Proceedings of the International MultiConference of Engineers and Computer Scientists 2010 Vol I,
IMECS 2010, March 17 - 19, 2010, Hong Kong

Motif and Anomaly Discovery of Time Series Based on Subseries Join

Yi Lin * and Michael D. McCool † Ali A. Ghorbani ‡

Abstract — **Time series motifs are repeated similar subseries in one or multiple time series data. Time series anomalies are unusual subseries in one or multiple time series data. Finding motifs and anomalies in time series data are closely related problems and are useful in many domains, including medicine, motion capture, meteorology, and finance.**

This work presents a novel approach for both the motif discovery problem and the anomaly detection problem. This approach first uses subseries join to obtain the similarity relationships among subseries of the time series data. Then the motif discovery and anomaly detection problems can be converted to graph-theoretic problems solvable by existing graph-theoretic algorithms. Experiments demonstrate the effectiveness of the proposed approach to discover motifs and anomalies in real-world time series data. Experiments also demonstrate that the proposed approach is efficient to process large time series datasets.

Keywords: ***pattern recognition, motif discovery, anomaly detection, time series, graph-theoretic algorithm***

1 Introduction

Time series are composed of sequences of data items measured at typically uniform intervals. Time series arise frequently in many scientific and engineering applications, including finance, medicine, digital audio, and motion capture. Motifs are approximately repeated subseries in a single time series data or a time series dataset. One example is shown in Figure 1(a). Anomalies are unusual subseries in a single time series data or a time series dataset. Anomalies can be outliers in time series that contains approximately periodic patterns, or subseries of deviations of a quantity from some mean. Either the mean is calculated for the entire time series, or only for parts of it. For example, in Figure 1(b), the underlined part is quite different from the other parts of the data that is sine-like.

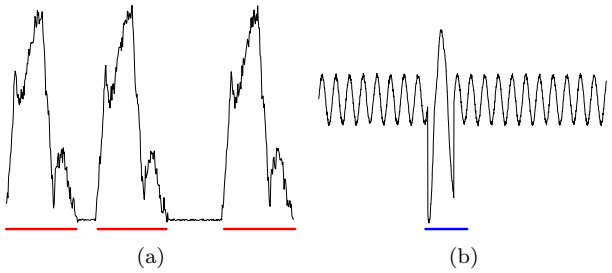


Figure 1: Simulated examples of time series motif and anomaly. (a) The approximate repeated subseries that are underlined are the motif of this time series. (b) The unusual subseries that is underlined is the anomaly of this time series.

Motif discovery and anomaly detection are fundamental unsupervised learning techniques and are useful in many real-world applications. For example, motifs in Web posts or comments are frequent term-based patterns. They are used in analyzing people's opinions or feedbacks. Motifs of stock prices are similar value patterns. These patterns may indicate future trends. Anomaly detection in cardiogram data can help a doctor to diagnose heart deceases. Anomaly detection in network traffic analysis will alert the network administrator of spam or malicious attacks. Given the wide use in so many domains, discovering motifs and anomalies is attracting increasing attentions from both academia and industry.

Existing approaches usually handle motif discovery and anomaly detection separately [1, 2]. Also they are limited in that they can only find similar patterns of the same length or tolerate a limited amount of phase-scaling. Some of them have high computational complexities.

In this work, we propose an approach for both motif discovery and anomaly detection. This approach can overcome the limitations of previous work. This work is based on a feature representation and a subseries join operation that were proposed in our previous work [3, 4]. Based on subseries join, the motif discovery and anomaly detection problems can be easily converted to graph-theoretic problems. Then motif discovery and anomaly detection problems can be easily solved together by using existing graph-theoretic algorithms.

The rest of the paper is organized as the follows. In Sec-

*Faculty of Computer Science, University of New Brunswick, 540 Windsor Street Fredericton, NB, Canada E3B 5A3. Email: ylin@unb.ca. Corresponding author.

†Intel Corporation. Email: michael.mccool@intel.com.

‡Faculty of Computer Science, University of New Brunswick, 540 Windsor Street Fredericton, NB, Canada E3B 5A3. Email: ghorbani@unb.ca.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

tion 2, we review previous work and underlying technology on time series motif discovery and anomaly detection. In Section 3, we introduce a feature representation and a subseries join operation. Section 4 proposes our motif discovery and anomaly detection methods. We perform empirical evaluations in Section 5. Finally Section 6 provides conclusions and suggestions for future work.

