2011 First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology

Salient Segmentation of Medical Time Series Signals

Jonathan Woodbridge, Mars Lan, and
Majid Sarrafzadeh
Computer Science Department
UCLA
Los Angeles, Ca USA
{jwoodbri, marslan, majid}@cs.ucla.edu

Alex Bui
Department of Radiological Sciences
UCLA
Los Angeles, Ca USA
buia@mii.ucla.edu

Abstract—**Searching and mining medical time series databases is extremely challenging due to large, high entropy, and multidimensional datasets. Traditional time series databases are populated using segments extracted by a sliding window. The resulting database index contains an abundance of redundant time series segments with little to no alignment. This paper presents the idea of "salient segmentation". Salient segmentation is a probabilistic segmentation technique for populating medical time series databases. Segments with the lowest probabilities are considered salient and are inserted into the index. The resulting index has little redundancy and is composed of aligned segments. This approach reduces index sizes by more than 98% over conventional sliding window techniques. Furthermore, salient segmentation can reduce redundancy in motif discovery algorithms by more than 85%, yielding a more succinct representation of a time series signal.**

Keywords:Time series signals, Segmentation, Indexing, Data mining

I. Introduction

The advent of remote and wearable medical sensing has created a dire need for efficient medical time series databases. Wearable medical sensing devices provide continuous patient monitoring by various types of sensors, such as accelerometers for activity monitoring; electrocardiogram (ECG) for heart monitoring; and pulse oximeters for blood oxygen saturation monitoring. These devices have the potential to create massive amounts of data. For example, there are currently over 3 million people worldwide implanted with a pacemaker [1]. If these systems had the ability to gather, store, and transmit an ECG signal, we could expect to receive over 560 terabytes of data per day, just from individuals with pacemakers (assuming a three channel ECG, sampling rate of 360 Hz, and 2 byte ADC). Medical data is also extremely time sensitive, requiring timely analysis from healthcare professionals to detect potential health emergencies. Therefore, medical time series databases must be able to store and index large datasets to enable information retrieval tasks that can promptly extract information.

Indexing time series signals to search and mine is an extremely difficult problem due to high dimensionality, high entropy, and massive datasets. Previous works have addressed these issues through reduction techniques (such as Piecewise Aggregate Approximation (PAA) [2], Fourier transform reductions [3], and Chebyshev polynomials [4]) and indexing techniques (such as special R*-Trees [5][6] and iSAX [7][8]).These techniques populate the index with dimensionally reduced segments extracted from raw time series signals. Segments are extracted by determining a window of size m that is slid along the time dimension, indexing each possible segment. Each slide (or translation) of the window is of D data points where $D \geq 1$. However, indexing in this manner leads to redundancy, as the same pattern may be indexed more than once with different translations.

This paper replaces the sliding window with a preprocessing segmentation technique. This technique generically segments a raw time series signal, returning only those segments determined as interesting (i.e., salient). Here, saliency is defined as the least probable segments within a region of interest. Time series signals are modeled as a Markov chain to calculate the probability of each segment's occurrence ($P(w_i)$). Specifically, saliency is defined as the negative log of each window's probability ($-\log P(w_i)$). Windows with a higher saliency than neighboring windows (local maximums) are considered salient and are inserted into the index. All other high probability segments are ignored.

Limiting the index to only salient segments improves data mining and search performance. Fig. 1 presents the top five ECG motifs returned by the motif discovery algorithm in [9]. Notice that there are only two unique patterns and three redundant patterns. Returning additional motifs only yields a larger majority of redundant segments. Medical time series signals are particularly prone to motif redundancy due to their cyclical nature. Hence, finding the true makeup of a medical time series signal using traditional motif discovery algorithms is difficult, at best. Salient segmentation reduces the amount of redundancy in characteristic motifs by more than 85%, yielding a more tangible representation of a medical time series signal. As such, signal search is improved by reducing queries to salient segments. A search of the index is initiated by selecting a region of interest (ROI) in a time series signal. An ROI is often very large and contains many segments. However, searching the index for all segments within an ROI can inundate the user with largely redundant

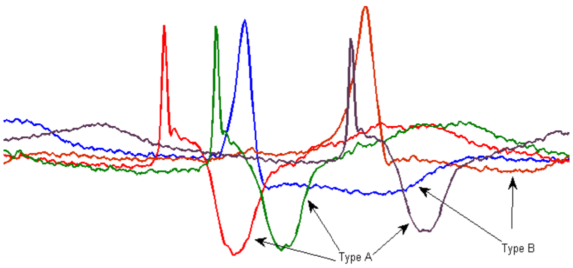

Figure 1. Top five motifs from a 2 minute ECG recording returned by the motif discovery algorithm in [9]. The results consist of only two unique patterns.

