Sperm Whale Signal Analysis: Comparison using the AutoRegressive model and the Wavelets Transform

Olivier Adam, Maciej Lopatka, Christophe Laplanche and Jean-François Motsch

***Abstract*—**This article presents the results using a parametric approach and a Wavelet Transform in analysing signals emitting from the sperm whale. The extraction of intrinsic characteristics of these unique signals emitted by marine mammals is still at present a difficult exercise for various reasons: firstly, it concerns non-stationary signals, and secondly, these signals are obstructed by interfering background noise.

In this article, we compare the advantages and disadvantages of both methods: AutoRegressive models and Wavelet Transform. These approaches serve as an alternative to the commonly used estimators which are based on the Fourier Transform for which the hypotheses necessary for its application are in certain cases, not sufficiently proven.

These modern approaches provide effective results particularly for the periodic tracking of the signal's characteristics and notably when the signal-to-noise ratio negatively effects signal tracking.

Our objectives are twofold. Our first goal is to identify the animal through its acoustic signature. This includes recognition of the marine mammal species and ultimately of the individual animal (within the species). The second is much more ambitious and directly involves the intervention of cetologists to study the sounds emitted by marine mammals in an effort to characterize their behaviour.

We are working on an approach based on the recordings of marine mammal signals and the findings from this data result from the Wavelet Transform. This article will explore the reasons for using this approach. In addition, thanks to the use of new processors, these algorithms once heavy in calculation time can be integrated in a real-time system.

***Keywords*—**autoregressive model, Daubechies Wavelet, Fourier Transform, marine mammals, signal processing, spectrogram, sperm whale, Wavelet Transform.

I. Introduction

MARINE mammals emit very unique sounds. These sounds distinguish them from other species but also enable individuals of the same species to be identified. The definition of an acoustic signature is key to identifying an animal but also in the endeavor of behavioral analysis. Our work is compatible with the cetologist's research geared towards tracking an animal (or a group of marine mammals), and aims to correlate their signals to specific life sequences: hunting, social behaviour, mating…

Real-life conditions make the recording of marine mammals difficult [1][2]. The noise-to-signal ratio is often unfavourable.

At present, scientific analysis of sounds emitted by marine mammals follows a classical approach based on the Fourier Transform [3][4][5][6][7]. In order to achieve a time-frequency representation, the spectrogram is of widespread use. It is relatively easy to interpret the obtained results by observing the evolution of frequencies during successive time windows. In addition, the spectrogram is easily applied and currently obtained through fast calculation: in our studies, we use the split radix method to reduce the calculation time. Also, the spectrogram is systematically used in most analyses of marine mammal signals without much attention to strict mathematical hypothesis necessary for its use. The Fourier transform is not optimal when the signal-to-noise ratio is deficient or when signals are extremely brief or staccato. Similarly, the Fourier transform is inadequate for non-stationary and non-linear signals. It is important to be critical, in interpreting the obtained results. While this estimating device can appear perfectly suitable in cases where the sounds contain principal characteristic frequencies (in vocalisation of killer-whales, for example), it can be less adapted to dealing with transitory or even impulsive sounds, as in the case of sperm whales [8][9][10]. Comparisons can be drawn with speech signals in human beings, between the voiced parts (which provide specific frequency peaks known as harmonics formants) and the unvoiced parts (for example plosive or fricative sounds).

In using the Fourier Transform the first step is to find a compromise between time and frequency resolution with the drawback that in favouring one, precision in the other is lost. As for harmonic signals, for which we know, a priori, the range of frequencies, the spectrogram is sufficient for arriving at a first estimation of the frequencies evolution. The mathematic formula is

Manuscript received Jan 17, 2005. This work was supported in part by the "Association Dirac" (France).

O. Adam is assistant professor. He is with the Laboratoire d'Informatique Industrielle et d'Automatique (LiiA), University Paris XII, Créteil, FRANCE (corresponding author to provide phone: (33)-1-45171774; fax: (22)-1-45171752; e-mail: adam@univ-paris12.fr).