2 Related Work

Time series motifs are widely used in various medical applications, including examining the utility of on-body monitoring sensors [5] and selecting maximally informative genes [6]. Time series motifs are also used in finding patterns in sports motion capture data [7] and video surveillance applications [8].

To the best of our knowledge, the first formal definition of time series motifs was proposed in 2003 [9]. Based on this work, a fast motif discovery algorithm was introduced later in [1]. However, the above work is only limited to discovering a motif whose subseries are of the same length. Yankov et al. [10] proposed a motif discovery method that uses a uniform scaling Euclidean distance and a symbolic representation based on thresholding. Uniform scaling is important for indexing and matching time series for motion-capture data [11] and music. The thresholds they used to convert a time series to a symbolic sequence are heuristically determined. This method is semi-automatic because the user also needs to specify the length of the motif subseries manually. This arbitrary segmentation may cause undesirable division of important features in the data into different segments. Overlapping sliding windows can be used [12] to avoid division of features but at the cost of a redundant representation. Generally, a better definition of a "motif" that does not depend on a priori knowledge of its shape or length is needed.

Wei et al. proposed an anomaly detection algorithm based on a symbolic representation of time series [13]. This representation was later used in fining unusual shapes in a large image database [2]. This symbolic representation divides a time series into segments of uniform length. However, the manual section of the segment length is not sensitively adapt to the actual behavior of the data. Specifically, wavelet analysis techniques have been widely used for network intrusion detection [14, 15]. However, wavelet analysis-based anomaly detection techniques often have high computation complexity due to wavelet transformation.

3 Subseries Join

Subseries join finds pairs of similar subseries of time series. The subseries join results include similar subseries and excludes dissimilar subseries. The similar subseries consist of motifs. The dissimilar subseries may contain anomalies. As a formal definition, subseries join should return all pairs of subseries drawn from two datasets that satisfy the similarity threshold and are also maximal-length [4].

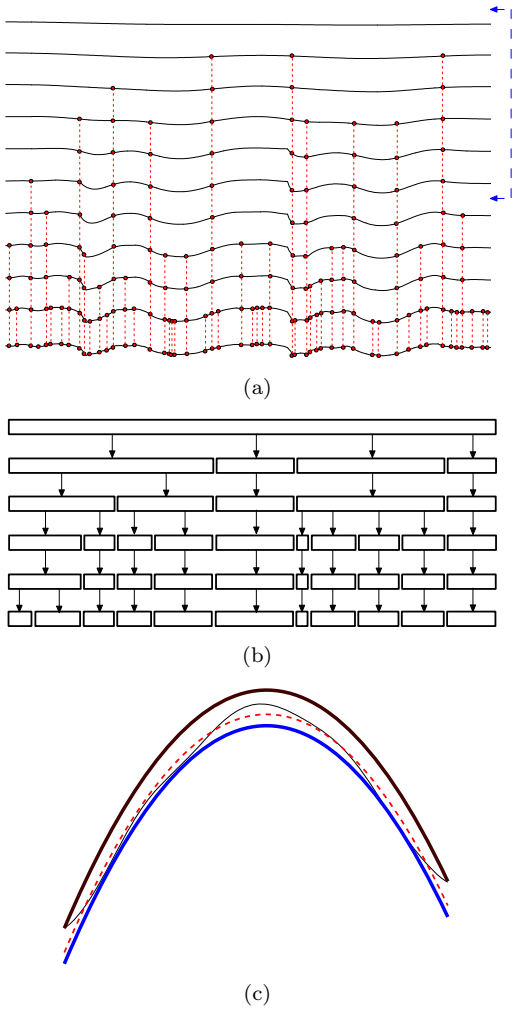


Figure 2: (a) The boundary are lined up by red dotted lines at the same positions in different scales. (b) The hierarchical structure of the scales marked by long dashed lines and arrows at the right side of (a). (c) The thin solid curve is the original time series. The red dashed curves are a quadratic polynomial that fits the original time series. The blue and black thick curves are the minimal quadratic polynomial envelope.