978-0-7695-4407-6/11 $26.00 © 2011 IEEE
DOI 10.1109/HISB.2011.41

results. In addition, an ROI may contain many common and uninteresting segments (such as an accelerometer at rest). Returning matches to such non-salient segments will yield little to no use to a user. In preliminary testing, the presented salient segmentation framework decreases the overall number of segments in a time series signal (and ROI) by more than 98%. Limiting searches to only salient segments within an ROI not only reduces the number of returned results, but also improves the quality of those results by ignoring redundant and uninteresting segments (thereby increasing relevancy of results).

Salient segmentation offers the following three properties:

1. All salient patterns are segmented;
2. All salient patterns are segmented consistently (i.e., alignment); and
3. Near linear algorithmic complexity

The first property ensures that all segments that are similar to a salient segment are also labeled salient. The second property ensures alignment and therefore removes redundancy. Salient patterns should only be indexed once unless a translation of that pattern adds significant information. The third property is required for large datasets: a high complexity algorithm would not only drain resources, but introduce a large delay from the time data was received to the time it becomes available for information retrieval tasks. Salient segmentation offers a near linear time algorithm, allowing efficient processing of a time series signal.

The remainder of this paper is organized as follows. Section II presents related work in information retrieval in time series data. Section III details the proposed algorithm and optimizations to improve performance. Section IV presents our experimental setup and Section V presents the results. A discussion of results and conclusions are in Sections VI and VII respectively.

II. Background

A. Definitions

Time series signal: A time series signal $T = \{t_1, t_2, \dots, t_n\}$ is composed of an ordered set of n points in the time domain.

Window: A window w_i is a contiguous set of points in $T$of length m where $m \leq n$ (generally $m \ll n$). A time series of size n has a total of $n - m + 1$ unique windows.

Segment: A segment s_i is a contiguous set of points in $T$of length m' where $m \leq m' \leq n$ (generally $m' > m$and $m' \ll n$). A segment s_i encapsulates a window w_i such that s_i is centered over w_i.

Window saliency: Saliency is a measure of the "interestingness" of a window. More interesting windows are assigned a higher saliency than less interesting windows.

Salient segmentation: Salient segmentation is the process of extracting the most salient (or interesting) segments from a time series signal.

Salient time series index: An index composed of only-salient segments from a time series signal.

B. Signal Search and Index

The GEMINI model proposed in [3] is a process for indexing and searching time series signals. In GEMINI, a time series signal is broken down into a series of segments. Each segment is dimensionally reduced and inserted into a multidimensional indexing structure.

Optimal reduction algorithms have the property of minimum bounds. Minimum bounds guarantee the distance between two reduced segments is less than or equal to the distance between the corresponding raw segments:

$$Dist(A_{reduced}, B_{reduced}) \leq Dist(A_{raw}, B_{raw}) \quad (1)$$

The minimum bounds property for distance guarantees that there are no false dismissals. Therefore, the set of all search results is a superset of all true matches. GEMINI based reduction algorithms include PAA [2], DFT [3], and Chebyshev Polynomials [4].

GEMINI methods segment time series signals with a sliding window approach. A window of size m is translated along the time dimension by a fixed D data points (where $D \geq 1$). Such an indexing scheme leads to oversized and redundant indices. GEMINI methods overcome the size of indices with multidimensional indexing structures such as R*-Trees [5] and iSAX [7][8]. However, little is done to remove the redundancy of entries within the index.

Users often search indices for regions of interests (ROIs). An ROI is a subsequence within a time series that is larger than a segment, but much smaller than the total length of a time series signal. ROIs contain many segments. However, searching for all segments within an ROI yields a large number of results many of which are redundant. A more practical search method chooses the most salient segments from an ROI (as done by this paper). If only salient segments are extracted from an ROI, then non-salient indexed segments serve little to no use. Hence, indices should remove non-salient segments to decrease index sizes and improve search speed.

The authors in [10] remove redundancy by localizing important signal features by characteristic landmarks. For example, local maximums and minimums could be treated as landmarks. Their method was shown to yield good reduction in the data and index size while keeping a low reconstruction error. However, the choice of landmarks is highly domain-specific, disallowing a generic solution. For instance, the definition of local maxima and minima may change dependent on the type of signal and the amount of noise. Some signals, such as EMG (electromyography) or voice, have little use for maxima/minima – these high frequency signals are better characterized by frequency components. Hence, a new set of landmarks must be derived for each type of signal and requires a varying level of optimization depending on the quality of the signal.

C. Motif Discovery

Motif discovery finds sets of similar segments in a time series signal [11]. Each segment in a time series signal is compared to all other segments. Matches are defined as two segments with a distance less than or equal to a given threshold, thr. The top motif is the segment with the most matches. The k-motif algorithm returns the k segments with the most matches. A simple quadratic algorithm to find the top motif is given in Fig. 3.

Motif discovery algorithms suffer from redundancy. Motif results are dominated by similar motifs with different translations. Such redundancy is exacerbated by cyclical and repetitive signals such as medical time series. Redundancy is mitigated by the removal of trivial matches. Two segments $\hat{\imath} = [i, i + m']$ and $\hat{\jmath} = [j, j + m']$ are trivial matches if $dist(\hat{\imath}, \hat{\jmath}) \leq thr$ and there is no segment $\hat{\imath}' = [i', i' + m']$ where $i \leq i' \leq j$ and $dist(\hat{\imath}, \hat{\imath}') > thr$ [11].

