Segmenting Sound Waves to Support Phonocardiogram Analysis: The PCGseg Approach

Hajar Alhijailan[1,2], Frans Coenen[3], Jo Dukes-McEwan[4], and Jeyarajan Thiyagalingam[5]

[1] Department of Computer Science, The University of Liverpool, Liverpool, UK
[2] College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
h.alhijailan@liverpool.ac.uk
[3] Department of Computer Science, The University of Liverpool, Liverpool, UK
coenen@liverpool.ac.uk
[4] Small Animal Teaching Hospital, The University of Liverpool, Leahurst Campus, Neston, UK
jdmcewan@liverpool.ac.uk
[5] Department of Electronics and Electrical Engineering, The University of Liverpool, Liverpool, UK
tjeyan@liverpool.ac.uk

Abstract. The classification of Phonocardiogram (PCG) time series, which is often used to indicate the heart conditions through a high-fidelity sound recording, is an important aspect in diagnosing heart-related medical conditions, particularly on canines. Both the the size of the PCG time series and the irregularities featured within them render this classification process very challenging. In classifying PCG time series, motif-based approaches are considered to be very viable approach. The central idea behind motif-based approaches is to identify reoccurring sub-sequences (which are referred to as motifs) to build a classification model. However, this approach becomes challenging with large time series where the resource requirements for adopting motif-based approaches are very intensive. This paper proposes a novel two-layer PCG segmentation technique, called as PCGseg, that reduces the overall size of the time series, thus reducing the required for generating motifs. The evaluation results are encouraging and shows that the proposed approach reduces the generation time by a factor of six, without adversely affecting classification accuracy.

1 Introduction

Time series analysis is concerned with the extraction of important parameters from a given time series, which is an extension of point series data [1]. The distinguishing feature of such data is that it comprises an ordered sequence of values. Although there a whole class of problems associated with time series analysis, the work presented in this paper is directed towards the classification of canine Phonocardiogram (PCG) time series. Some example time series fragments are given in Fig. 1. Within PCG time series, certain recurring sub-sequences, known as motifs, characterize certain heart conditions among canines. As such, identifying these motifs is a basic requirement prior to more detailed diagnosis.

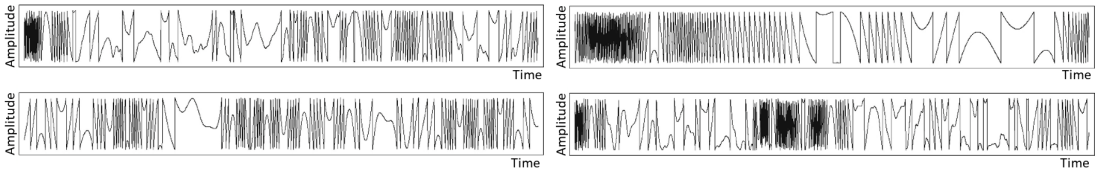

Fig. 1. PCG time series examples

A number of motif discovery techniques have been proposed in the literature [2–5]. However, in the context of PCG data, two significant issues are: (i) the time to generate/identify these motifs and (ii) the accuracy of resulting classification. The first problem is often exacerbated by the size of PCG time series. The second is concerned with the quality of the identifiable motifs, so that they are appropriate enough to be reasonable differentiators of different classes. These two issues are inter-linked, albeit being data dependent. And hence, if motif generation can be made more efficient, additional resource will be available to select "better" motifs given a time constraint.

A suggested solution to the first, and consequently provide for the second, is to pre-process the time series so as to reduce their size in such a way that salient features are preserved [6]. One common pre-processing technique is time series segmentation [6]. The basic process is to divide the time series into blocks. The intended advantage is that the resulting segmented time series is much shorter than the original while preserving necessary salient features. From the literature a variety of algorithms have been proposed to obtain a good high level segmentation of time series data [6–12]. An extensive coverage of the literature is provided in Sect. 2.

