A New Approach for Analyzing Physiological Time Series

Dong Mao, Yang Wang, and Qiang Wu

July 28, 2010

1 Introduction

An understanding of physiological time series such as the heart-beat intervals is important to many areas, like heart-attack prediction, cardiovascular health, sport and exercise, etc. The study of time series can reveal underlying mechanisms of the physiological system, which usually contains both deterministic and stochastic components. Therefore the analysis of time series is very complicated because of the nonlinear and non-stationary characteristics of physiological time series data. Over the past years, time series analysis methods are applied to quantify physiological data for identification and classification (see [7, 12]). The application of physiological time series analysis commonly focus on measuring different aspects of time series data such as complexity, regularity, predictability, dimensionality, randomness, self similarity, etc. The tools used in these techniques include but not restrict to the mean, standard deviation, Fourier transform, wavelet, entropy, fractal

dimension, pattern detection (see [8, 13]).

Recently a new mathematical tool, empirical mode decomposition (EMD), was proposed by Norden Huang et al (see [5, 6]). It decomposes a time series into a finite sum of intrinsic mode functions (IMF) that generally admit well-behaved Hilbert transforms. This decomposition is based on the local characteristic time scale of the data, which makes EMD applicable to analyze nonlinear and non-stationary signals. EMD and Hilbert transform together, called the Hilbert-Huang transform (HHT), usually allow to construct meaningful time-frequency representations of signals using instantaneous frequency of the data. EMD and HHT have been applied with great success in many application areas such as biological and medical sciences, geology, astronomy, engineering, and others (see [5, 1, 3, 6, 11, 10]). Another interesting set of examples is the work of L.Yang, who has successfully applied EMD based techniques for texture analysis and Chinese handwriting recognition (see [16, 17, 15, 18]).

The main purpose of this paper is to develop a new approach for the analysis of physiological times series. Our approach is motivated by two intuitions and coupled with modern machine learning techniques. The first intuition comes from a belief that a physiological system should contain a deterministic part that reflects the basic mechanism for the system to survive and a stochastic part that represents the variability of resilience. Mathematically they can be represented by the low frequency and high frequency components of a physiological signal. This motivates the application of methods of decomposing signals into various components according to frequencies in the quantitative analysis of physiological time series. Examples include

the Fourier transform, wavelets, EMD. In our method we will use an iterative convolution filter which is an alternative of EMD. The second intuitions comes from a statistical perspective of irregularity. A lot of study has proved that normal physiological systems show irregularity due to the existence of stochastic components while the decrease of irregularity usually imply the abnormality. From statistical perspective, irregularity of a data set is represented by the "outliers". This motivates us to study the statistics of outliers in physiological time series. However, we must be careful in dong so. Practical physiological times series usually contains noise which may also appear as outliers. We have to guarantee the "outliers" we examined are not pure noise. This is possible because true outliers do not have informative structures and could be detected. The second intuition is the motivation for our feature construction in section 2.2.

These two intuitions enable us to decompose the physiological times series and construct features for our quantitative analysis. Combining with the well established feature selection techniques in machine learning we can remove the redundancy of the features and find relevant statistics for classification of physiological time series. SVM-RFE (Support Vector MachineRecursive Feature Elimination) is suggested in this paper for linear classification problems. The details of our approach will be described in Section 2.

We will use our approach to the study of congestive heart failure problems. The purposes is two-fold: The first is to build good classifier to enable good diagnosis. The second is to find what kind of irregularity is related to the heart health. The results and discussions are summarized in Section 3.

The novelty of our method is mainly the following two points. Firstly,

although we decompose the time series into components of different frequencies, we do not compare them from the frequency domain. Secondly, we proved that the outliers in a physiological time series are usually not true outliers but are informative instead.

2 Method

2.1 Signal decomposition

Let L be a low pass filter. Denote by T the weak limit of the the operator $(I-L)^n$ as $n \to \infty$, i.e., for a discrete signal X and time t

$$T(X)(t) = \lim_{n\to\infty} (I-L)^n(X)(t).$$

Using this operator iteratively, a signal X can be decomposed as follows: Let $F_1 = T(X)$ and for $k \geq 2$,

$$F_k = T\left(X - \sum_{i=1}^{k-1} F_i\right).$$

After m steps we get $F_1, \ldots, F_m$ which we call mode functions and the residual

$$R = X - \sum_{i=1}^{m} F_i.$$

Then we have

$$X = F_1 + F_2 + \ldots + F_m + R.$$

In this decomposition, roughly speaking the former mode functions are noise or high frequency components and the latter mode functions are low frequency components and R is the trend.