Removing such trivial matches is too passive of a filter as trivial matches generally result in the removal of very few segments. A more aggressive approach is to set a constant w such that any segment $\hat{\imath}'$ is trivial with respect to $\hat{\imath}$ if $|\hat{\imath} - \hat{\imath}'| \leq w$ [9]. But a constant w is too rigid of a measure;and a small w does very little to remove redundancy while large windows remove motifs that may not actually be trivial.

D. *Salient Segmentation in Imaging*

Saliency has been successfully used in image processing. For example, the Scale Invariant Feature Transform (SIFT) algorithm [12] computes a series of Gaussian convolutions over an image. SIFT takes the distance between convolutions to find large gradients that are deemed salient. Gradients from unknown images can be compared to gradients from known images to perform classification.

III. Salient Segmentation

A. *Algorithm*

A time series signal is modeled as a Markov chain such that the following property holds.

$$Pr(T_{n+1}|\, T_1 = t_1, T_2 = t_1, \dots, T_n = t_n) = Pr(T_{n+1}|\, T_n = t_n) \quad (4)$$

The transition distribution for a time series signal is calculated by taking a histogram of all possible transitions for each state. However, calculating the transition distribution for each state individually is memory inefficient and requires an extremely large signal to ensure an accurate distribution. While a time series database may be significantly large, individual time series may not. Fortunately, the transition distributions for proximal states are quite similar for most time series signals. Using this observation, close states are grouped together such that each grouping has an equal number of samples to estimate their respective transition distributions. Grouping is accomplished by first estimating the distribution of states. Next, the state space is divided into groups such that each group has an equal probability (i.e., the area under the probability curve for each group is equal). Fig. 2 gives an example grouping with an ECG signal. Each shaded region represents a distinct grouping (or bin) and has an equal probability of occurrence using the distribution displayed to the left.

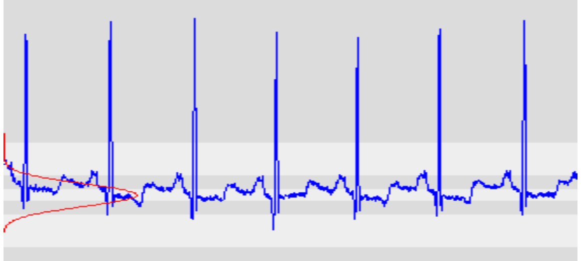

Figure 2. The state probability ditribution function is displayed to the left. Each of the seven shaded regions represents a distinct group where each group has an equal probability of occurrence. The transition distribution function is calculated for each distinct group resulting in seven difference transition distributions.

Each state T_i uses the transition distribution computed for its corresponding bin Bin_{T_i} and is defined as:

$$P(T_i|T_{i-1} = t_{i-1}) = P(T_i|Bin_{T_{i-1}} = Bin_{t_{i-1}})(5)$$

Note that grouped states are only used for the prior state and not for the transition states. The transition states use the entire range to improve sensitivity.

Each point t_i in a time series corresponds to a window w_i where w_i includes the points $[t_{i-\lfloor m/2 \rfloor}, t_{i+\lfloor m/2 \rfloor}]$. The probability for each window w_i is calculated as:

$$P(w_i) = \prod_{j=i-\lfloor m/2 \rfloor}^{i+\lfloor m/2 \rfloor} P(T_j|Bin_{T_{j-1}} = Bin_{t_{j-1}}) \quad (6)$$

With each point's corresponding saliency as:

$$Saliency_i = -\log P(w_i) \quad (7)$$

```
$function$ FindTopMotif($S, m', thr$)

    $topMotifCount = 0$
    $topMotifIndex = 0$

    $for\ i = 1$: $\textbf{length}(S) - m + 1$
        $count = 0$
        $for\ j = 1$: $\textbf{length}(S) - m + 1$
            $if\ thr \geq \textbf{dist}(S(i{:}\, i + m'), S(j{:}\, j + m')$
              $count{+}{+}$
            $end$
        $end$

        $if\ count \geq topMotifCount$
           $topMotifCount = count$
           $topMotifIndex = i$
        $end$
    $end$

    $return\ topMotifIndex$
$end$
```

Figure 3: Top Motif Algorithm

```
$function$ SalienctSegment($inputSignal, bins, m, m', \alpha, \delta$)
    $range =$ QuantizeRange($inputSignal, bins$)

    $pdf =$ ComputeTransitionDistribution($inputSignal, range$)

    $tdf =$ ComputeTSF($inputSignal, pdf, m$)

    $filteredTDF =$ RDP($tdf, \alpha$)

    $salientPoints =$ FindMaximums($filteredTDF$)

    $segments = ExtractSegments(salientPoints, \delta, m')$

    return $segments$
$end$
```

Figure 4: Salient segmentation algorithm

A time series saliency function (TSF) is constructed by concatenating each successive point's saliency such that:

$$F_{saliency} = \{Saliency_1, Saliency_{2,} \dots, Saliency_{n-m+1}\} \quad (8)$$

Each local maximum i in $F_{saliency}$ is considered salient and its corresponding segment ($[t_{i-\lfloor m'/2 \rfloor}, t_{i+\lfloor m'/2 \rfloor}]$) is inserted into the index. However, saliency functions often contain a significant amount of noise, thus resulting in over segmentation (too many maximums).