In the following, we briefly describe the subseries join algorithm proposed in our previous work [3, 4]. This subseries join algorithm will be applied to efficiently solve the joint problem motif discovery and anomaly detection. Figure 2 illustrates some kernel steps to conduct subseries join. First, each time series in the dataset is smoothed by an anisotropic diffusion analysis [16], which generates a scale-space of smoothed time series. Then all the time series in the scale space are segmented using the Canny

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

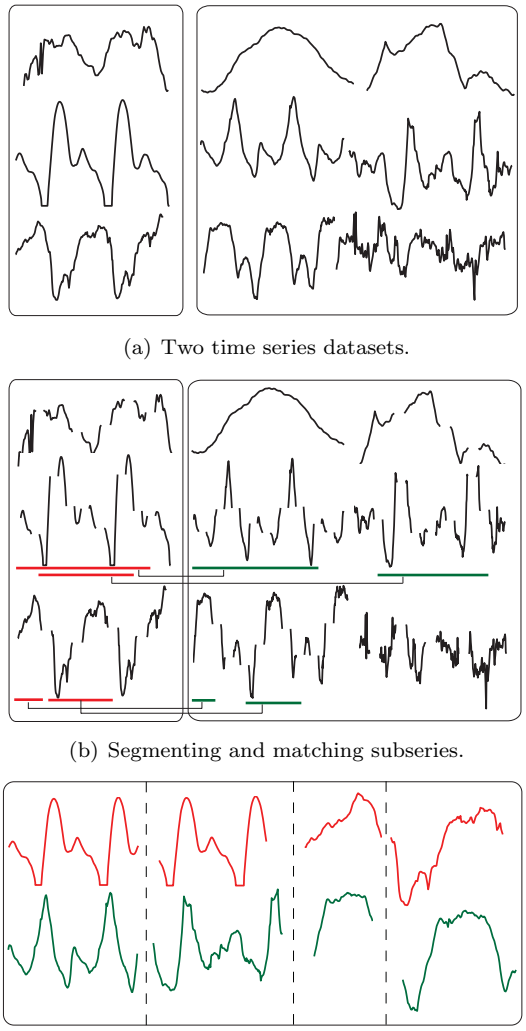
edge detector [17]. The boundary points between pairs of segments are lined up at the same positions of different scales. The scale space of time series and segmentation are shown Figure 2(a). The segments at different scales give a hierarchical structure, which is shown Figure 2(b). A minimal polynomial envelope is used to represent each segment in a reduced-dimensionality space, which is shown in Figure 2(c). We call such a representation a *feature*. A novel distance function is defined over features. This feature representation automatically segments time series by keeping its feature continuities, which is not possible in most of previous work. All features in the dataset are indexed in an R-tree. R-tree leaves are associated with the features in the feature sequences instead of saving feature representations redundantly. Pairs of matching features are found by an R-tree join operation [18]. These matching features generate candidate matching feature sequences. We use a dynamic programming algorithm to compare and align two candidate matching feature sequences. The aligned matching feature sequences can locate the matching subseries.

To give the reader a better view of the subseries join approach, Figure 3 illustrates middle results of subseries join. Given two time series datasets shown in Figure 3(a), time series in the datasets are non-uniformly segmented and pairs of matching subseries are found using dynamic programming, which is shown in Figure 3(b). Figure 3(c) shows a set of pairs of matching subseries that satisfy a similarity threshold and are maximal-length.

4 Motif Discovery and Anomaly Detection

In this section, we will discuss how to use the subseries join algorithm to solve the problems of motif discovery and anomaly detection. For example, the results of subseries join of one or multiple time series can be converted a graph to find motifs and anomalies, as shown in Figure 4. Subseries A matches E. Subseries A', E', and G are found to be similar. The same is for subseries B, D, and F. Subseries C has no matches. The relationship of subseries can be mapped to an undirected graph given in Figure 5(a). Every vertex represents a subseries. Every edge between two vertices represents a matching relationship between a pair of subseries. Then, we remove self-loops from the graph. Also, there may exist overlapped subseries, for example, A and B. If the overlap percentage is above a certain threshold e.g., our prototype system used 70% as a threshold, the overlapped subseries, such as A and A', E and E', are merged. After this processing, Figure 5(a) turns into Figure 5(b).

Now motif discovery can be defined in graph-theoretic terms. Several alternative graph-theoretic definitions of a motif can be considered. One possible definition is based on the *maximal cliques*. A *clique* in an undirected graph is a subgraph in which every vertex is connected to every other vertex in the subgraph. A maximal



(a) Two time series datasets.

(b) Segmenting and matching subseries.

(c) Subseries join results.