Amongst all these techniques, the most appropriate segmentation mechanism is often problem-specific and dependent on the application domain under consideration. The application domain at which the work presented in this paper is the classification of canine PCGs according to a variety of heart conditions that can be identified from PCG data. A PCG is a two dimensional time series, where the values are amplitudes. The average number of points (length) in the collected time series was over $355{,}000$; a significant number and too many to be processed easily using established motif generation mechanisms. As mentioned before, some form of pre-processing, segmentation, to reduce the overall file size was thus considered to be appropriate. by adopting this approach, using some form of segmentation, the intention is to reduce the size of the collected time series so that a useful classification model can be generated in an efficient manner. The main contribution is thus a bespoke, two layer, PCG time series segmentation mechanism, PCGseg, which significantly reduces the processing time (by a factor of six) without adversely affecting the resulting accuracy. The approach is fully described and evaluated using a realistic motif-based classification scenario.

The rest of this paper is organized as follows. Section 2 gives a review of the previous work specific to the application domain. A formalism for the problem domain is given in Sect. 3. The proposed segmenting method, which is specifically designed for PCG data recorded in WAVE file format is presented in Sect. 4. Section 5 presents the evaluation of the proposed PCGseg approach. Some discussion is then presented in Sect. 6. This paper is concluded with some summarising remarks and directions for future work in Sect. 7.

2 Previous Work

This section provides a review of existing work concerning the segmentation of time (point) series. The section is divided into two subsections. Subsection 2.1 considers previous work concerning the segmenting of point series, whist Subsection 2.2 examines previous work on PCG segmentation.

2.1 Segmenting Point Series

As noted in the introduction to this paper, segmentation is the process of dividing a whole entity into constituent or distinct elements, the term is frequently used in the context of image analysis [13]. Techniques for achieving effective and efficient segmentation remain an area of current research. The main objective of segmentation is to reduce the amount of data so that the features of interest are retained. In the context of time series, a number of mechanisms have been used to achieve segmentation, these include: (i) Fourier Transforms [9], (ii) Wavelets [10], (iii) Symbolic Mappings [8] and (iv) Piecewise Linear Representation (PLR) [7, 12]. PLR is the most common of these four mechanisms. It is also of relevance with respect to the work presented here because it is achieved in a similar manner (using a moving window approach). The fundamental idea of PLR is to translate a given time series T into a model $\bar{T}$ which comprises a number of "best fitting" straight lines (segments). However, the nature of PCG curves (point series), see Fig. 1, is such that PLR is not appropriate for the PCG application considered in this paper. Line fitting, whatever form this might take, is not sufficiently descriptive for the purpose of PCG classification.

Regardless of the segmentation approach used, each can be implemented using one of three basic mechanisms as follows [6]: Sliding Window, Top Down and Bottom Up. The first approach is faster than the other two. However, decisions are made without any deep exploration of the time series as in the case of the later two approaches; this in turn may affect the quality of the segmentation. Both Top Down and Bottom Up give good results, but they tend to be impractical given large data sets as they require a scan of the entire time series. The method proposed in this paper uses the Sliding Window mechanism, however the segmentation is not related to similarity with the underlying point series but with how accurately the segments represent the "shape" of the underlying point series sub-sequence.

2.2 PCG Segmentation

As noted above, segmentation is the term used to subdivide a signal trace into sub-sequences according to some criteria. A PCG trace can be divided into single heart (cardiac) cycles each comprised of four principal components: (i) first cardiac sound (S1), (ii) *systole* , (iii) second cardiac sound (S2) and (iv) *diastole* . The Extraction of PCG segments from several cardiac cycles (heart beats) is usually achieved with help of an ECG signal recorded at the same time and/or a Carotid Pulse (CP) which can then be used as a reference signal [14, 15]. In the case of the work considered in this paper, no such reference signals are available.

There is some reported research directed at the segmentation of PCGs without a reference signal [16–19] but with some limitations. Some of this research [17] concentrates mainly on finding first and second heart sounds ($S1$ and $S2$) in a "waveshape", which is a reflection of the PCG as a form of five visible deflections per each heart beat. Other work, such as [18], is directed at extracting heart events ($S1$, $S2$, $S3$ and $S4$) in a spectrogram. However, finding heart sounds (events) in a few seconds of a recorded PCG signal is not suited to Time Series Motif Discovery; a much longer time series is required with respect to canine PCG data, the target domain used with respect to the evaluation presented later in this paper. In [20] it was observed that even a 10 second series, featuring 12 to 27 beats for a normal resting dog, is not adequate for motif discovery. In addition, the focus of the majority of the existing work on PCG segmentation is directed at heart event detection in normal cardiac activity using the energy of the signal which in turn is affected by noise [19] or murmurs [16]; whereas this paper aims at classifying regular and irregular cardiac activities.