M. Lopatka and C. Laplanche are PhD students with the LiiA, University Paris XII, Créteil, FRANCE (e-mail: maciej.lopatka@xl.wp.pl – laplanche@univ-paris12.fr).

JF. Motsch is Professor and the director of iSNS in the LiiA, University Paris XII, Créteil, France (e-mail: motsch@univ-paris12.fr).

 COPYRIGHT © 2005 ENFORMATIKA

$$S_x(nT, f) = \left| \int_{nT}^{(n+1)T} x(t)\, g(t - nT) \exp(j 2\pi f t)\, dt \right|^2 . \quad (1)$$

Where $x(t)$ is the signal, $g(t)$ the time window, T and f the coordinates in the time-frequency representation ($n \in Z$). This representation has some drawbacks:

1) It is irreversible.
2) It often results in an over-sampling of the original signal in order to detect rapid signal fluctuations.
3) It does not allow for a time localization in every frequency.
4) Data regarding inter-spectral phases is lost.

The above formula (eq. 1) provides a representation of time-frequency. However, the time window length must be determined from the outset. Thus, the approach is less than ideal if we consider that the whole of frequencies varies with time and so it is under and overestimated at random. This is a major disadvantage when dealing with sperm whales as these particular signals are brief and so both rich in frequency and time domains (see figure 1). These sounds are produced through a very unique mechanism: they are pneumatic in origin and the spermaceti figure importantly [11][12][13][14].

The spectrogram of the sperm whales signal (figure 1) is shown in figure 2.

We note the different segments of the click in both figure 1 and 2. In these 2 illustrations, we can distinguish 4 segments; the first being a sequence of impulses of the strongest amplitude; followed by a segment of residue; the third and fourth segments repeating the first two but with lesser amplitude.

To arrive at this spectrogram, we had to choose the two following parameters for the time window: length and shape. The length corresponds to the number of click samplings. This number is actually a compromise between time and frequency precision. We put forth an a priori hypothesis regarding the scope of the frequency range. Consequently, we can propose the most appropriate size for the time window reserved for this particular frequency range. The disadvantage is that through this fixing of the time window size, the precision on the higher frequencies deteriorates.

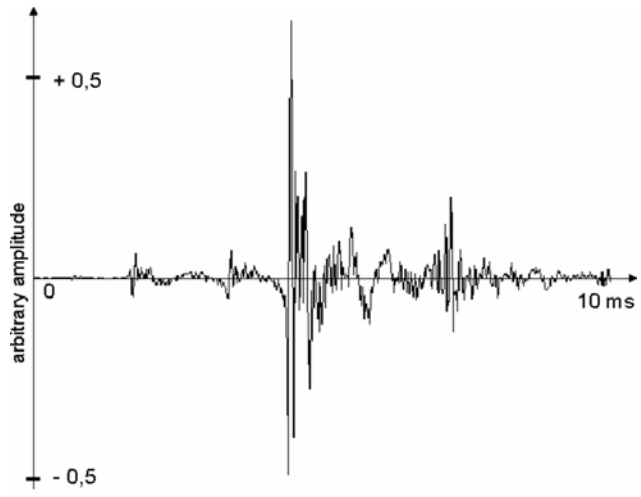

Fig. 1. Sperm whale Click. The above signal illustrates the 4 segments including the first impulse and final residue.

The spectrogram (figure 2) is calculated based on 64 samplings which constitute the shortest time window that can be chosen in order to consider a stationary signal for the length of a millisecond. This choice can be debated when referring to the first (and the third) segment of the click. Other frequency estimators have been defined in signal processing theory to avoid the disadvantages in using the spectrogram [15][16]. We chose to employ the Wavelet Transform [17][18][19]. We will justify this choice in the following section. Then we will illustrate the acoustic signatures obtained through the Wavelet Transform. Before concluding, we will show the results of its performance when adding the noise factor.

