DOI: 10.21917/ijivp.2011.0048

COMPUTATION OF IMAGE SIMILARITY WITH TIME SERIES

V. Balamurugan[1] , K. Senthamarai Kannan[2] and S. Selvakumar[3]

[1]*Department of Computer Science and Engineering, Sri Vidya College of Engineering and Technology, Tamil Nadu, India*
E-mail: bala_vm@yahoo.com
[2,3]*Department of Statistics, Manonmaniam Sundaranar University, Tamil Nadu, India*
E-mail: [2]senkannan2002@gmail.com and [3]selvamsuy@yahoo.co.in

Abstract

Searching for similar sequence in large database is an important task in temporal data mining. Similarity search is concerned with efficiently locating subsequences or whole sequences in large archives of sequences. It is useful in typical data mining applications and it can be easily extended to image retrieval. In this work, time series similarity analysis that involves dimensionality reduction and clustering is adapted on digital images to find similarity between them. The dimensionality reduced time series is represented as clusters by the use of K-Means clustering and the similarity distance between two images is found by finding the distance between the signatures of their clusters. To quantify the extent of similarity between two sequences, Earth Mover's Distance (EMD) is used. From the experiments on different sets of images, it is found that this technique is well suited for measuring the subjective similarity between two images.

Keywords:

Similarity Search, Vector Quantization, Similarity Measures, Clustering, EMD

1. INTRODUCTION

For the past two decades, similarity search and retrieval of data in the time series database has received a considerable attention by the researchers, because of its widespread applications in several fields [21]. The domain where similarity search is useful includes the following:

- Meteorologist frequently come across a situation where he needs to find those periods in a given time series, that have similar weather conditions as of the given query period.
- Medical practitioners may be interested to know the details about the patients who have similar electrocardiogram (ECG) of another patient.

Searching for similar patterns in a set of time series helps us to perform data mining activities such as prediction, classification, hypothesis testing, retrieval, indexing, change detection, frequent pattern mining, and segmentation of time series [5]. Time series is a sequence of real numbers representing values at specific time points. Examples for the time series include, meteorological data observed on hourly basis, web logs containing navigational patterns of the internet users, stock prices, clinical datasets, etc. Digital image, audio and video signals can also be treated as a time series, since time stamps are implicitly attached with these data. In case of digital image, the scan time of the electron beam at a particular pixel is used as a time stamp. Similarity search in time series database is a problem of finding data sequences from the database, that is similar to the given query sequence. Similarity can be computed by measuring the distance between the two sequences. To quantify the similarity, various similarity measures viz. Euclidean distance, Minkowski distance, Manhattan distance, edit distance, etc have been used. Some of the challenges in similarity search are: selection of suitable similarity measure, high dimensionality, false dismissal owing to data approximation, and selection of index structure.

In the real time applications, the time series will have voluminous data. For example, if a grey level digital image of size 256 X 350 is converted into time series, then the resultant time series will have 89600 values. Processing the entire range of values to find the similarity will lead to increase in computational time. To address this issue, dimensionality reduction techniques are used. There are various techniques viz. Piecewise Constant Approximation (PCA), piecewise linear approximation, piecewise vector quantization, derivative time series segment approximation, signature extraction, and data transformation are used to reduce the data dimension. After reducing the size of the data, clustering is performed on the time series. The resultant series can be indexed based on clusters. Further, the features [7] can be extracted and compared to find the similarity. The running time for the search can be reduced by finding distance between the extracted features instead of considering all the data points. The main focus of this work is to exploit the techniques that are available in the time series analysis to image retrieval.

The indexed time series of features can now be treated as a text and text retrieval algorithm [12, 25] can be applied to retrieve the similar time series. In the proposed work, PCA and K-Means Clustering are used to approximate the time sequence. The concept is extended to digital images for the retrieval of image and the EMD is used to measure the distance between two time series.