Figure 3: Middle results of subseries join.

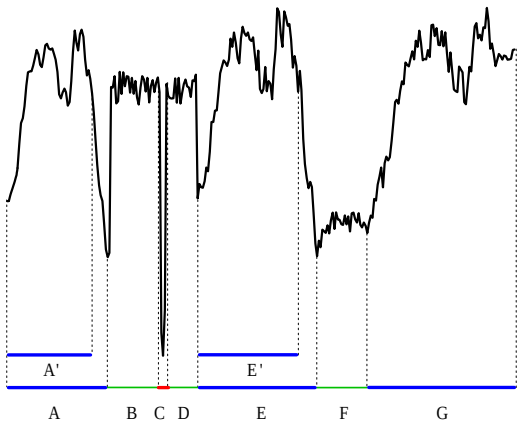
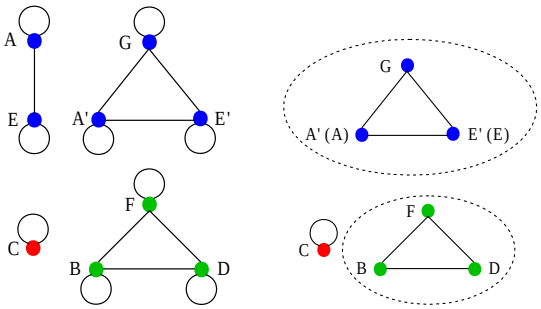


Figure 4: Subseries join results of a time series.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)



(a) The graph of subseries join results.

(b) Remove the self-loops and merge the overlapped vertices A and A', E and E'.

Figure 5: Convert the subseries join results into a graph. The maximal cliques that are framed by dashed circles give the motifs. The isolated vertex C is an anomaly.

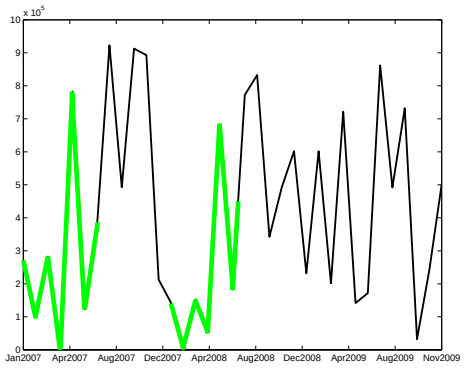
clique is a complete subgraph that is not contained in any other complete subgraph. Unfortunately, the maximal clique problem is an NP-complete problem, and the clique enumeration problem is NP-hard. However, there exist many heuristic algorithms to approximately solve the clique enumeration problem [19], including parallel algorithms [20], and polynomial-time approximation algorithms [21]. Another possible definition of motif is based on the *k-connected subgraph*. A k-connected subgraph in an undirected graph is a subgraph in which every vertex is connected to at least k other vertices in the subgraph. Although k-connected subgraph problem is also NP-complete, many existing approximation algorithms can efficiently solve this problem [22].

Anomaly detection can be defined as finding the isolated vertices in the graph. For example, in Figure 5, the vertex C is an isolated vertex. Finding isolated vertices is simpler than finding maximal cliques or k-connected subgraphs in a graph, because it requires only linear computational time. We used the algorithms in the Boost Graph Library [23] for our prototype system implementation.

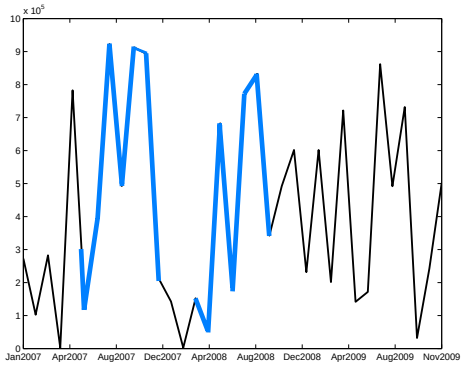
Note that we assume that an anomaly pattern does not repeat in a time series, because if an anomaly pattern repeats, it may be considered as a motif. In real applications, determining whether a pattern is a motif, anomaly, or something else is often subjective, related to different data, domains, and tasks. However, our approach can help the user to efficiently find candidate targets of interest.