II. METHOD

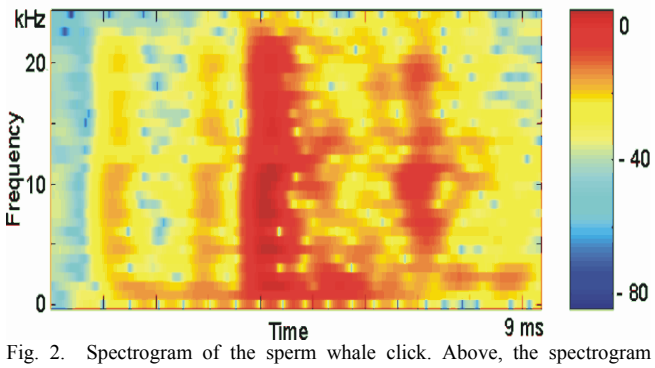

Fig. 2. Spectrogram of the sperm whale click. Above, the spectrogram results from the click recorded and shown in figure 1. In order to calculate this spectrogram we favored time precision (as opposed to frequency precision) by choosing a large time window.

To avoid inaccuracies in time-frequency localisation and compensate for the drawbacks of the spectrogram, different mathematic approaches are defined based on a projection of the signal on various vectors each having different time lengths. The objective of this projection is that it be adaptable to the frequencies we plan to research.

We have chosen the AutoRegressive parametric models (AR models) for the following reasons: firstly, this approach provides a representation of pertinent desired information in a set of coefficients which can then be used either directly in signal analysis or as input in an expert pattern recognition system. Secondly, this model is resistant to noise and can, by tracking coefficients, distinguish the presence of a marine mammal.

We have also chosen the Wavelet Transform. This transform is used with the objective of having a precise time-frequency representation and the resistance to noise.

Following is a presentation of both methods.

A. AutoRegressive Model

The parametric model is an approach used to provide a

 COPYRIGHT © 2005 ENFORMATIKA

representation of time signals [20][21][22]. Essentially, the model is a linear combination of previous signal samples or of a noise. Whether dealing with speech processing, theory of automatic domain, or a prediction of time series, a complete set of models has been defined. The basic formula is the following [20][22]:

$$Y(z) = \frac{z^{-n_l} B(z)}{A(z)} U(z) + \frac{C(z)}{A(z)} E(z). \quad (2)$$

Where $E(z)$ represents the Adjusted Mean (AM), generally it is a white noise that is used as input in the parametric model; where $Y(z)$ represents the AutoRegressive part (AR), past samples of the time signal are concerned; where $U(z)$ represents the eXogenous part (X): it is an external input, for example, the tendency of a time series or the need for comparison with another series is liable to influence the current model.

Figure 3 provides a graphic illustration of this basic parametric model.

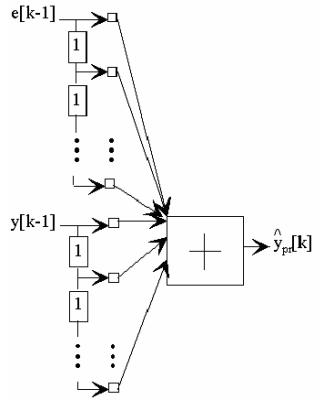

Fig. 3. Design of the ARMA model.

In order to deduce the pertinent signal information in a set of coefficients, it is best to use algorithms which cause the coefficients to converge to where optimal values are attained, according to chosen criteria of error. It is the least square average of error that is commonly used. Many algorithms can be employed; Yule-Walker, Levinson, or adaptive algorithms.

This parametric model is a generalisation of all models used in the identification of a time series. We can deduce which model to use, AR model, MA model or ARX model depending on which of the three inputs described above is more pertinent to the analysis.