A set of linear approximations calculated by the Ramer-Douglass-Peucker (RDP) algorithm [13][14] is used to filter the TSF. The RDP algorithm begins with a linear approximation with endpoints at the first point p_1 and the last point p_n of the TSF. Next, the distance between the linear approximation and each point between the first and last point is calculated. If the point with the largest distance p_i is above a given threshold, thr, the signal is estimated by two linear approximations: $p_1 \Rightarrow p_i$ and $p_i \Rightarrow p_n$. The algorithm is repeated on both segments and continues until no point is more than thr from its linear approximation. Here, the value of thr is a function of the standard deviation of the estimated TSF:

$$thr = \alpha\sigma \quad (9)$$

TSF's with higher standard deviations generally have both a larger amount of noise and disparity between peaks and valleys. Hence, TSFs with a large standard deviation require a more aggressive filter. A larger value of α will result in fewer maximums while a smaller value will result in more maximums. This paper assumes $\alpha = 1$.

The RDP approximation of the TSF results in a slight misalignment between segments. This means that the salient points' locations calculated after the RDP approximation has a slight variation from the true locations. Therefore, an additional δ points are added before and after a segment s_i to create an elastic window defined by the range $[i - \lfloor m'/2 \rfloor - \delta, i + \lfloor m'/2 \rfloor + \delta]$.When the segment is searched, a window of size m' will compare all windows located between $i - \lfloor m'/2 \rfloor - \delta$ and $i + \lfloor m'/2 \rfloor + \delta$. As $\delta \ll n$, the elastic window results in only a small decrease in performance.

The relationship between the elastic window, segment and window is shown in Fig. 5. The salient segmentation algorithm is summarized in Fig. 4.

B. Complexity

Quantization of the range and calculation of the transition distribution is done in linear time. The TSF is computed in time $O(mn)$ in Equation 6. However, the TSF can be calculated as a sliding window where each slide results in one division and one multiplication yielding an $O(n)$ runtime. The RDP filtering technique has an average runtime of $O(n \log n)$ and has been improved to have an upper bound of $O(n \log n)$ in [15]. Finding maximums and extracting segments can both be done in linear time yielding an overall upper bound of $O(n \log n)$.

IV. Experimental Setup

Two experiments were conducted: search and motif discovery. The search experiment demonstrates the first two properties for salient segmentation. The motif discovery experiment further proves the second property by demonstrating the removal of redundancy by salient segmentation. The third property was proven in Section III.

The datasets used by this paper are as follows:

1. MIT-BIH Arrhythmia Database (ECG) [16]. This dataset contains several 30-minute segments of two-channel ambulatory ECG recordings. These sample included arrhythmias of varying significance.
2. Gait Dynamics in Neuro-Degenerative Disease Database [17][18]. This dataset contains data gathered from force sensors placed under the foot. Healthy subjects as well as those with Parkinson's disease, Huntington's disease, and amyotrophic lateral sclerosis (ALS) were asked to walk while the data was recorded. Data includes 5-minute segments for each subject.
3. WALK. This dataset contains a series of annotated recordings from a tri-axial accelerometer worn in a subject's pants pocket. Data was recorded while subjects travelled through the interior of a building.

No reduction algorithms or advanced filtering was used in the analysis of salient segmentation. Each segment was inserted into the index with only a moderate low pass filter (non-weighted averaging window). Reduction algorithms and advanced filtering techniques were excluded to remove any biasing of results.

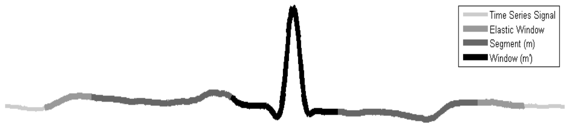

Figure 5. Displays the relation of a window, segment, and elastic window.

A. Search

An index was created for each signal in the test data sets using the salient segmentation technique. Each segment in the index was compared to all possible segments in its respective time series signal using a sliding window with $d = 1$. The closest matches from the sliding window were stored as the true closest matches for each segment. Next, each segment in the index was compared to all other segments in the index. The closest matches in the index were compared to the sliding window's closest matches to create precision-recall curves.

Each dataset was run with three elastic windows: 5, 10, and 20 data points. The elastic window was introduced to account for misalignments resulting from the linear approximations. Additional segmentation parameters for each dataset are given in Table 2.