3 Formalism

A time (point) series P comprises a sequence of n data values $\{p_1, p_2, \ldots, p_n\}$. Using PCGseg a segmented point series S consists of a sequence of m segments $\{S_1, S_2, \ldots, S_m\}$ where each segment S_i represents some sub-sequence of P. Each segment S_i is defined in terms of a tuple of the form $\langle S_p, S_C \rangle$, where S_p is the parent segment and S_C is a set of constituent sub-segments $\{S_{c_1}, S_{c_2}, \ldots\}$. S_P is defined in terms of a tuple of the form: $\langle shape, type, length \rangle$, where: $shape$ is the nature of the point series defined by the segment, $\{slant, vertical, dome, flat\}$; $type$ is the direction of the shape, either up or $down$; and $length$ is the number of points represented by the segment. More specifically the length of a segment is the difference between the start and end index values. Thus a shape comprised of two points will have length 1 and so on. Each element $S_{c_i} \in S_C$ is represented by a second tuple of the form: $\langle type, length, depth \rangle$, where: $type$ is the nature of the point series defined by the sub-segment $\{up, down, flat\}$, $length$ is the length of the sub-segment (calculated as described above); and $depth$ is the difference between the maximum and minimum amplitude values represented by the sub-segment in question. Given this representation a motif M is then some subset of S ($M \subset S$) that repeats and is thus deemed to be representative of the underlying point series.

A collection of x segmented, and labeled, point series is given by $\mathbf{D} = \{D_1, D_2, \ldots, D_x\}$, where each D_i is a tuple of the form $\langle S_i, c_i \rangle$ where S_i is a segmented point series and c_i is a class label taken from a set of class labels C. The aim is to identify a collection of motifs $\mathbf{L} = \{L_1, L_2, \ldots\}$ such that each L_i is a tuple of the form $\langle M_i, c_i \rangle$ where M_i is a motif and c_i is a class label. The set $\mathbf{L}$ can then be used to build a classification model of some kind.

4 PCGseg

The proposed PCGseg algorithm is underpinned by the idea of capturing the "shapes" that exist in a PCG sequence. The fundamental idea is that a PCG time sequence can be conceptualised in terms a series of shapes and sub-shapes (segments and sub-segments). From Fig. 1 four distinct shapes can be identified: (i) slant, (ii) vertical, (iii) dome and (iv) flat. It is also important to note that the vertical shape always occurs between any two other shapes. In other words it can be regarded as a separator; this feature is used in the context of the proposed PCGseg mechanism to identify the start and end points of segments.

As already noted in Sect. 3, a segment is defined by a tuple of the form $\langle shape, type, length \rangle$, and as also already noted above, the possible values for the $shape$ variable are: $\{vertical, slant, dome, flat\}$. Each segment is defined in terms of a conceptual Minimum Bounding Box (MBB) surrounding it. The x-dimension of the MBB of a segment corresponds to the value of the $length$ variable associated with the segment. More specifically, each shape is defined as follows:

1. **slant**: A slant shape comprises a sequence of three or more points ($length \geq 2$) such that the start and end points are at opposite corners of the MBB and the difference between the start and end point amplitude values is greater than a threshold $t2$.
2. **vertical**: A vertical shape is a special case of the slant shape whose $length$ is 1. This shape appears very often, always between two other shapes; it can thus be viewed as a separator.
3. **dome**: A dome shape comprises a sequence of three or more points ($length \geq 2$) such that the start and end points are on the same side (top or bottom) of the associated MBB. This is defined in terms of the difference between the start and end point amplitude values which must be less than a threshold $t2$.
4. **flat**: A flat shape is a special case of the dome shape, comprised of two or more points ($length \geq 1$) and whose depth (maximum difference between amplitudes) is less than a predefined threshold $t3$.