The rest of the paper is organized as follows: Section 2 provides information on related works that deals with time series similarity. Section 3 describes the notation, definitions related to similarity measure, dimensionality reduction, vector quantization and image retrieval [24, 28]. Section 4 describes the methodology in finding the similarity. In Section 5, experimental results and analysis are presented. Finally the concluding remarks and the directions for future work are furnished in section 6.

2. RELATED WORKS

Similarity search over time series data has been studied for many years. In similarity search, for a given query sequence, all the sequences in the database that are similar to the given query are found. The similarity search in time series was attempted initially by Rakesh Agarwal et al. [1, 2], where they proposed an indexing method for time sequences in order to process similarity queries. They transformed the time series to frequency domain by applying Discrete Fourier Transformation (DFT). Features were extracted from the frequency domain and indexed using R – Trees. Dimensionality reduction plays a vital role in enhancing the speed of operation during similarity search. Faloutsos et al. [6], proposed a signature based technique for similarity-based queries, where approximate matching was done for a whole sequence using key words called signatures. As an alternate method, Keogh and Pazzani [15], represented the time series as a piecewise linear segments for fast as well as accurate classification and clustering in a relevance feedback framework. The results were further improved by Adaptive piecewise constant modeling that was suggested by Scargle et al. [20] for the signals in multidimensional spaces. Qiang Wang and Vasileios Megalooikonomou [18] proposed a dimensionality reduction techniques in which piecewise vector quantization was used for data reduction. This technique improved the speed of matching and reduced the false dismissal. Khanh et al. [16] proposed a novel nonlinear transformation scheme named bounded approximation for extracting the features. A better dimensionality reduction technique named Derivative Segment Approximation (DSA) was introduced by Francesco Gullo et al. [9] to represent the time series.

The transformation based techniques such as wavelet transformation is also useful in similarity matching. Chan and Fu [4], proposed a method for matching two time series by transforming the time series into frequency domain using wavelet transformation. Franky et al. [11] applied the Haar wavelets along with time warping for efficient similarity search in time series. Also, Ivan Popivanov and Miller [14] worked on similarity search in time series using wavelets. While most of the research works focused on finding the similarity between two given sequences, Xiang Lian [27] focused on efficient similarity search over the future stream time series. As another approach, Sangjun et al. [19], proposed a method to compute the minimum distance queries for time series data. The above work assumed that both the series have the same length. However, there may be some occasions where the two series may have different length. Tamer and Ambuj [23], were working on arbitrary length time series as well as optimization of the similarity search. Durga Toshniwal and Joshi [5], proposed a method for similarity search in time series data using time-weighted slope. Though there was a considerable improvement in speed of operations the presence of accuracy loss due to approximation was always there in the above methods. Some of the researchers focused on time series segmentation. In [26], Xiaoyan Liu et al. developed two online piecewise linear segmentation methods viz. Feasible Space Window (FSW) and the Stepwise Feasible Space Window (SFSW) for time series segmentation. Bautista-Thompson et al.[3], designed a shape similarity index based on histogram that shows the statistical distribution of point to point differences between two time series. In [24], Xiao-Li Dong et al. presented a shape-based discrete symbolic representation and a novel distance measure that was used to measure the similarity between two time series. Hyo-Sang Lim et al. [13], proposed a similar sequence matching method that efficiently supports variable-length and variable-tolerance continuous queries on time series data stream. He used the window construction mechanism that divides long sequences into smaller windows for indexing and searching the sequences. The most related work to our paper is the dimensionality reduction techniques for efficient time series analysis [18]. The work, proposed in [18] is enhanced along with incorporation of techniques such as K – means clustering, interpolation and extended to measure the similarity between two images.

So far the research works have been progressed in two diverse areas viz. time series analysis and image similarity analysis. As a novel approach we adapt an interdisciplinary approach, where the techniques of both domains are amalgamated in order to exploit the best practices. In addition the current limitation of having equal length for both time series is overcome by applying the interpolation.

3. NOTATIONS AND DEFINITIONS

This section provides various concepts, notations and definitions that are needed to understand the current work.