5 Experimental Results

We evaluate the effectiveness of our proposed approach for discovering motifs and anomaly in real-world data including stock prices [24] and Electrocardiograph (ECG) data [25]. Figure 6 shows motifs discovered by our prototype system in a time series of stock prices. Figure 7 shows an anomaly discovered by our prototype system in a time series of Electrocardiograph (ECG) data. The found motifs have similar shapes of different lengths. Parts of the found motifs are also overlapped. The found anomaly's properties, such as height and length, can be controlled by parameters in the prototype system. The user can choose different parameters for different data, domains, and tasks. Compared to most previous work [1, 10], our approach can easily adapt to different application by parameter tuning.



(a) One motif is marked by thick green lines.



(b) Another motif is marked by thick blue lines.

Figure 6: Motifs discovered in stock prices: NASDAQ Monthly High, Jan 1, 2007∼Nov 30, 2009.

To evaluate the performance of our approach, the prototype system was also tested on a 3.15GB motion capture dataset [26] containing 3,962,581 frames. A frame is an element in a motion time series. The dataset contains various kinds of motion time series data. Each motion data is a time series data ranging in length from about 300 to 23000 frames. All experiments were run on Linux Kernel 2.6 PC with 500MB RAM and Pentium IV 3.0GHz CPU. The prototype generated 101,397 segments of the dataset. The prototype system needs about 3.5 hours for representation and index construction for the whole dataset. We conduct the proposed subseries join operation on the whole dataset to discover motifs, i.e., motion subseries of different styles. The total computational time is 35.6 minutes. The average computational time per time se-

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

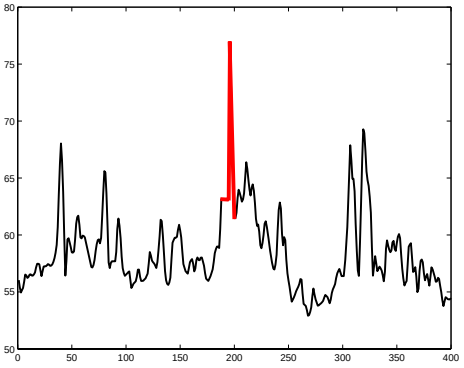


Figure 7: An anomaly marked by thick red lines discovered in Electrocardiograph (ECG) data.

ries is 0.86 seconds. Figures 8-10 show some motifs that are similar motions found in the whole dataset. From the figures, we can see that the motif motions have similar but different poses, e.g., the second elements in the running motions. The motif motions also have different lengths, i.e., different speeds of motions. This demonstrates that our approach can tolerate different lengths and phase-scaling.

(a) Length: 91 frames

(b) Length: 85 frames

(c) Length: 109 frames

Figure 8: A motif found in the dataset that are running motions.

6 Conclusions

In this work, we have discussed motif discovery and anomaly detection for time series data. We propose a novel approach for both problems. This approach is based on a new definition of subseries join and an algorithm to compute subseries join efficiently. Experiments have demonstrated the effectiveness of our proposed approach for discovering motifs and anomalies in real-world time series data. Experiments also show that our approach can efficiently find motifs and anomalies in a large dataset.

Our work provides a general solution motif discovery and anomaly detection on any time series data. Therefore, our prototype systems need some user-defined parameters that are tuned for data in a specific domain. At present, the parameters are selected empirically without any domain knowledge involved. In the future, we will apply this work in some specific domains, such as Web mining and network security. Domain knowledge will be used to automatically tune the system parameters.

(a) Length: 148 frames

(b) Length: 143 frames

(c) Length: 126 frames

Figure 9: A motif found in the dataset that are jumping motions.

(a) Length: 119 frames

(b) Length: 127 frames

(c) Length: 153 frames

Figure 10: A motif found in the dataset that are kicking motions.

References

[1] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover, "Exact Discovery of Time Series Motifs," *SIAM International Conference on Data Mining (SDM'09)*, 2009.

[2] L. Wei, E. Keogh, X. Yi, and M. Yoder, "Efficiently Finding Unusual Shapes in Large Image Databases," *Data Mining and Knowledge Discovery*, Vol. 17, No. 3, pp. 343 - 376, 2008.

[3] Y. Lin and M. D. McCool, "Nonuniform Segment-Based Compression of Motion Capture Data," *Pro-*

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

ceedings of the 3rd International Symposium on Visual Computing (ISVC'07), pp. 56-65, 2007.

[4] Y. Lin, "Subseries Join and Compression of Time Series Data Based on Non-uniform Segmentation," Ph.D. Dissertation, *School of Computer Science, University of Waterloo*, 2008.