The $type$ of a shape is determined by the "direction" of the shape. The possible values for the $type$ variable are $\{up, down\}$. The value for the $type$ variable is defined by the first two points in the sequence. If the amplitude of the second point is greater than the first, the type is up. If it is less than the first, the type is $down$. Where the first two points have the same amplitude value, a rare occurrence, this is dealt with by considering their location within the overall time series. If both amplitudes are below the average value, a threshold $t4$, they are considered to have the type up and otherwise, the type is $down$.

Whatever the case, a sub-segment, as noted in Sect. 3, is described by a tuple of the form $\langle type, length, depth \rangle$. The possible values for the type variable are: $\{up, down, flat\}$. Note that the values for the $type$ variable associated with a sub-segment are not the same as those associated with a segment. The value for the $type$ variable is defined by all points in the sub-segment. If the amplitude of all points is increasing, the type is up. If it is decreasing, the type is $down$. Otherwise, the type is $flat$.

From the foregoing, we have a set of five thresholds, $\{t1, t2, t3, t4, t5\}$, which are used to segment point series. They are particularly used for detecting vertical shapes, slant shapes, dome shapes, dome and flat shape types and sub-segment types respectively.

The motivation for the proposed PCGseg mechanism was to represent time series, and PCG point series in particular, in terms of their constituent shapes and sub-shapes in a two level hierarchy instead of averaging values taken periodically [21] or extracting trends [22]. The conjectured advantages were that:

1. The main information, which would be lost in the case of the application of the averaging method, is preserved. The significance is that the rate of change in PCG records is quite high. Averaging may still work if a very narrow window is used however this would defeat the objective of the segmentation, to reduce the data volume.
2. A more succinct segmentation would be produced, than that produced by earlier segmentation mechanisms, by considering "parent" and "child" shapes (segments and sub-segments) where the child shapes represent "trends" in the parent shape. Recall that a parent shape can have any number of sub-shapes (trends); this will be the case where the parent shape features many irregularities and fluctuations.
3. Prediction using point series requires a substantial amount of matching of point series sub-sequence. The proposed two-level hierarchical segmentation allows for "early abandonment" where the parent segment does not fit the comparator segment (no need to go down to the next level).

4.1 Motif Detection

Once the entire data set had been segmented, motifs can be identified. The idea was to store the identified motifs, together with an associated class label, in a "bank" of motifs which could then be used to classify (label) previously unseen PCG records. The adopted approach was founded on the MK motif discovery algorithm [23], a well established motif detection algorithm that operates by identifying and testing a number of candidate motifs and selecting the first n where matches are found. The MK algorithm operates using a window of size ω (measured in terms of a number of points) and a reference value r. A sequence of r random windows are generated, of length ω, and a best similarity value is obtained for each by comparing it with all other sub-sequences of length ω in the given point series. The top n are then selected. In [23], $n = 2$ was used, this value was also used with respect to the evaluation reported on later in this paper.

The original version of the MK algorithm, as described in [23], was designed to operate with point series, and used Euclidean Distance as a measure of similarity. However, this measure would clearly be unsuitable in the case of segmented data and thus an alternative distance function was required so that distances between motifs comprised of segments could be obtained. To measure how well two potential

motifs match, five criteria were considered, in turn, in such a way that a policy of early abandonment could be adopted. The first criterion is the number of (parent) segments. The second is segment shapes and types. Then, the number of (children) sub-segments followed by sub-segment types. Lastly, average sub-segment lengths and depths. Thus when comparing two motifs, if any one of the first four criteria was not satisfied, the comparison would be abandoned without further comparison being required. In more detail, both sets of parent segments must be the same shape and type, in the same order. The number of sub-segments that feature in each segment should be identical, as should the order, shapes and types of the sub-segments.