3.1 NOTATIONS

Let T be the original time series with n time-stamped feature values observed over the time. For a given feature e, let e_i be the value of feature at time stamp i. Then the time series of feature e is represented as, $T = e_0, e_1, e_2, \ldots. e_{n-1}$. For example the feature in a time series for stock price might be daily closing stock price of a specific company. In real time, the time series data set contains huge volume of data that is not suitable for further processing. The given series needs to be approximated and the approximated form of the original time series be, $T' = e_0, e_1, \ldots.,$ e_{l-1} where l is the length of the approximated time series. The value of the l is far less than n. The other notations are: N – number of time series in the dataset, D – the Euclidean distance between two series and EMD – Earth Mover's Distance.

3.2 DEFINITIONS

Quantisation: A process in which the continuous range of values of an analog signal is sampled and divided into non-overlapping sub-ranges and a discrete, unique value is assigned to each sub-range [10].

Vector Quantization (VQ): It is a process that divides large set of points into groups having approximately the same number of points, closest to them. The idea is that in image and voice, many of the data points that describe a particular pixel or sound are similar, and hence, not much information will be lost if each vector is replaced by its closest reference vector. In addition, the information can be represented in a much smaller space. VQ is useful in pattern recognition since it reduces the number of patterns that need to be considered.

Histogram. Histogram of a set [22] with respect to a measurement represents the frequency of quantified values of that measurement. It is a mapping from a set of 'd' dimensional integer vectors 'I' to the set of non-negative reals. Vectors 'I' represent bins.

Signature. The signature [6, 8] of a set is a lossless representation of its histogram in which bins of the histogram that have the value 0 are expressed implicitly. Alternately, a signature $\{S_j = (m_j, w_{m_j})\}$, represents a set of feature clusters. Each cluster is represented by its mean (or mode) m_j, and by the fraction w_{m_j} of pixels that belong to that cluster. The integer subscript j ranges from one to a value that varies with the complexity of the particular image.

Ground distance. Ground distance can be defined as a distance between the basic features that are aggregated into distributions. This is called ground distance.

Earth Mover's Distance. EMD was initially Proposed by Peleg, Werman and Rom [17] and reinvented for CBIR by Yossi Rubner in 2000 [28]. It is a true metric and it is based on the minimal cost that must be paid to transform one distribution into the other. In [28], Rubner et al. computed the EMD by solving the transportation problem. In general, the transportation problem deals with finding the least expensive flow of goods from the suppliers to the consumers that satisfies the consumer's demand. However, in case of EMD, the distance between two histograms is defined as the minimum cost required moving pixels of one bin to other bin of first signature so that the resulting signature is same as second one. EMD between two signatures is expressed as,

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \tag{1}$$

In Eq.(1), $P = \{(p_i, w_{p_1}).........(p_m, w_{p_m})\}$ is the first signature with m clusters, P_i the cluster representative and w_{p_1} is the weight of the cluster. $Q = \{(q_j, w_{q\,j}),....(q_n, w_{q_n})\}$, denotes the second signature with n clusters. The ground distance between p_i and q_i is represented as $[d_{ij}]$ and f_{ij} is the flow between p_i and q_j.

4. MATERIALS AND METHODS

The problem of similarity search is concerned with efficiently locating subsequences in large archives of sequences or in a single long sequence and it may be stated as follows: Given a query sequence *Q*, a set of time sequence *T*, a distance measure *d*, and tolerance ε, find the set *R* of sequences similar to *Q*. To find exact matches, one can use sliding window approach or other techniques. However, in case of the similarity search, one needs to locate approximate matching and there is a need to quantify the extent of similarity between two sequences. Similarity search can be categorized into two types such as whole sequence matching and subsequence matching. In whole sequence matching, it is assumed that all sequences to be compared are of the same length. In the proposed work, whole sequence matching is used to find similarity between two images. From mathematical point of view, distance is defined as a quantitative degree of how far apart two objects are. The choice of similarity measure is domain dependent. There are many similarity distance viz. L_p norm such as Euclidean distance, time warping, cosine distance, Bhattacharya distance, jaccard distance, Hamming distance, longest common subsequence etc. Out of these, Euclidean distance is one of the popular metrics used widely for measuring the similarity between two elements. However, in many occasions the euclidean distance is unable to capture subjective similarities effectively. The distances based on non-parametric test statistics such as Kolmogorov-Smirnov distance, Cramer/Von Miser type statistics and the distances based on information theory such as Kullback Leibler divergence, Jeffrey divergence, etc. are also available in practice. Distances based on ground distance measures such as Quadratic Form and Earth mover's distance have also been used in the recent works. The main advantages of the EMD are as follows:

- Robust in comparison to the other histogram matching techniques.
- It suffers from no quantisation problems due to rigid binning.
- It tolerates the shift in the feature space.
- Increased precision for image retrieval.
- It supports partial matching.
- It uses only signatures so better memory usage.

4.1 DIMENSIONALITY REDUCTION

Dimensionality reduction plays an important role in fast similarity searches in large databases. When the image is converted as time series, its dimension is normally large and there is a need for approximation. One way to approximate the time series is to break the series with constant segments and to approximate the series with its mean value. However, the mean value is affected by its extremity. Therefore, mean value along with its coefficient of variations (CV) are used to approximate the time series. VQ is applied on the resulting series and it results in histogram. EMD is applied on the histograms to find the similarity between two series.

4.2 CLUSTER ANALYSIS

Cluster analysis groups observations based on the information found in the data describing the observations or their relationships. The aim of the cluster analysis is to make the observations in a group similar to one other and different from the observation in other group. The best definition for the cluster depends on the type of data and the desired results. Similarity based cluster definition states that a cluster is a set of objects that are *similar*, and objects in other clusters are not *similar*. Several clustering techniques have been proposed over the years. They are: i) Hierarchical vs. partitional, ii) Divisive vs. agglomerative and iii) Incremental vs. non-incremental. The clustering can be treated like an optimization problem. One approach to optimizing a global objective function is to rely on algorithms, that find effective solutions, but not optimal. An example of this approach is the K-means algorithm which tries to minimize the sum of the squared distances between objects and their cluster centers. In the proposed work K-means clustering is used to convert the time series into series of clusters. It is based on the idea that a center point can represent a cluster. The main issue in the K-means algorithm is to find a suitable value for K. There are many methods viz. Rule of thumb, Elbow Distance method, Information criterion approach, Information theoretic approach, etc deal with the selection of K.

5. EXPERIMENTAL RESULTS

From the experimental results it is observed that, EMD on clustered time series provides an appropriate measure for similarity between two time sequences. Let us consider two time series X and Y as given below. The time stamps of X and Y are not explicitly given.

$$X = [4, 4, 3, 3, 1, 1, 2, 2]$$
$$Y = [1, 1, 2, 2, 3, 3, 4, 4]$$

The EMD between the x and y without taking the timestamp into consideration is ‘0’. Let $X1$ and $Y1$ be the two time sequences with time stamps attached explicitly.

$$X1 = [[1,4],[2,4],[3,3],[4,3],[5,1],[6,1],[7,2],[7,2]]$$
$$Y1 = [[1,1],[2,1],[3,2],[4,2],[5,3],[6,3],[7,4],[8,4]]$$

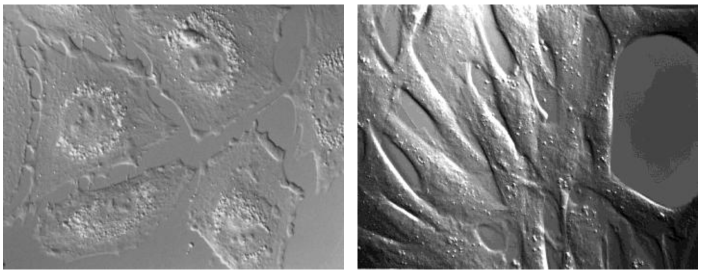
The EMD computed for *X1* and *Y1*, after clustering with two clusters is 1 and with five clusters is 6. Since all the eight values of X and Y are different with respect to their time stamp, EMD computed after clustering provides a better measure. While clustering many time series, the instances of all the time series are averaged and the cluster centers are found based on their average values. In the present case, the resultant time series C obtained by averaging *X1* and *Y1* is clustered.