This procedure follows the spirit of the traditional EMD introduced in [5]. In the traditional EMD, the low pass filter L is chosen as the average of the upper envelope (the cubic spline connecting the local maxima) and the lower envelope (the cubic spline connecting the local minima). This method, although has been successfully used in many applications, is lack of theoretical foundation and has its limitations.

In [9] a new approach is proposed. In this new approach the low pass filter is a moving average generated by a mask $\mathbf{a} = (a_j)_{j=-N}^{N}$ that gives the $L(X)$ as the convolution of a and X, i.e.,

$$L(X)(t) = \sum_{j=-N}^{N} a_j X(j+t).$$

With this choice of L we call the operator T an iterative convolution filter. A rigorous mathematical foundation and convergence analysis is in [9, 14]. Note the mask $\mathbf{a}$ is finitely supported on $[-N, N]$ and N is called the window size. The flexibility to choose the window size is crucial in applications and forms a main advantage of this method.

Similar to decompositions by many other methods like Fourier transform and wavelets , the trend and low frequency components are usually assumed to characterize the profile of the signal and the high frequency components characterize the details. In different applications we need the features of difference components.

2.2 Feature extraction

After decomposing the signal into the mode functions and the trend, we need to extract statistics that can characterize the essential features of these

components. This step requires a priori knowledge of the problem under consideration. It could be rather weak. But without any priori knowledge, it is difficult to get proper statistics. Also, this step is strongly problem dependent. In the following let us use the heart-beat intervals as an example to illustrate how to construct the features.

For each mode function F_i, we first get its mean m_i and standard deviation σ_i. By the previous studies [2] the healthy heart beats more irregularly than the unhealthy heart. This motivates us to design the statistics to measure the irregularity. To this end, we consider the terms that are larger than $m + \sigma$ and find their mean and standard deviation. We also find the mean and standard deviation of the terms that are larger than $m + 2\sigma$. Symmetrically we also get the mean and standard deviation of those terms that are smaller than $m - \sigma$ and $m - 2\sigma$. Note all these terms are in some sense "outliers" and it is natural to use the statistics of the outliers as the characterization of the irregularity.

Next we consider the local maxima and local minima of F_i. These two series measure the local upper amplitude. For each series we consider the ten statistics as those for F_i.

Therefore for each component we get 30 statistics.

Unlike in [2], we use the whole 24-hour heart beat time series and assume we do not know the periods for different activities such as sleeping and walking. We think the statistics for different periods should be different and not all of them represent the difference between the healthy and unhealthy people. This motivates the idea of split the whole time series into subseries. Suppose we have K subseries for each patient. Then we get K subcompo-

nents for each mode function which will be denoted by $F_{ij},\ j = 1, \ldots, K$. For each subcomponent F_{ij} we also get the 30 statistics as for F_i. For the same statistics, we have K values from the K subcomponents. We compute the mean of all values, the mean of lower half and upper half, respectively. This gives 90 statistics as summary. So for each component we get 120 statistics.

For physiological signals, we believe the trend and low frequency components are determined by the fundamental mechanism while the individual differences should be reflected by the high frequency components. In case that we do not have much knowledge about the disease to be diagnosed we may assume the features may also comes from the trend. So the same 120 statistics are also computed for the trend component.

2.3 Feature subset selection

After the above two steps we have get many features for the data. Usually only a small part of them are related to the diagnosis and the physiological mechanism of the disease. The task of the third step is to find the relevant ones. This will be realized by eliminating the irrelevant ones step by step.

Firstly, if a statistic is almost constant, then it is useless in the diagnosis and should be eliminated. For example, the means of the mode functions m_i are all approximately zero and should be eliminated.

Next we use the SVM-RFE method [4] to rank the features. In this method, given a set of training samples, we first train linear SVM to get a classifier and then rank the features according to the weights. Because of large feature size and small training samples, the classifier might not be as good. Also, the high correlation between the features may result the

relevant features to have small weights. These reasons could lead the rank to be inaccurate. In order to refine the rank we eliminate the least important feature and repeat the process to re-rank the remained features. Running this process iteratively we finally get the refined rank of the features.