Lastly, the fifth criterion, the Root Mean-Square Distance (RMSD) over the lengths and depths of the sub-segments was calculated and a similarity index, sim, was produced using (1) where X_{c_i} and Y_{c_i} are sub-segments in the two motifs X and Y being compared and j is the number of sub-segments.

$$sim = \frac{\sqrt{\sum_{i=1}^{j} (length_{X_{c_i}} - length_{Y_{c_i}})^2} + \sqrt{\sum_{i=1}^{j} (depth_{X_{c_i}} - depth_{Y_{c_i}})^2}}{j} \quad . \tag{1}$$

The strategy of using five criteria is obviously very strict. This was deemed appropriate for motif discovery. However, unsuitable for measuring the similarity between the test and training data in the classification stage. In early experiments (not reported here), it was found that when using this strict strategy, the vast majority (74%) of the test data was not being classified at all because of the strictness of the matching. Therefore, an alternative, more tolerant, approach was adopted for the classification stage when comparing a motif $M1$ identified in a record to be classified and a motif $M2$ in the bank of labelled motifs extracted from the training data (details not included here for reasons of space limitation).

5 Evaluation

This section presents the evaluation of the proposed PCGseg approach. Recall that the motivation for the proposed approach was to speed up the classifier generation process without loss of accuracy. The operation of the proposed PCGseg approach was compared with the use of unsegmented data in terms of a PCG multi-class classification problem (the data set used is described in Subsection 5.1). In both cases, the MK algorithm [23], presented earlier, was used to generate motifs. The evaluation was conducted in terms of accuracy and runtime; runtime to establish whether the PCGseg approach was faster or not, and accuracy to determine whether adoption of the PCGseg approach had a negative effect on accuracy or not. Recall also that the MK algorithm operates using two parameters: (i) the desired window size ω and (ii) the number of candidates to be generated, r, known as the *reference value*. Using the PCGseg approach, the value for ω was defined in terms of a number of segments, a range of values for ω was considered $\{25, 50, 75\}$ with respect to the evaluation reported here. With respect to the comparator approach, using unsegmented data, ω was defined in terms of a number of points, a range of values for ω was considered $\{25000, 50000, 100000\}$, selected so as to achieve broad equivalents with the selected number of segment ω values. In both cases, a range of values for r was also considered $\{2, 4, 6\}$; although it should be noted that in [23] it was reported that the value of r was not critical and that any value greater than five makes little difference. The three values for both ω and r thus combined to give nine combinations, hence nine sets of experiments. As suggested in [23], the top two best motifs were retained for each record.

5.1 Evaluation Data Set

The data set used for the evaluation was a set of canine Mitral Valve disease Phonocardiograms (PCGs), encapsulated as WAVE files and (for the purpose of the evaluation) interpreted as time series such that the y-axis comprised *amplitude* values. The PCG data was collected, by staff at the University of Liverpool Small Animal Teaching Hospital, using electronic stethoscope equipment. In some cases, the recordings were done in stereo, in others in mono. In the case of the stereo WAVE files, two point series were extracted, one for each channel. This resulted was a 72 point series dataset. Each point series had a class label associated with it selected from the class attribute set $\{B_1, B_2, C, Control\}$. The first three class attributes are stages of Mitral Valve disease that appear in the collected data as defined by the European College of Veterinary Internal Medicine (ECVIM) [24]. The last class attribute represents PCGs that did not feature any disease (used for control purposes). The average length of a single point series was $355{,}484$ points. Once PCGseg had been applied, it was reduced to $83{,}569$ segments. A reduction in size by a factor of 4.25.

5.2 Runtime Evaluation

The runtime results obtained using the proposed PCGseg approach coupled with the MK algorithm are presented in Fig. 2. The runtimes were considerably shorter than those obtained using the MK algorithm applied to time series without segmentation. With segmentation, the average runtime for each experiment was about 14 hours, without segmentation (results not included here for reasons of space limitation) the average runtime per experiment was some 90 hours (more than six times greater than when using segmented data). From Fig. 2, it can be seen that when using $r = 2$ and $r = 6$ similar runtime patterns are produced, whereas when using $r = 4$, the behaviour is different. For $r = 2$ and $r = 6$, the best runtime was obtained using a window size of $\omega = 50$; whilst for $r = 4$ using a window size of $\omega = 25$ produced the best runtime.