We have numerous criteria at our disposal for a subsequent evaluation of these results: converging of the squared error, normalised averaged error, white test of error. In addition we can use Akaïke criteria also known as Final Prediction Error [23]. It is a criterion which provides a Performance/Complexity ratio:

$$FPE = \frac{N + p + 1}{N - p - 1} e^2. \quad (3)$$

Where N, p are respectively the number of signal samples studied and the number of model parameters (otherwise known as its order). e is the error between signal samples and the values calculated by this model. We note that as the order increases, the criteria increases. Therefore, for a given error, a model having many parameters is considered to be less optimal than a weaker order model.

To summarize, there are 3 steps to carrying out this approach most effectively (see figure 4): selection of model, selection of algorithm, and validation.

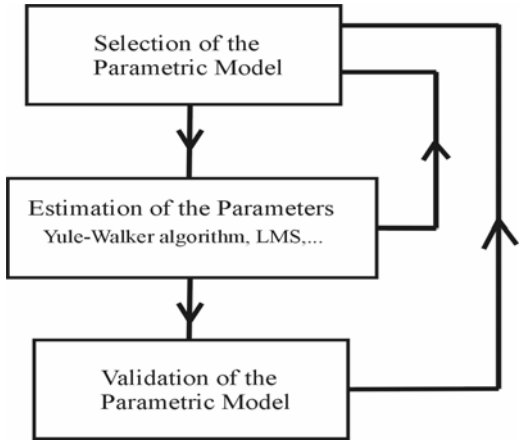

Fig. 4. Parametric approach. After having selected both model and parametric algorithm, the performance has to be evaluated. This method facilitates the selection of the best suited model for the planned application.

This approach has many advantages, particularly as it corresponds to the original idea that all existing physical processes can be modeled (ad least in approximation) through a more or less complicated linear system. In many applications, we have to go further by choosing non-linear models only in instances where the results obtained through parametric models are unsatisfactory [20]. We will take this into consideration in further studies at a later date.

What advantages does this approach provide? We side-step the drawbacks of the Fourier estimator. It is no longer necessary to hypothesize as to the frequency of the analysed signal. It is also no longer necessary to resort to time windows of analysis. Finally, through the use of the adaptive algorithm, it is not necessary to make hypotheses in reference to the stationary nature of the signal. It is certainly this latter argument which is the preponderant for our application: the model coefficients can evolve with each introduction of a new sample making it possible, through analysis of the value of this weight, to distinguish a marine mammal's zones of presence or absence.

Finally, if for the novice, the evolution of coefficients does not provide any results, or the results are too difficult to interpret, one can arrive at the spectral representation using the following formula

 COPYRIGHT © 2005 ENFORMATIKA

$$P_{AR}(f) = \frac{\sigma^2}{|A(f)|^2}, \quad (4)$$

with $A(f) = 1 + \sum_k a_k e^{-j2\pi kf}$. (5)

We note that it is easy to go from time representation to frequency representation (eq. 4). In addition, the model gives a spectrum having finer resolution than that of the Fourier Transform. We also note that modelling the time signal is a question of modelling its spectrum.

Also, as with the spectrogram, one can provide a time-frequency representation. We show the evolution of the spectrum we calculated using each new coefficient value (obtained with each new signal sample, for example). But one can also provide a time-model representation with a direct visualisation of the time evolutions of a set of coefficients. We apply these two modes of representation in our work.

We have chosen to develop this method for use on signals emitted by marine mammals and in particular, by sperm whales, for two principle reasons: firstly, to attribute one model to each individual cetacean and secondly, because it is more resistant to noise than the spectrogram.

Our method provides a satisfactory approximation by presenting a linear model of the process, with a period of calculation which allows one to easily imagine its real-time application.

B. Wavelet Transform

Morlet, for example, introduced time windows of various lengths which are inversely proportionate to the desired frequency. Through his approach, Morlet maintained a time precision with resolution independent of the frequency and did so even for non-stationary signals. The Wavelet Transform [24] is based on the same approach: the result is represented in a time frequency graph of varying resolution. This method provides frequency resolution (through low frequencies analysis) and time resolution (through high frequencies analysis).