The parameter m was chosen as the average size of an "interesting" pattern. For example, a step in the gait data set (heal down to toe up) was approximately 80 data points. Some care must be taken to avoid rounding errors. In the ECG data set, m was chosen to match the approximate size of the QRS complex (100 data points) instead of the entire heart beat (300-360 data points) to avoid rounding errors. The parameter m' was chosen to give context before and after a salient region. This parameter has no affect on the location of salient points and should be chosen based on the user's need.

B. Motif Discovery

The motif discovery experiment compares results between the motif discovery algorithm in [9] and a modified motif discovery algorithm using only salient segments. These methods were compared with two metrics: redundancy and coverage. Redundancy measures the percentage of the time series signal that is represented by more than one motif (i.e., the total amount of data points that exists are repeated in two or more motifs). Coverage measures the percentage of the time series signal that is represented by the returned set of motifs. Both motif algorithms were run with increasing k (where k is the k closest motifs) until no new motifs were returned. The parameters listed in Table 1 were used for the experiment.

TABLE 2: SEGMENTATION PARAMETERS

Dataset	P	m	Filter Size	m'
ECG	8	100	10	600
Gait	4	80	10	600
Walk	2	25	5	300

TABLE 1: MOTIF DISCOVERY PARAMETERS. ALL OTHER PARAMETERS ARE IDENTICAL TO TABLE 2

Dataset	m'	R	Reference Points
ECG	300	2	10
Gait	80	2	10
Walk	60	2	10

Medical time series are cyclic by nature. The parameters chosen to segment the signal were tuned to find individual cycles, such as one heartbeat or one step. The time series signals used for this paper contain 90-100% activity. Therefore, motif results should contain high coverage as most of the signal contains interesting patterns.

To assess redundancy, m' was reduced from the search experiment. m' is set to the average complete cycle time for each signal. For example, one heartbeat takes approximately 300 samples in the ECG dataset. Two steps in the WALK dataset (left and right) take approximately 60 data points, and one step in the gait dataset takes 80 data points. Only one step is used for the gait dataset as each channel measures only one foot. Reducing m' focuses the comparison on both methods' abilities to isolate individual cycles. A small m' should result in low overlap. Note, m' has no affect on localizing salient points in salient segmentation as shown in Equation 6.

The modified motif discovery algorithm finds the two closest segments $\hat{\imath}$ and $\hat{\jmath}$ in the index such that $dist(\hat{\imath},\hat{\jmath}) = \min_{\forall a,b \in I} dist(\hat{a},\hat{b})$ (where I denotes the set of all segments in the index). Any segment $\hat{\imath}'$ with $dist(\hat{\imath}',\hat{\imath}) \leq 2 \cdot dist(\hat{\imath},\hat{\jmath})$ or $dist(\hat{\imath}',\hat{\jmath}) \leq 2 \cdot dist(\hat{\imath},\hat{\jmath})$ are deemed to be in the neighborhood of $\hat{\imath}$ or $\hat{\jmath}$ respectively and are grouped together with $\hat{\imath}$ and $\hat{\jmath}$. All segments in the neighborhood of $\hat{\imath}$ and $\hat{\jmath}$ are removed from the search and the algorithm is repeated until no segments are left. The algorithm was modified from the classical motif discovery algorithm presented in Section II.C to closely resemble more recent work in [9]. The algorithm in [9] utilizes a fixed window w such that any segment $\hat{\imath}'$ where $|\hat{\imath}' - \hat{\imath}| \leq w$ is a trivial match to $\hat{\imath}$. The experiments for [9] were repeated for $w = m', m'/2$, and $m'/4$.

V. RESULTS

A. Search

Precision-recall curves for the WALK, ECG, and gait datasets are shown in Fig. 6. Increasing the elastic window size has a small improvement for the gait and ECG datasets. The WALK dataset has a small improvement with a 10-point elastic window and decreased improvement with a 20-point elastic window.