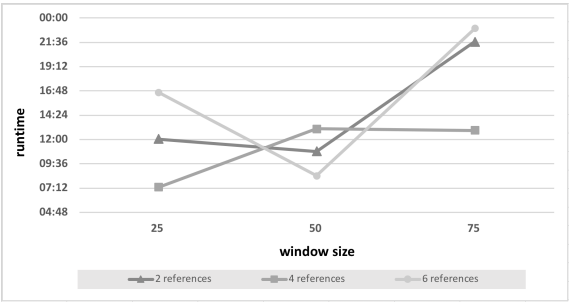

Fig. 2. Runtime plots for Motif Generation using the MK algorithm applied to segmented PCG time series

5.3 Classification Accuracy

For the experiments to compare the classifiers accuracy, two different classification models were used: (i) the well known k Nearest Neighbour (KNN) classification model [25, 26] and (ii) Smallest Average Classification (SAC), a variation of KNN developed by the authors. KNN classification operates by finding the k most similar existing (labelled) records to a previously unseen record to be classified. For the experiments reported here, $k = 1$ was used; thus the class associated with the most similar record was assigned to the record to be classified. KNN classification was chosen because it is frequently used in point series analysis [27, 28]. The SAC model calculates the similarity (average distance) between a new record to be classified and all previously stored records for each class; the new record will then be classified with the class label associated with the most similar class. Similarity was calculated using the approach presented at the end of Sect. 4 above.

Classification accuracy was measured in terms of Accuracy, Precision, Recall and F-Score; metrics all commonly utilised to evaluate classification models [29]. The evaluation was conducted using Ten Cross Validation (TCV) throughout. The average results obtained are presented in Tables 1 and 2, best results associated with each ω value highlighted in bold font. Table 1 gives the results obtained using the proposed PCGseg approach (and both KNN and SAC), whilst Table 2 presents the results obtained using the raw, unsegmented data (and both KNN and SAC). From the tables, it can be seen, firstly, that the performance using segmented and unsegmented data was similar, a recorded best and worst accuracy using segmented data of 70.8% and 54.7%, and without segmentation a best and worst accuracy of 71.9% and 57.6%. In terms of classification model, there was also a very little difference in operation between the two. In terms of the ω parameter, an argument can be made that, when using PCGseg, a value of $\omega = 50$ was the most appropriate. In terms of the experiments using unsegmented data, $\omega = 50000$ produced the best results. The chosen value for r seemed to have little effect, conforming the observation made in [23].

6 Discussion

Looking at the results presented in Tables 1 and 2 in more detail, a great deal of variability can be discerned. This variability of the results was attributed to the fact that the MK algorithm selects motifs in a random manner and this randomness probably does not work well with high data volumes (segmented or otherwise).

Table 1. Classification performance using segmented data and both KNN and SAC classification

ω	r	KNN				SAC			
		Accuracy	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score
25	2	**0.631**	**0.072**	**0.262**	**0.112**	0.596	0.088	0.250	0.118
	4	0.581	0.040	0.250	0.068	0.666	**0.120**	0.212	0.147
	6	0.596	0.047	0.250	0.077	**0.673**	**0.120**	**0.254**	**0.155**
50	2	**0.701**	0.100	0.250	0.142	0.701	0.104	0.250	0.145
	4	0.629	0.097	0.170	0.120	0.604	0.110	0.166	0.131
	6	**0.701**	**0.123**	**0.262**	**0.161**	**0.708**	**0.152**	**0.278**	**0.191**
75	2	**0.701**	0.100	0.250	0.142	0.547	0.065	0.120	0.069
	4	0.673	0.172	0.262	0.206	0.562	0.075	0.183	0.099
	6	0.681	**0.191**	**0.299**	**0.218**	**0.639**	**0.173**	**0.312**	**0.212**

Table 2. Classification performance using unsegmented data and both KNN and SAC classification