As is evident in figure 5, the spectrogram gives a uniform time frequency resolution. The Wavelet Transform resolution

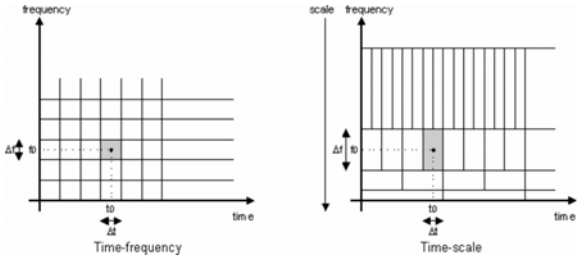

Fig. 5. The time-frequency graph

is contingent on the frequency.

The wavelets analyse finite time signals whose average is zero. Their particular shape is suited to discontinuous signals or signals with quick impulses. In fact, the results stemming from the Wavelet Transform constitute a multi-scale approach (rather than time-frequency graph, figure 5).

The notion of frequency is replaced by that of *scale*: to consider high and low frequencies, the wavelet $\Psi(t)$ is either contracted or dilated (figure 6). We note that a given wavelet width corresponds to a fixed resolution also called *scale*. Consequently, this wavelet is projected onto the entire length of the signal for analysis.

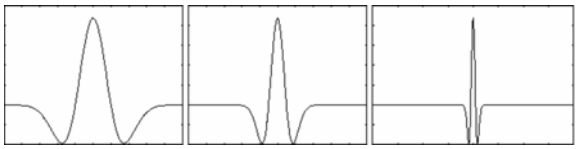

Fig. 6. Dilatation of wavelet. Once a wavelet has been selected (here, the Mexican hat) it is either contracted or dilated in order to consider the presence of all frequencies susceptible to being contained in the signal (highs and lows).

The family of wavelets is created by:

$$\Psi_{mn}(t) = a_0^{-m/2}\Psi\left(a_0^{-m}t - nb_0\right) \quad (6)$$

where $m, n \in Z$; $a_0 > 1$; $b_0 \in R^+$.

$\Psi(t)$ is called mother wavelet. Another expression is

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right), \quad (7)$$

where $\Psi_{a,b}(t)$ is the wavelet obtained from dilatation by using term $a > 0$ and by shifting by using term $b \in R$ of the mother wavelet $\Psi(t)$. We therefore have two parameters that characterize the wavelet: the scale (or dilation) *a*, associated with frequency, and the shifting term b, associated with the temporal (time) position. The greater a, the more dilated the wavelet. As a result, the greater values of a will be associated to low frequencies and the lesser values to higher frequencies.

As with all transforms, the obtained coefficients can serve to reconstruct the signal. In our case, they quantify the similarity of the wavelet to the analysed signal. In determining these coefficients we calculate a Continuous Wavelet Transform (CWT). These coefficients allow us to better visualize the signal's content exactly as the representation of the spectre did based on the coefficients of the Fourier Transform.

As regards the CWT, we must calculate the coefficients $C_{a,b}$, as they relate to the signal analysis $x(t)$ over the domain *D* by the wavelet $\Psi_{a,b}(t)$:

$$C_{a,b} = \left\langle x, \Psi_{a,b} \right\rangle_D = \int_D x(t)\overline{\Psi}_{a,b}(t)dt. \quad (8)$$

Even if these coefficients lack physical meaning, their absolute value will be higher than the similarity between $x(t)$ and $\Psi_{a,b}(t)$ will be important.

 COPYRIGHT © 2005 ENFORMATIKA

To calculate a signal's Wavelet Transform, therefore, it is necessary to calculate a series of scalar products to a specific scale at any given time (or the signal's projection onto the wavelet). The steps to follow are

1) select a mother wavelet $\Psi(t)$;
2) select a range of scale factors (values of a) ;
3) choose a range for time shifts (values of b) ;
4) create a wavelet $\Psi_{a,b}(t)$ for each pair (a,b) and calculate the scaled product to obtain a coefficient $C_{a,b}$.