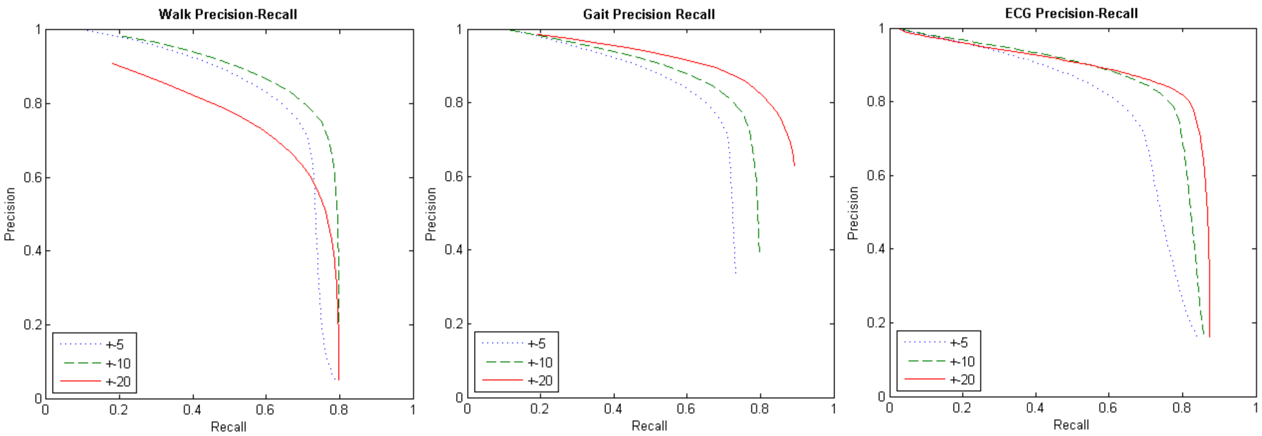

Figure 6: Displays precision-recall for the Walk, Gait, and ECG datasets. An elastic window of 5, 10, and 20 data points is displayed for each signal. An increasing elastic window has a small improvement demonstrating salient segmentation's ability to consistently localize salient points.

There are two sources for the variability in the locations of salient points. First, the RDP approximation adds variability to the location of salient points. Second, similar patterns are not necessarily exact. Therefore, the calculation of saliency may yield slightly different locations of the most salient points in similar patterns. However, both sources of variability are quite small, requiring only small elastic windows to correct alignment.

Precision-recall results will eventually decrease as the elastic window expands. With a constant m', increasing the elastic window increases the probability that a new pattern (not originally localized by salient segmentation) may be matched. The WALK dataset has a quicker drop-off in performance with respect to the elastic window due to smaller pattern sizes (in terms of data points). The average pattern size (one step) for the WALK dataset is approximately 20-25 data points. An elastic window of 20 or more data points will include an additional pattern (or step) on each side of the isolated segment. These additional patterns cause false positives for segments that lie close to true matches. This phenomenon is shown in Fig. 6. The WALK data shows a decrease in precision with a large elastic window, but recall is not affected.

The RDP linear approximation of the TSF curve added little variability to search performance. However, close inspection of the TSF curve reveals that the RDP algorithm suppressed a small percentage of peaks (salient points). This suppression resulted in a minor degradation in recall performance. The gait dataset had the simplest time series signals (lowest entropy), with large differentials at the beginning and end of patterns. This causes large peaks in the TSF curve, resulting in the filtering of very few salient points. In contrast, the WALK dataset had the most variable signal (highest entropy) with the smallest differential between the start and end of patterns. The relationship between entropy and recall is shown in the figures with increased recall for lower entropy signals.

The precision-recall results are notable, considering the number of segments suppressed by salient segmentation. Table 3 shows the percentage of indexed signal for each time series data set (the elastic window was not included for coverage calculations). Salient segmentation resulted in index sizes below 2% of a sliding window index. However, segments within the index spanned near 100% of each time series signal.

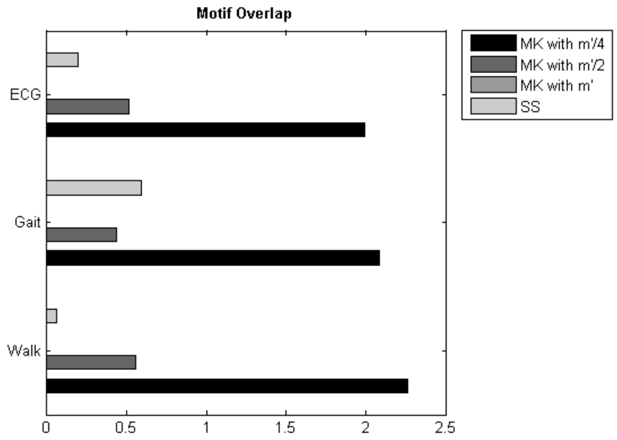

Figure 7. Displays the percentage of time series that is represented more than once.

B. *Motif Discovery*

The top 5 motifs for signals from each of datasets for both salient segmentation and the motif discovery algorithm in [9] are given in Fig. 8. Motif discovery with salient segmentation yields aligned motifs, while the comparison algorithm from [9] yields motifs in an arbitrary alignment. Misaligned results demonstrate a poor representation of a signal's