With this rank of features we can conclude which statistics are useful for the diagnosis and characterize the essence of the underlying physiological mechanism. Good classifiers can then be built to make accurate diagnosis.

3 Experiments and Results

In this section we apply our new method described in Section 2 to the hear beat interval times series and report our results and conclusions.

3.1 The data set

The data set includes the heart beat interval time series of 72 healthy people and 43 CHF patients. For each people the heart beat interval is measured for 24 hours under various activities. In our experiment we will assume the activity period is not known. The average ages of these two groups are both 55 years. The standard deviation of age of CHF patients is 11 years and which of healthy people is 16 years. If divide CHF patients into 4 degrees where the degree I is a slight CHF and the degree IV is a severe CHF, most CHF patients are of the degree III.

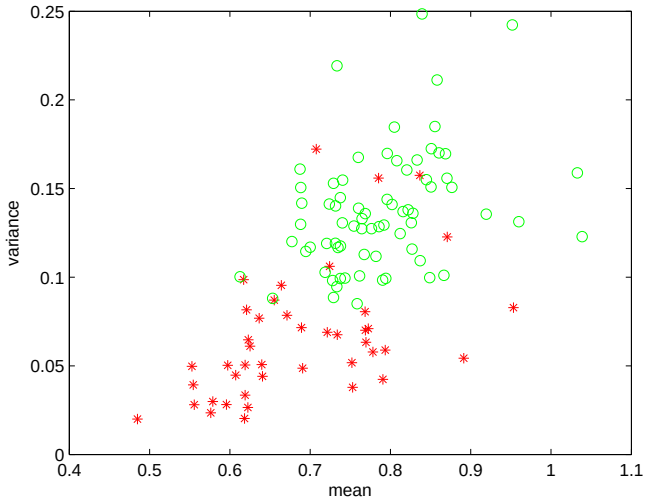

Figure 1: The mean and variance (in second) of the times series, 'o' for healthy people and '*' for CHF patients.

3.2 A primary study

Before using our new method, we study the classification ability of two simple statistics: mean and variance. In Figure 1 we plot the mean and variance of the heart beat intervals for the healthy people and CHF patients. We see that the healthy people and the CHF patients can be roughly separated. The average heart beat interval of healthy people is larger and so is the variance. It shows the heart of healthy people beats slower and more irregularly. This observation coincides with the previous study.

At the same time, we notice that several cases falling into the healthy people show to be severe CHF patients. So we conjecture that the mean and variance might not reflect the essence of the underlying mechanism, although

they have good separability.

3.3 Experiment: feature extraction

For each time series, we use the iterative convolution filter to realize the signal decomposition. In this step we need to specify the window size of the mask. It turns out it should be chosen between 50 and 100 to be stable. In our experiment it is chosen to be 50.

We then calculate the statistics proposed in Subsection 2.2. Here we need to specify the parameter K, the number of subseries. If a statistic really captures the essence of the data set, it should be stable and independent of the choice of K once it is chosen within a reasonable interval. Our experiments show that $K = 50$ is a good choice. Most heart beat signals were recorded for a little bit more than 24 hours. Thus when $K = 50$, each subseries is around 30 minutes of record.

Previous studies have shown that healthy heart beats irregularly. In statistics, irregularity could be measured by statistics of "outliers" that are not due to noise. This motivates us to consider the upper half mean and the lower half mean of the fluctuations. At the same time, from the study in Section 3.2 we find that a healthy heart beats slower than an unhealthy heart in average. These two intuitions enlighten us to conjecture that those larger heart beat intervals (i.e. slower heart beats) in the times series characterize the difference between the healthy people and CHF patients. To confirm this, we do a correlation analysis.

For the first two IMFs of the 50 components of each time series, we calculate and sort the mean and standard deviation of those terms larger

than mean plus standard deviation and those terms larger than mean plus two times standard deviation. For each statistic we compute its correlation to the CHF disease. The result is plotted in Figure 2 in red color. We compute the same indices for those items smaller than the mean minus one and two times standard deviation. The result is plotted in Figure 2 in blue color. From the comparison we see that, in average, correlations of the statistics associated to the larger fluctuations are larger and the upper half mean of these statistics are stable. This observation motives us to disregard the smaller fluctuations and the statistics for those.