$$C = [[1, 2.5], [2, 2.5], [3, 2.5], [4, 2.5], [5, 2], [6, 2], [7, 3], [8, 3]]$$

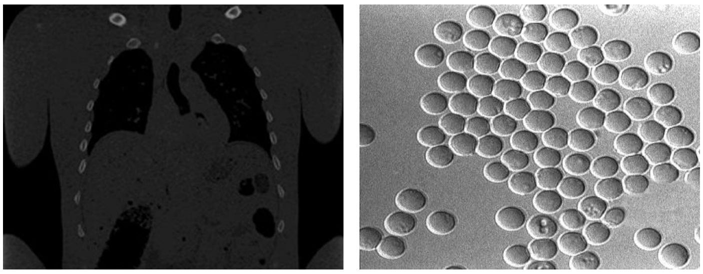
5.1 EXPERIMENTS WITH IMAGES

The histograms of two images differ, if the two images are different. However, the reversal is not true. There are some possibilities that the two histograms are equal for two different images. The Fig.1 shows four test images for which the EMD is computer after clustering. The dotted line shows the histogram of the Endocell and black line indicates the Muscle. Computation of EMD directly on a histogram of a set of images leads to a wrong conclusion about the similarity between two images. The reason is the spatial coordinates of the pixel values are ignored. This can be avoided by converting the image into time series and by finding the EMD on the clustered time series. Interpolation is used to make the dimensions of different images equal.

To reduce the computational time, PCA is used to reduce the dimension of the time series. The mean value is obtained from all the time series of the test images and the clustering is performed on the resultant data.



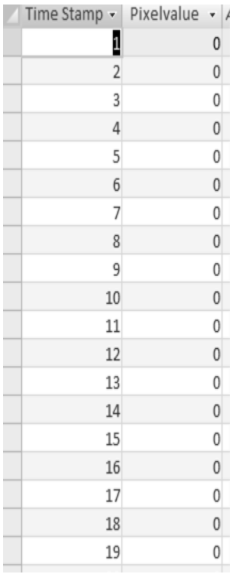
(a) (b)



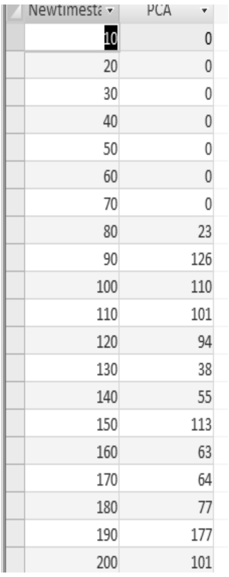
(c) (d)

Fig.1. Test Images with dimensions (a) Endocell (615 x 416) (b) Muscle (652 x 444), (c) CT Bone (483 x 410), (d) RB Cell (432 x 389)

The Fig.2 illustrates the snapshot of the time series obtained from the input image *CT Bone*. The total number of pixels in the resultant time series is 198030. The resultant time series that was obtained after applying the PCA is shown in Fig.3.



Time Stamp	Pixelvalue
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0



Newtimesta	PCA
10	0
20	0
30	0
40	0
50	0
60	0
70	0
80	23
90	126
100	110
110	101
120	94
130	38
140	55
150	113
160	63
170	64
180	77
190	177
200	101

Fig.2. Time series of CT Bone Fig.3. PCA of *CT Bon*

To cluster the time series, K-Means clustering with $K = 256$ is used. Setting the value for K is still a main issue in K-Means clustering. In the present case K is set to the value which is slightly greater than square root of the number of elements in the time series and it should satisfy the condition $K=2^z$ where z is an integer. There are two advantages in the present clustering approach. First, a rough similarity measure is obtained by grouping the pixel values that have similar value. Second, the spatial coordinates are considered, as the time stamp of the sequences are utilized in clustering. The whole sequence matching is used to determine the similarity between two sequences.