ω	r	KNN				SAC			
		Accuracy	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score
25000	2	**0.711**	0.105	0.250	0.146	**0.686**	0.094	**0.241**	0.133
	4	0.694	**0.218**	**0.308**	**0.247**	0.576	0.094	0.162	0.100
	6	**0.711**	0.105	0.250	0.146	**0.686**	**0.096**	**0.241**	**0.135**
50000	2	0.711	0.105	0.250	0.146	0.661	0.132	**0.283**	0.170
	4	**0.719**	**0.132**	**0.262**	**0.170**	0.584	0.090	0.237	0.115
	6	0.711	0.105	0.250	0.146	**0.711**	**0.153**	**0.283**	**0.191**
100000	2	0.694	0.099	0.233	0.137	0.584	0.096	0.212	0.109
	4	**0.711**	**0.105**	**0.250**	**0.146**	**0.652**	**0.168**	**0.283**	**0.198**
	6	**0.711**	**0.105**	**0.250**	**0.146**	0.635	0.107	0.245	0.141

It is conjectured that the limited performance effectiveness, as reported above, was as a consequence of the random manner that candidate motifs were chosen when using the MK algorithm, which may in turn have led to the selection of motifs that were not especially indicative of a class. The MK algorithm chooses the best two similar sub-sequences to be motifs whilst not taking into consideration whether these motifs are in fact good indicators of class or not. In other words, the chosen motifs may not be the best representatives of class. This might be solved by finding the most frequent motifs as these might be a better indicator of class. Furthermore, it is conjectured that smoothing and filtering may contribute to classification effectiveness, as this will serve to reduce runtime and remove noise.

Given the results presented in the previous section, it can be concluded that the classification, using the MK algorithm and segmented and non-segmented time series, was similar (best accuracy values of approximately 70% were obtained). However, using PCGseg significantly less runtime was required; the runtime was improved by a factor of six. This runtime advantage is then the principal benefit offered by the proposed PCGseg approach.

7 Conclusions

In this paper, a novel time-series segmentation method, PCGseg, has been proposed for segmenting the PCG time series data for facilitating motif-based time-series analysis. More specifically, the MK algorithm can be used to identify appropriate motifs in the PCG data prior to the classification process. The objective of the segmentation was to reduce the amount of data, by removing fine details, so as to provide for a more tractable representation while retaining all salient features. The performance of the proposed approach was evaluated by applying it to a realistic multi-class PCG classification problem. The evaluation shows that the proposed approach has a very promising performance gains, as much as six times of speedup compared to the traditional approach, while offering an acceptable level of accuracy, within 70% of the best recorded accuracy. For future work, the authors would like to consider the usage of alternative motif-based time series classification techniques, more specifically techniques where motifs are selected according to frequency of

occurrence. The intuition here is that such motifs will be better class differentiators in the context of PCG classification scenarios.

