A
| | | |
G G A _ T C _ G _ _ A

```






International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

V. EXPERIMENTAL RESULTS

Table Below shows the experimental result of Adopting biological alignment tools in 2D shape recognition

IMAGE	SAMPLE IMAGES	MATCHING %
		100
		42.5
		47.5
		12.5

IV. CONCLUSION

In this paper, we have presented a technique for recognizing two dimensional image that are taken from database and compared it by using Freeman chain code algorithm and Needleman Wunsch algorithm. Recognizing Two dimensional image has got 4 important stages. In the first step shapes have to be collected, scanned and resized. In the next stage the collected images have to undergo pre-processing. Features will be extracted in the later stage. And then Sequence generation and sequence alignment step.

We preliminary examined the idea of developing bioinformatics tools to solve Pattern Recognition difficulties. In specific we cast the two dimensional shape analysis problem converted into the biological sequence alignment problem, in which a huge amount of methods have been proposed in the bioinformatics area. Obtained results inspire us to go ahead with this research line.

REFERENCES

- [1]G. Andreu, A. Crespo, and J. Valiente. Selecting the toroidal self-organizing feature maps (TSOFM) best organized to object recognition. In *Proc. of IEEEICNN97*, volume 2, pages 1341–1346, 1997.
- [2] M. Bicego, A. Martins, V. Murino, P. Aguiar, and M. Figueiredo. 2d shape recognition using information theoretic kernels. In *Proc. Int. Conf on Pattern Recognition*, pages 25–28, 2010.
- [3] A. Bronstein, M. Bronstein, and R. Kimmel. The videogenome, arXiv:1003.5320v1, 2010.
- [4] L. Chen, R. Feris, and M. Turk. Efficient partial shapematching using smith-waterman algorithm. In *Proc.Int. Conf on Computer Vision and Pattern Recognition*,2008.
- [5] M. Daliri and V. Torre. Shape recognition based onkernel-edit distance. *Computer Vision and Image Understanding*,114(10):1097–1103, 2010.
- [6]R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biologicalsequence analysis: probabilistic models of proteinsand nucleic acids. cambridge univ, 1998.
- [7] R. Huang, V. Pavlovic, and D. Metaxas. A profile hiddenmarkov model framework for modeling and analysisof shape. In *Proc. Int. Conf on Image Processing*,pages 2121 –2124, 2006.
- [8] C. Kemena and C. Notredame. Upcoming challengesfor multiple sequence alignment methods in the highthroughputera. *Bioinformatics*, 25(19), 2009.
- [9]H. Li and N. Homer. A survey of sequence alignmentalgorithms for next-generation sequencing. *Briefings inBioinformatics*, 11(5):473–483, 2010.
- [10]S. Loncaric. A survey of shape analysis techniques. *PatternRecognition*, 31(8):983–1001, 1998.
- [11]S. Madeira and A. Oliveira. Biclustering algorithms forbiological data analysis: a survey. *IEEE Trans. on ComputationalBiology and Bioinformatics*, 1:24–44, 2004.



ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

- [12]Y. Mingqiang, K. Kidiyo, and R. Joseph. A survey of shape feature extraction techniques. In P.-Y. Yin, editor, *Pattern Recognition Techniques, Technology and Applications*. 2008.
- [13]S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.
- [14]T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*.