[5] D. Minnen, T. Starner, I. Essa, and C. Isbell, "Discovering Characteristic Actions from On-body Sensor Data," *The 10th International Symposium on Wearable Computers (ISWC'06)*, pp. 11-18, 2006.

[6] I. Androulakis, J. Wu, J. Vitolo, and C. Roth, "Selecting Maximally Informative Genes to Enable Temporal Expression Profiling Analysis," *Proceedings of Foundations of Systems Biology in Engineering*, 2005.

[7] Y. Tanaka, K. Iwamoto, and K. Uehara, "Discovery of Time-series Motif from Multi-dimensional Data Based on MDL Principle," *Machine Learning*, Vol. 58, No. 2-3, pp. 269-300, 2005.

[8] R. Hamid, S. Maddi, A. Johnson, A. Bobick, I. Essa, and C. Isbell, "Unsupervised Activity Discovery and Characterization from Event-streams," *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI'05)*, Lecture Notes in Computer Science, Springer, Vol. 3669/2005, pp. 119-130, 2005.

[9] B. Chiu, E. Keogh, and S. Lonardi, "Probabilistic Discovery of Time Series Motifs," *Proceedings of the 9th International Conference on Knowledge Discovery and Data mining (KDD'03)*, pp. 493-498, 2003.

[10] D. Yankov, E. Keogh, J. Medina, B. Chiu, and V. Zordan, "Detecting Time Series Motifs under Uniform Scaling," *Proceedings of the 13rd International Conference on Knowledge Discovery and Data Mining (SIGKDD'07)*, pp. 844-853, 2007.

[11] E. Keogh, T. Palpanas, V. Zordan, D. Gunopulos, and M. Cardle, "Indexing large human-motion databases," *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB'04)*, pp. 780-791, 2004.

[12] Y-S. Moon and K-Y. Whang and W-S. Han, "General Match: A Subsequence Matching Method in Time-series Databases Based on Generalized Windows," *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, pp. 382-393, 2002.

[13] L. Wei, N. Kumar, V. Lolla, E. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Efficiently Finding Unusual Shapes in Large Image Databases," *Proceedings of the 17th International Conference on Scientific and Statistical Database Management*, pp. 337 - 340, 2005.

[14] C-T. Huang, S. Thareja, and Y-J. Shin, "Wavelet-based Real Time Detection of Network Traffic Anomalies," *Proceedings of Workshop on Enterprise Network Security and the 2nd International Conference on Security and Privacy in Communication Networks*, pp. 1-7, 2006.

[15] W. Lu and A. A. Ghorbani, "Network Anomaly Detection Based on Wavelet Analysis," *Journal on Advances in Signal Processing*, Vol. 2009, 2009.

[16] P. Perona and J. Malik, "Scale-space and Edge Detection using Anisotropic Diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 7, pp. 629-639, 1990.

[17] J. Canny "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 8, No. 6, pp. 679-698, 1986.

[18] S. Shekar and S. Chawla, "Spatial Databases: a Tour," *Prentice Hall*, 1st Edition, 2003.

[19] J. M. Byskov, "Algorithms for k-colouring and Finding Maximal Independent Sets," *Proceedings of the 14th ACM and SIAM Symposium on Discrete Algorithms*, pp. 456-457, 2003.

[20] N. Du, B. Wu, L. Xu, B. Wang, and X. Pei, "A Parallel Algorithm for Enumerating all Maximal Cliques in Complex Network," *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM'06)*, pp. 320-324, 2006.

[21] H. Ito, K. Iwama, and T. Osumi, "Linear-Time Enumeration of Isolated Cliques," *Proceedings of 13rd Annual European Symposium on ALgorithms (ESA'05)*, pp. 119-130, 2005.

[22] Z. Nutov, "An almost $O(\log k)$-approximation for k-connected subgraphs," *Proceedings of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms*, pp. 912-921, 2009.

[23] The Boost Graph Library, `http://www.boost.org/doc/libs/1_41_0/libs/graph/doc/index.html`.

[24] Stock Price Indices, `http://www.economagic.com/sp.htm`.

[25] Heart Rate Time Series, `http://ecg.mit.edu/time-series/index.html`.

[26] Carnegie-Mellon MoCap Database, *Graphics Lab, Carnegie-Mellon University*, `http://mocap.cs.cmu.edu`.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)