3.4 Feature ranking and subset selection

To rank the features, we randomly split the data set into two subsets as the training set and the test set, respectively. In the training set we have 50 healthy subjects and 30 CHF subjects and in the test set there are 22 healthy and 13 CHF subjects. We use the training set to build the SVM classifier and use the test set to control the accuracy. Using the SVM-RFE methods described in Subsection 2.3 we rank the features. To guarantee the stability of the rank we repeat this procedure 1000 times and choose the statistics that appear most frequently in the model.

In all 1000 repeats, the classification error on the test data set is summarized in the following table:

number of errors	0	1	2	3	4	5
number of repeats	823	116	42	14	4	1

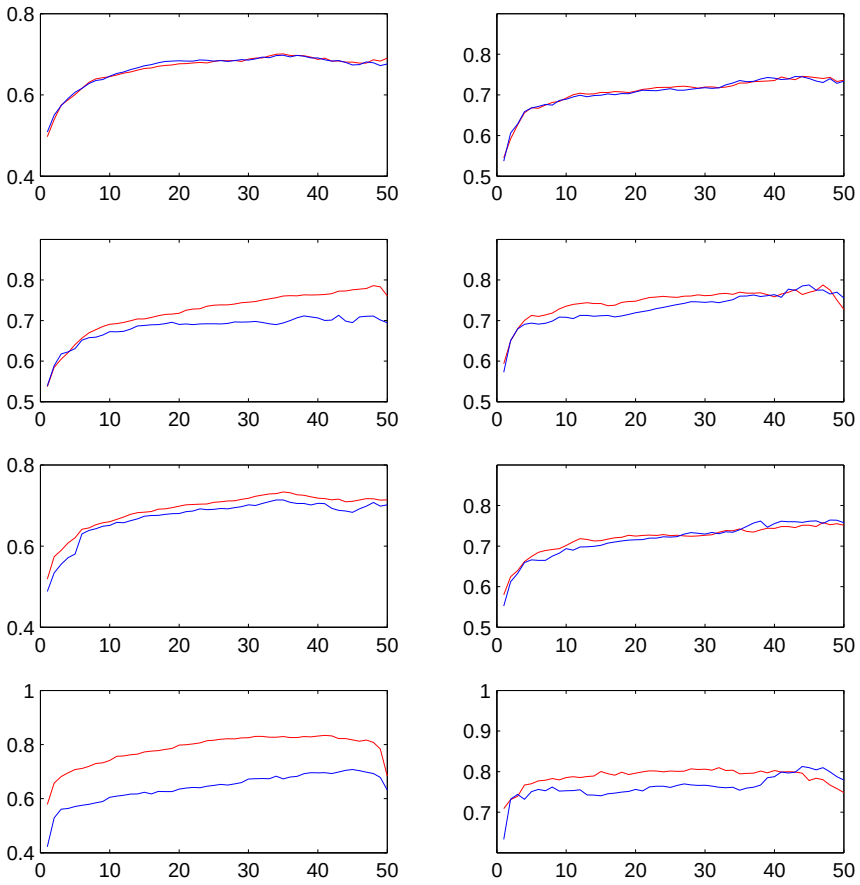

Figure 2: The correlations of various statistics to the CHF disease. The first column is for the first IMF and the second column is for the second IMF. The first line is for the mean of those items larger than the mean plus standard deviation (red line) and those items smaller than the mean minus the standard deviation (blue line). The second line is for the standard deviation of two types items. The third line is for the mean of those items larger than the mean plus 2 times standard deviation (red line) and those items smaller than the mean minus 2 times standard deviation (blue line). The forth line is for the standard deviation of two types of items.

We list the top 10 statistics selected by this procedure:

1. IMF 1: For the subseries consisting of local maxima, find all terms which are greater than the mean plus two times standard deviation, then compute the standard deviation.

2. IMF 1: For the subseries consisting of local maxima, find all terms which are less than the mean minus two times standard deviation, then compute the standard deviation.

3. IMF 1: Equally divide the series into K subseries, for each subseries find all terms which are less than the mean minus two times standard deviation, compute the standard deviation, then take the mean of these K standard deviations.

4. IMF 1: Equally divide the series into K subseries, find local maxima of each subseries, find all terms of local maxima which are greater than the mean plus two times standard deviation, compute the standard deviation, then take the mean of these K standard deviations.