TABLE 3. INDEX SIZE AND COVERAGE

Dataset	Index Size	Coverage
ECG	0.4%	99.9%
Gait	1.8%	96.7%
Walk	1.6%	98.3%

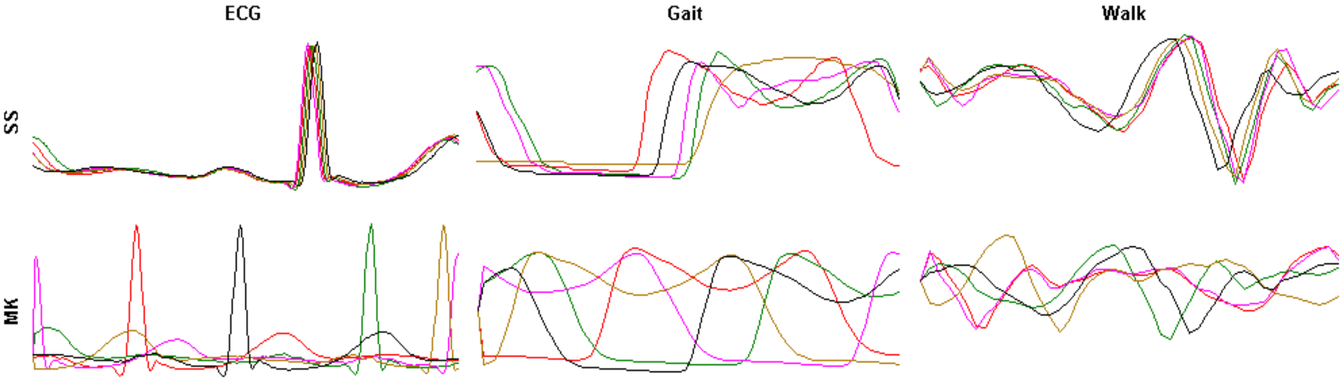

Figure 8. Displays the top 5 motifs returned by the motif discovery algorithm with salient segmentation (SS) and the motif discovery algorithm in [9] (MK). Salient segmentation returns sets of aligned motifs unlike [9].

true makeup. By way of illustration, the results for motif discovery with salient segmentation show that all three signals are composed largely of the same pattern. The results from [9] appear to have five distinct motifs for each signal – but all five patterns are extremely similar and are just returned in different alignments.

Overlap and coverage for all three data sets are shown in Fig. 7 and Fig. 9 respectively. The algorithm in [9] (denoted as MK) shows no overlap when using a window of size m′ to filter trivial matches. This is expected as any points lying m′ data points from a motif are considered invalid for future rounds of motif discovery, therefore avoiding overlap. Also expected, decreasing m' increases both overlap and coverage. Salient segmentation yields lower overlap than the algorithm in [9] for the ECG and WALK datasets, and similar overlap for the gait data.

The gait dataset resulted in a significantly higher overlap than the WALK and ECG datasets for salient segmentation. The gait dataset measures pressure from shoes as subjects walked. The on-off pressure is sharp and abrupt, often resulting in two salient points for each step (heel down and toe up). This double segmentation of some steps results in an increased amount of overlap.

Reduced coverage was exhibited for the gait and WALK datasets. Poor coverage had two causes. First, both datasets have small regions of no activity (e.g., such as the subjects standing still). These regions account for 2-10% of these datasets and result in no segmentation within these regions. Therefore, coverage is extremely low in these regions (as expected). Second, the WALK and gait datasets had the highest variability in the length of a pattern cycle. For example, the WALK dataset comprised subjects traversing hallways, and ascending/descending stairways, resulting in variable step lengths. The gait dataset's subject pool consisted of neurologically impaired patients (such as Parkinson's disease and ALS), resulting in inconsistent gait. Coverage can be improved by decreasing the α parameter for the RDP algorithm (Eq. 9) or increasing the size of m'. However, these changes have a trade-off of with redundancy (overlap).

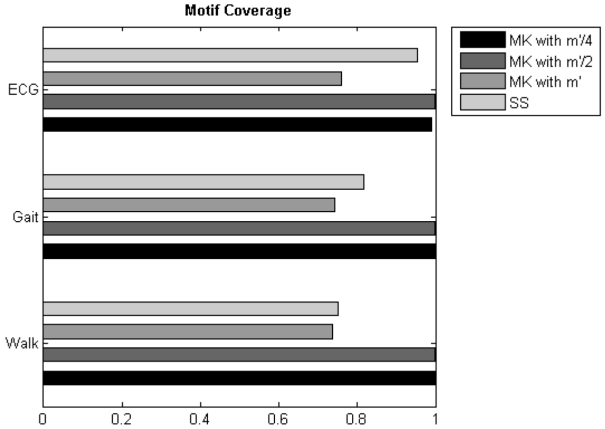

Figure 9. Displays the percentage of time series signal represented by the motif discovery algorithms.

The overlap and coverage results are particularly compelling as no assumptions are made about trivial matches when using salient segmentation. All and only salient segments are matched in the motif discovery algorithm. No assumption is made on the proximity of one segment to another removing the arbitrary measurements proposed in [9][11]. These results are even more encouraging when considering the alignment offered by salient segmentation. All similar motifs are in alignment vastly improving the quality of motif discovery of algorithms from those in [9][11].

VI. Discussion

This paper presented experiments with a fixed parameter m. This parameter was fixed due to prior knowledge of the test datasets. Real world systems will not have as much *a priori* knowledge of an incoming signal. For instance, the ECG dataset consists of heartbeats taken during low activity periods. A more realistic ECG dataset consists of more variable heartbeats; and in general, medical databases must expect a higher variability of incoming signals. Fortunately, salient indices are relatively small and the algorithmic complexity of salient segmentation is low. Therefore, running salient segmentation on several values of m can populate the index with differing levels of granularity to account for variability.