By sampling the two parameters a and b of the wavelet, the principal is to separate the signal into two components, one representing its general form (also called *approximation*) and the other representing its *details*. This is to say that the general shape of a function is represented by its low frequencies, whereas detail is represented by its high frequencies. It is the same as using 2 filters simultaneously (low and high pass filters).

A wavelet $\Psi(t)$ and a scale function of $\Psi_{a,b}(t)$ are associated with each pair of filters. With each transform, we move from a signal length N to two signals of length $N/2$. This is referred to as passing to an inferior resolution. By repeating this method, the whole range of resolutions can be accessed. The minimal decomposition leaves only one value for both the *approximation* and the *details*. In practice, it is in analysing the results obtained through various resolutions that we fix the value of the level of decomposition.

III. RESULTS AND COMMENTS

Following is a presentation of results obtained by using AutoRegressive Models and the Wavelet Transform.

A. AutoRegressive Model

In this section, we will present the results obtained when applying this method, beginning with the research of the model order, coefficients evolution and spectrogram comparisons and finally noise resistance.

After a first analysis of the spectrum of sperm whale clicks, we opted to use the AutoRegressive model. We used the FPE Akaïke criterion (referred to above) to obtain the optimal model order (see figure 7).

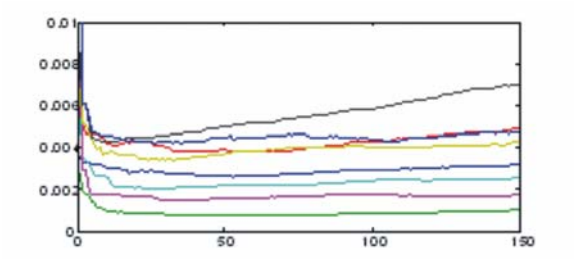

Fig. 7. Application of Akaïke criteria. The above value depicts the axis corresponding to the amplitude of Akaïke criteria as it pertains to the order of the AR model, of 8 clicks.

Figure 7 shows Akaïke criteria of less than 5.10^{-3} as soon as the order of AR models is higher than 25. With an order between 25 and 50, the criteria stagnate. It is in this range that we choose our model.

To avoid an unjustifiable increase of time of calculation, we set the order of AR models to 32.

We modelled the click (shown in figure 1): figure 8 shows the evolution of coefficients (on both vertical and horizontal axes) and figure 9 shows the evolution of the spectrum calculated from the coefficients.

Figure 9 is an AR spectrogram illustrating the evolution of spectrums with each modification of the model coefficients. In any case, this mode of representation is more easily interpreted by the cetologist. He can identify the frequency evolutions. This mode of presentation is extracted from figure 8 and it engenders increased calculation time. So, in keeping with the objective of building an expert system for detecting automatically the presence of marine mammals, the analysis of coefficients is sufficient and we can bypass this mode of presentation.

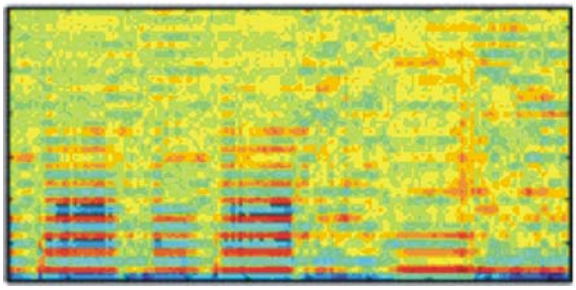

Fig. 8. Evolution of the coefficients. Time is represented on the horizontal axis. The 32 coefficients are on the vertical axis. Their amplitude is projected onto a multicolored scale.

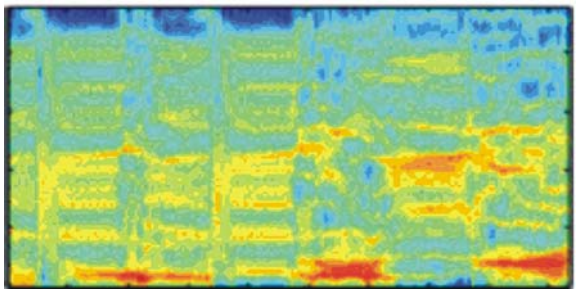

Fig. 9. Evolution of the spectrum. Time is represented on the horizontal axis. The spectrums calculated from the 32 coefficients are represented on the vertical axis. The amplitude is projected onto a multicolored scale.