5. IMF 1: Equally divide the series into K subseries, find local minima of each subseries, find all terms of local minima which are greater than the mean plus two times standard deviation, compute the standard deviation, then take the mean of these K standard deviations .

6. IMF 2: Find all terms which are greater than the mean plus two times standard deviation, then compute the standard deviation.

7. IMF 2: Equally divide the series into K subseries, for each subseries find all terms which are greater than the mean plus two times standard deviation, compute the standard deviation, then take the mean of these K standard deviations.

8. IMF 2: Equally divide the series into K subseries, find local maxima of each subseries, find all terms of local maxima which are greater than the mean plus two times standard deviation, compute the standard deviation, then take the mean of these K standard deviations.
9. IMF 2: Equally divide the series into K subseries, find local minima of each subseries, find all terms of local minima which are less than the mean minus two times standard deviation, compute the standard deviation, then take the mean of these K standard deviations.
10. Trend: Equally divide the series into K subseries, find local maxima of each subseries, find all terms of local minima which are greater than the mean plus standard deviation, compute the standard deviation, then take the mean of these K standard deviations.

These 10 statistics that appear most frequently in the model all measure the irregularity of the local amplitude. Take Statistics 1 and Statistics 7 as the example. They are obtained as the following. To get Statistics 1, for the first IMF F_1, find the local maxima u and compute the mean m and the standard deviation σ of u. Then we choose terms greater than $m + 2\sigma$ and find their standard deviation. To get Statistics 7, for the subseries of the second IMF $F_{2j}, j = 1, \ldots, K$, compute the mean m_{2j} and the standard deviation σ_{2j} of F_{2j}. Then we choose terms greater than $m_{2j} + 2\sigma_{2j}$ of F_{2j} and find their standard deviations. Then we compute the mean of K such standard deviations. In the following figure we show the distribution of the healthy people and CHF patients using these two statistics. From this figure it is easy to see that healthy people and CHF patients are well separated.

Observing these two statistics, we find that both of them measure the

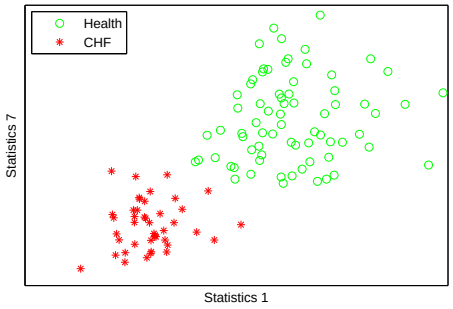

Figure 3: CHF * vs Healthy ∘. The x-axis is Statistics 1 and the y-axis is Statistics 7.

ability of the heart beat to become extremely slower than usual. Our result shows that the strong adaptability of extremely slower heart beat might be the irregularity that characterizes the healthy hearts.

3.4.1 Reliability of the top features

We have found that the most relevant features are statistics for the "outliers", i.e., those items larger than mean plus two times standard deviations, or items less than mean minus two times standard deviations for IMFs. A natural question arises: "Is this accidental?" This is equivalent to ask whether the outliers taken into account are noise or informative.

In order to answer this question we further analyze these outliers. Firstly we notice that the up and down fluctuations are not balanced for both healthy people and CHF patients. The percentage of items larger than mean plus two times standard deviation for healthy people is 2.84% and those items smaller than the mean minus stand deviation is only 2.35%. For CHF patients the

percentages are 2.49% and 2.17%, respectively. This observation is the first evidence that outliers are not due to noise because otherwise they should be balanced distributed. Moreover, recall for normal distribution the percentage of one-side outliers outside the two times standard deviation is 2.28%. We see the outliers for CHF is closer to it due to noise while those for healthy people are much more and probably due to not only noise and hence are informative.

To further confirm our conclusion, we do the following test: we calculate the statistics for the items larger than the mean plus v times standard deviation with the variable v changes from 0 to 2 and investigate their correlation to the CHF disease. Here we consider three quartile of the 50 standard deviations of these items in the 50 components. The correlation is plotted in Figure 4. From this analysis, we see the correlation increases with v. Such a trend appears also in other statistics. This clear trend implies that the relevancy between these statistics and the CHF disease is not accidental. Instead, we should consider the outliers informative and their properties characterize the essence difference between healthy people and CHF patients.