Experiments have been conducted with the distances such as Manhattan distance, Bray-Curtis distance, Canberra distance, Cosine distance and EMD. The distances among the given test images are presented in the Table.1 and it reveals that the image endocell and RB Cell are highly similar with EMD 1195. The distances between the CT Bone and other images are high. Experiments have been carried out with different set of images and its efficiency is found good.

Table.1. EMD Matrix for Input Images

Image-I	Image-II	Manhattan Distance	Bray Curtis Distance	Canberra Distance	Cosine Distance	EMD
CT Bone	Endocell	96	0.347826	0.638163	0.999014	6649
CT Bone	Muscle	266	0.596413	0.634460	0.998749	8860
CT Bone	RB Cell	133	0.424920	0.618042	0.998776	7185
Endocell	Muscle	170	0.313653	0.182729	0.999984	4373
Endocell	RB Cell	37	0.0904645	0.0959158	0.995595	1195
Muscle	RB Cell	133	0.229706	0.192117	0.995053	3198

6. CONCLUSION

A method that uses time series similarity analysis to compute similarity between two images is introduced. Grey level test images have been converted into time series. The dimensions of test images were made equal by interpolating the respective time series. Dimensionality of a time series is reduced by applying PCA. Average value of all the test time series is found and the resultant series is clustered using *K*-Means clustering. The drawbacks while using the Euclidean distance to measure the similarity is rectified by the use of EMD. Since the EMD captures subjective similarity effectively, our approach exhibits its efficiency in finding the similarity between images. Experiments were performed on several gray level images to demonstrate the efficiency. This approach can easily be applied to other time series applications like Stock analysis, meteorological data analysis, content based image retrieval, etc.

REFERENCES

[1] R. Agarwal, Christos Faloutsos and Arun Swami, "Efficient similarity search in sequence databases", *Proceedings of the 4th International Conference of Foundations of Data, Organization and Algorithms (FODO),* Vol. 730, pp. 69 – 84, 1993.

[2] R. Agarwal, K. Lin, H. S. Sawhney and K. Shim, "Fast similarity search in the presence of noise, scaling and translation in time-series database", *Proceedings of the 21st International Conference on Very Large Data Bases,* pp.490-501, 1995.

[3] Bautista. E Thompson and S. Santos De la Cruz, "Shape similarity Index for Time series based on features of Euclidean Distance Histogram", *Proceedings of the 15th IEEE International Conference on Computing,* pp. 60-64, 2006.

[4] K. Chan and A. W Fu, 1999, "Efficient time series matching by wavelets", *Proceedings of the 15th International Conference on Data Engineering*, pp. 126–133, 1999.

[5] Durga Toshniwal and R. C. Joshi, "Similarity Search in Time Series Data using Time Weighted Slope", *Informatica*, Vol. 29, pp. 79 – 88, 2005.

[6] C. Faloutsos, M. Ranganathan, A.O Mendelzon, and T. Milo, " A signature technique for similarity-based queries", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 419-429, 1994.

[7] Florence Duchene, Catherine Garbay and Vincent Rialle, "Learning recurrent behaviors from heterogeneous multivariate time-series", *Artificial Intelligence in Medicine*, Vol. 39. No. 1, pp. 25 – 47, 2007.

[8] Francesc Serratsa and Alberto Sanfeliu, "Signature versus histogram: Definitions, distances and algorithms", *Pattern Recognition*, Vol. 39, pp. 921-934, 2006.

[9] Francesco Gullo, Giovanni Ponti, Andrea Tagarelli and Sergio Greco, "A time series representation model for accurate and fast similarity detection", *Pattern Recognition*, Vol. 42, No. 11, pp. 2998–3014, 2009.

[10] Francisco J. Ruiz, Cecilio Angulo and Nuria Agell, "A Supervised Interval Distance-Based Method for Discretization", *IEEE Transaction on Knowledge and Data Engineering,* Vol. 20, No. 9, pp. 1230–1238, 2008.