References

1. Bezruchko, B., Smirnov, D.: Extracting Knowledge From Time Series:An Introduction to Nonlinear Empirical Modeling. Springer-Verlag Berlin Heidelberg (2010)
2. Chiu, B., Keogh, E., Lonardi, S.: Probabilistic discovery of time series motifs. In: Proc. 9th ACM SIGKDD Int. Con. on Knowledge Discovery and Data Mining, ACM (2003) 493–498
3. Gao, Y., Lin, J., Rangwala, H.: Iterative grammar-based framework for discovering variable-length time series motifs. In: Proc. 15th IEEE Int. Conf, on Machine Learning and Applications (ICMLA'17), IEEE (2017) 111–116
4. Serra, J., Arcos, J.: Cparticle swarm optimization for time series motif discovery. Knowledge-Based Systems **92** (2016) 127–137
5. Torkamani, S., Lohweg, V.: Survey on time series motif discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **7**(2) (2017) 1–8
6. Keogh, E., Chu, S., Hart, D., Pazzani, M.: Segmenting time series: A survey and novel approach. In Kandel, A., Bunke, H., Last, M., eds.: Data mining in Time Series Databases. World Scientific (2001) 1–22
7. Huang, G., Zhou, X.: A piecewise linear representation method of hydrological time series based on curve feature. In: Proc. 8th Int. Conf. on Intelligent Human-Machine Systems and Cybernetics (IHMSC'16),. (2016) 203–207
8. Anirudh, R., Turaga, P.: Geometry-based symbolic approximation for fast sequence matching on manifolds. International Journal of Computer Vision **116**(2) (2016) 161–173
9. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra: Dimensionality reduction for fast similarity search in large time series databases. Journal of Knowledge and Information Systems **3**(3) (2001) 263–286
10. Patel, A., Bullmore, E.: A wavelet-based estimator of the degrees of freedom in denoised fmri time series for probabilistic testing of functional connectivity and brain graphs. NeuroImage **142** (2016) 14–26
11. Zhao, H., Dong, Z., Li, T., Wang, X., C., P.: Segmenting time series with connected lines under maximum error bound. Information Sciences **345** (2016) 1–8
12. Zhao, H., Li, G., Zhang, H., Xue, Y.: An improved algorithm for segmenting online time series with error bound guarantee. Int. Jo. of Machine Learning and Cybernetics **7**(3) (2016) 365–374
13. Belevich, I., Joensuu, M., Kumar, D., Vihinen, H., Jokitalo, E.: Microscopy image browser: A platform for segmentation and analysis of multidimensional datasets. PLOS Biology Journal **14**(1) (2016) 1–13
14. Oliveira, J., Sousa, C., Coimbra, M.: Coupled hidden markov model for automatic ecg and pcg segmentation. In: Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP'17). (2017) 1023–1027
15. Quiceno, A., Delgado, E., Vallverd, M., Matijasevic, A., Castellanos-Domnguez, G.: Effective phonocardiogram segmentation using nonlinear dynamic analysis and high-frequency decomposition. In: Proc. Computers in Cardiology, IEEE (2008) 161–164
16. Ahlstrom, C.: NonLinear Phonocardiographic Signal Processing. PhD thesis, Linkoping University, Sweden (2008)
17. Dokur, Z., lmez, T.: Heart sound classification using wavelet transform and incremental self-organizing map. Digital Signal Processing **18**(6) (2008) 951–959
18. Gavrovska, A., Paskas, M., D., D., Reljin, I.: Region-based phonocardiogram event segmentation in spectrogram image. In: Proc. Neural Network Applications in Electrical Engineering (NEUREL'10), IEEE (2010) 69–62
19. Moukadem, A., Dieterlen, A., Hueber, N., C., B.: Comparative study of heart sounds localization. In: Proc. Bioelectronics, Biomedical and Bio-inspired Systems, SPIE Proceedings Vol. 8068 (2011) 9 pages
20. Helton, W.: Canine Ergonomics: The Science of Working Dogs. CRC Press (2009)
21. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing sax: a novel symbolic representation of time series. Data Mining and Knowledge Discovery **15**(2) (2007) 107–144
22. Sklansky, J., Gonzalez, V.: Fast polygonal approximation of digitized curves. Pattern Recognition **12**(5) (2007) 327–331
23. Mueen, A., Keogh, E., Zhu, Q., Cash, S., Westover, B.: Exact discovery of time series motifs. In: Proc. SIAM Int. Conf. on Data Mining. (2009) 473–484
24. Nakamura, K., Kawamoto, S., Osuga, T., Morita, T., Sasaki, N., Morishita, K., Takiguchi, M.: Left atrial strain at different stages of myxomatous mitral valve disease in dogs. Journal of Veterinary Internal Medicine **31**(2) (2017) 316–325
25. Chen, C., Pau, L., Wang, P.: Handbook of Pattern Recognition and Computer Vision. World Scientific (1993)
26. Kuncheva, L.: Combining Pattern Classifiers: Methods and Algorithms. John Wiley and Sons, 2nd ed (2014)
27. Wang, X., Fang, Z., Wang, P., Zhu, R., Wang, W.: A distributed multi-level composite index for knn processing on long time series. In: Proc. Int Conf. Database Systems for Advanced Applications (DASFAA'17), Springer, LNCS 10177 (2017) 215–230
28. Stojanovic, M., Bozic, M., Stankovic, M., Stajic, Z.: A methodology for training set instance selection using mutual information in time series prediction. Neurocomputing **141** (2014) 236–245
29. Witten, I., Frank, E., Hall, M., Pal, C.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2016)