4 Conclusions and discussions

In this paper we developed a new approach for the analysis of the physiological times series. The motivation comes from that the physiological times series usually contains both deterministic and stochastic parts and they can be represented by the low and high frequency components of the times series. Our new method uses an iterative filter to realize the decomposition

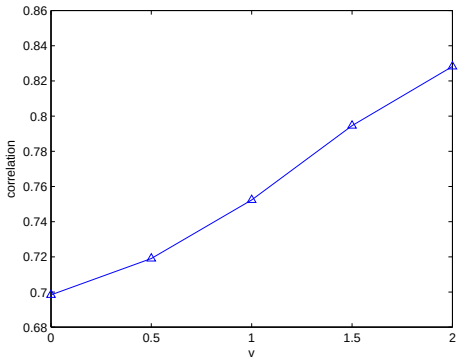

Figure 4: Corrections of the statistics described in Section 3.4.1 with v varying from 0 to 2.

of the times series into high and low frequency components and study their statistics. SVM-RFE is then used to select highly relevant features.

Our method is applied to analyze the heart beat interval time series for CHF disease. The top features are found to measure the ability of heart to beat extremely slowly. Healthy heart show strong ability which we conjecture are due to the strong resilience to the environment and human activities.

References

[1] Q. Chen, N. Huang, S. Riemenschneider, and Y. Xu. A B-spline approach for empirical mode decompositions. *Adv. Comput. Math.*, 24(1-4):171–195, 2006.

[2] M. Costa, A. L. Goldberger, and C.-K. Peng. Multiscalce entropy analysis of biological signals. *Physical Review, E*, 71:021906, 2005.

[3] J. Echeverria, J. Crowe, M. Woolfson, and B. Hayes-Gill. Application of empirical mode decomposition to heart rate variability analysis. *Medical and Biological Engineering and Computing*, 39:471–479, 2001.

[4] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[5] N. Huang, S. Shen, Z.and Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and L. H.H. Theempirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London, A*, 454:903–995, 1998.

[6] N. E. Huang, Z. Shen, and S. R. Long. A new view of nonlinear water waves: the Hilbert spectrum. In *Annual review of fluid mechanics, Vol. 31*, pages 417–457. Annual Reviews, Palo Alto, CA, 1999.

[7] H. Kantz, J. Kurths, and G. Mayer-Kress. *Nonlinear techniques in physiological time series analysis*. Springer series in synergetics. Springer, Heidelberg, 1998.

[8] H. Kantz and T. Schreiber. *Nonlinear time series analysis*, volume 7 of *Cambridge Nonlinear Science Series*. Cambridge University Press, Cambridge, 1997.

[9] L. Lin, Y. Wang, , and H. Zhou. A new approach to empirical mode decomposition. preprint, 2008.

[10] B. Liu, S. Riemenschneider, and X. Y. Gearbox fault diagnosis using emperical mode decomposition and hilbert spectrum. preprint.

[11] D. Pines and L. Salvino. Health monitoring of one dimensional structures using empirical mode decomposition and the hilbert-huang transform. In *Proceedings of SPIE*, volume 4701, pages 127 – 143, 2002.

[12] T. Schreiber. Interdisciplinary application of nonlinear time series methods. *Phys. Rep.*, 308(2), 1999.

[13] H. Tong. *Nonlinear Time Series Analysis*. Oxford University Press, Oxford, 1990.

[14] Y. Wang and Z. Zhou. Toeplitz operators in $\ell^\infty(F)$ and their applications to empirical mode decompositions. preprint, 2008.

[15] L. Yang, Z.and Yang and Y. Tang. Illumination-rotation-invariant feature extraction for texture classification based on hilbert-huang transform. preprint.

[16] Z. Yang, L. Yang, D. Qi, , and C. Suen. An emd-based recognition method for chinese fonts and styles. *Pattern Recognition Letter*, 27:1692–1701, 2006.

[17] Z. Yang, L. Yang, and D. Qi. Detection of spindles in sleep eegs using a novel algorithm based on the hilbert-huang transform. In T. Qian, M. I. Vai, and Y. Xu, editors, *Wavelet Analysis and Applications*, Applied and Numerical Harmonic Analysis, page 543C559. Birkhauser, 2006.

[18] T. Zheng, L. Xie, and L. Yang. Integrated extraction on handwritten numeral strings in form document. *Pattern Recognition and Artificial Intelligence*, 2009. in press.