[11] Franky Kin-Pong Chan, Ada Wai-chee Fu, and Clement Yu, "Haar wavelets for efficient similarity search in time series: with and without time warping", *IEEE Transaction on Knowledge and Data Engineering,* Vol. 15, No. 3, pp. 686–705, 2003.

[12] Hung Chim and Xiaotie Deng, "Efficient Phrase-Based Document Similarity for Clustering", *IEEE Transaction on Knowledge and Data Engineering*, Vol. 20, No. 9, pp. 1217–1229, 2008.

[13] Hyo-Sang Lim, Kyu-Young Whang and Yang-Sae Moon, "Similar Sequence Matching supporting variable-length and variable-tolerance continuous queries on time series data stream", *International Journal on Information Sciences*, Vol. 178, pp. 1461-1478, 2008.

[14] Ivan Popivanov and Renee J. Miller, "Similarity search over time series data using wavelets", *Proceeding of the 18th International Conference on Data Engineering,* pp.212-221, 2002.

[15] K. J. Keogh and M. J. Pazzani, "An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback", *Proceedings of 4th International Conference on Knowledge Discovery and Data Mining,* pp. 239-243, 1999.

[16] Khanh Vu, Kien A. Hua, Hao Cheng, and Sheau-Dong Lang, "Bounded Approximation: A new criterion for

dimensionality reduction approximation in similarity search", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 6, pp. 768 – 783, 2008.

[17] S. Peleg, M. Werman and H. Rom, "A unified approach to the change of resolution: Space and gray-level", *IEEE Transaction on pattern Analysis and Machine Intelligence*, Vol. 11, pp. 739-74, 1999.

[18] Qiang Wang and Vasileios Megalooikonomou, "A dimensionality reduction techniques for efficient time series similarity analysis", *Journal of information system, Elsevier*, Vol. 33, No. 1, pp. 115-132, 2008.

[19] Sangjun Lee, Dongseop Kwon and Sukho Lee, "Minimum distance queries for time series data", *The journal of systems and software*, Vol. 69, No. 1-2, pp. 105 -113, 2004.

[20] J. Scargle, Jackson. B and Norris. J.,"Adaptive Piecewise Constant modeling of signals in Multidimensional spaces PHYSTAT.", pp. 157 – 161, 2003.

[21] Srivatsan, B. Laxman, and P.S Sastry, "Survey of temporal data mining", *Sadhana Academy of Proceedings in Engineering Scheme,* Vol. 31, No. 2, pp. 173-198, 2006.

[22] Sung-Hyuk Cha and Sargur N. Srihari, "On measuring the distance between histograms", *Pattern Recognition,* Vol. 35, No. 6, pp. 1355-1370, 2002.

[23] Tamer Kahveci and Ambuj K. Singh, "Optimising Similarity Search for arbitrary length time series queries", *IEEE Transaction on Knowledge and Engineering,* Vol. 16, No. 4, pp. 418- 433, 2004.

[24] Xiao-Li Dong, Cheng-Kui Gu and Zheng-Ou Wang "Research on shape based time series similarity measure", *15th IEEE International conference on Machine Learning and Cybernetics,* pp. 1241 – 1246, 2006.

[25] Xiaojun Wan, "A novel document similarity measure based on earth mover's distance", *International Journal for Information Sciences*, Vol. 177, No. 18, pp. 3718-3730, 2007.

[26] Xiaoyan Liu, Zhenjiang Lin and Huaiqing Wang. "Novel online methods for time series segmentation", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 12, pp. 1618-1626, 2008.

[27] Xiang Lian and Lei Chen, "Efficient similarity search over future stream time series", *IEEE Transaction on Knowledge and Data Engineering*, Vol. 20, No. 1, pp. 40-53, 2008.

[28] Yossi Rubner, Carlo Tomasi and Leonidas J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval", *International Journal of Computer Vision*, Vol. 40, No. 2, pp. 99-121, 2000.