As shown with the gait dataset, salient segmentation sometimes suffers from over-segmentation. Such segmentation results in some redundancy over multi-segmented patterns. Some care should be taken when mining for patterns to avoid misleading results from redundancy. However, the redundancy is manageable due to the alignment property offered by the salient segmentation framework. For example, steps in the gait dataset that are segmented twice would exist with two different, but consistent, alignments. When using the algorithm in [9] similar segments will lie in arbitrary alignments resulting in arbitrary measures to remove redundancy (such as fixed windows for filtering trivial matches).

VII. Conclusion

This paper presented salient segmentation for medical time series signals. Salient segmentation models time series signals as a Markov chain. The probability of each segment within the signal is computed using this model. Segments with the lowest probabilities within a local region are considered salient and are inserted into the index. All other higher probability segments are ignored. Salient segmentation consistently segments similar patterns with similar alignments and runs with a $O(n \log n)$ complexity.

Salient segmentation probabilistically determines the most important features of a time series signal. Constraining searches to only these most important features improves the quality of search while reducing redundancy. Search performance was assessed using precision-recall curves on the gait, WALK, and ECG datasets. All three signals displayed both high precision and high recall.

Redundancy and consistency of segmentation was also assessed by comparing motif discovery with salient segmentation to the motif discovery algorithm in [9]. Salient segmentation had a consistently lower amount of redundancy with high signal coverage. Motif results were also aligned, unlike [9], yielding a good representation of the time series signal.

Acknowledgment

The authors would like to thank Professor Eamonn Keogh for providing us with their implementation of the motif discovery algorithm presented in [9]. This publication was supported by Grant Number T15 LM07356 from the NIH/National Library of Medicine Medical Informatics Training Program.

References

[1] M. A. Wood, K. A. Ellenbogen. Cardiac Pacemakers From the Patient's Perspective. American Heart Association, 105(18):2136-8, 2002.

[2] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and Information Systems, 3(3):263–286, 2001.

[3] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. ACM SIGMOD Record, 23(2):419–429, 1994.

[4] Y. Cai and R. Ng. Indexing spatio-temporal trajectories with Chebyshev polynomials. In Proceedings of the 2004 ACM SIGMOD international conference onManagement of data, page 610. ACM, 2004.

[5] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger. The R*-tree: an efficient and robust access method for points and rectangles. ACM SIGMOD Record, 19(2):322–331, 1990.

[6] A. Guttman. R-trees: a dynamic index structure for spatial searching. In Proc. of the SIGMOD Conf., pages 47-57, 1984.

[7] J. Shieh and E. Keogh. iSAX: indexing and mining terabyte sized time series. In Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 623–631. ACM, 2008.

[8] J. Shieh and E. Keogh. iSAX: disk-aware mining and indexing of massive time series datasets. Data Mining and Knowledge Discovery, 19(1):24–57, 2009.

[9] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover. Exact discovery of time series motifs. In Proceedings of the SIAM International Conference on Data Mining (SDM 2009), pages 473–484. Citeseer.

[10] C.S. Perng, H. Wang, S.R. Zhang, and D.S. Parker. Landmarks: a new model for similarity-based pattern querying in time series databases. In icde, page 33, 2000.

[11] J. Lin, E. Keogh, S. Lonardi, and P. Patel. Finding motifs in time series. Proc. of 2nd Workshop on Temporal Data Mining (KDD'2), 2002.

[12] D.G. Lowe. Distinctive image features from scale invariant keypoints. International journal of computer vision, 60(2):91–110, 2004.

[13] U. Ramer. An interative procedure for the polygonal approximation of plane curves. Computer Graphics and Image Processing, pages 244-256. Elsevier, 1972.

[14] D. Douglass and T. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its Caricature.The Canadian Cartographer 10(2), 112–122 (1973).

[15] J. Hershberger and J. Snoeyink. An O (n log n) implementation of the Douglas-Peucker algorithm for line simplification. Proceedings of the tenth annual symposium on Computational geometry, pages 383-384. ACM, 1994.

[16] G. Moody and R.G. Mark. The MIT-BIH Arrhythmia Database CD-ROM. Proceedings Computers in cardiology: Chicago, Illinois, September 23-26, 1990, page 185, 1991.

[17] J.M. Hausdorff, A. Lertratanakul, M.E. Cudkowicz, A.L. Peterson, D. Kaliton, and A.L. Goldberger. Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. Journal of applied physiology, 88(6):2045, 2000.

[18] J.M. Hausdorff, S.L. Mitchell, R. Firtion, CK Peng, M.E. Cudkowicz, J.Y. Wei, and A.L. Goldberger. Altered fractal dynamics of gait: reduced strideinterval correlations with aging and Huntington's disease. Journal of Applied Physiology, 82(1):262, 1997.