We have focused on the resistance factor in our approach for different signal-to-noise ratios (performance levels are kept stable despite the added noise factor). To reiterate, the results obtained by the Fourier spectrogram are not always easily interpreted when the signal-to-noise ratio is very weak. For the purpose of our study, we added white noise to the reference clicks. While this approach is purely theoretical, the

 COPYRIGHT © 2005 ENFORMATIKA

results obtained under unfavourable conditions can serve as a reference point. In practice, we consistently use a filter before recording the signals. Figure 10 and 11 show the results obtained with signal-to-noise ratios equal to 0 and -5 dB.

The influence of added noise is apparent in the preceding two figures. We have established a threshold to keep only the

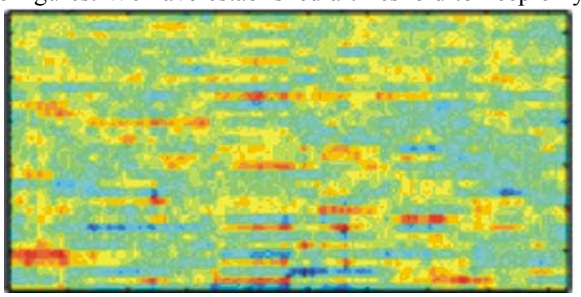

Fig. 10. Scattering of coefficients. White noise has been added. The signal-to-noise ratio is 0 dB.

important coefficients because we believe it necessary to resort to an expert system to treat these signals under unfavourable conditions. This will be the subject of an upcoming publication by our team of researchers.

B. *Wavelet Transform*

For the purpose of our study, we applied the Daubechies

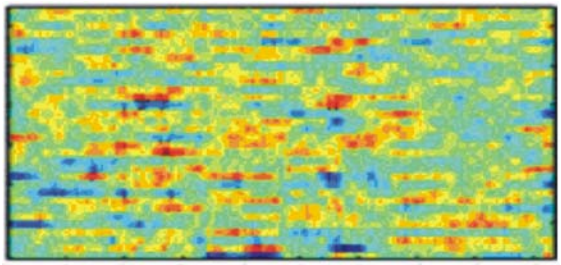

Fig. 11. Scattering of coefficients. White noise has been added. The signal-t-noise ration is -5dB.

wavelet (order 15) [25]. We chose this wavelet for its great similarity to the shape of the referenced click [19]. We carried out the analysis based on a scale of 1 to 64 (at intervals of 1). The number 1 mark indicates half of the sampling frequency (24 kHz). One can easily distinguish the shape of the time-scale representation of the clicks.

We applied this technique to the click presented in figure 1. Figure 12 depicts the multi-scale representation.

1) Analysis of the shape 'fanning out': In figure 12, we note a fan shape symbolizing the acoustic signature of the emissions from the sperm whale. This shape is the characteristic result of the Daubechies wavelet applied on the sperm whale click. This shape is entirely different from what we can discern through a killer whale's shrill whistling, for example. Thus we have laid out two distinct objectives. The first is to distinguish the species, or indicate when a sperm whale signal has been detected. This can be done using continuous monitoring mechanisms in a given marine zone, for instance, based on global characteristics that the Wavelet Transform provides. A range corresponding to this opening or fanning out which represents the click needs to be established. We have embarked upon research that would address this. At present, the study has only been applied to a limited number of recorded sperm whales and thus needs to be further explored before it can be validated. Secondly, it is necessary to identify a single sperm whale. We could compare our method of recognition to those of digital impressions: i.e. the superimposing of new findings on those provided in a data base. We chose 10 identical points serving as references to verify the identity of an animal.

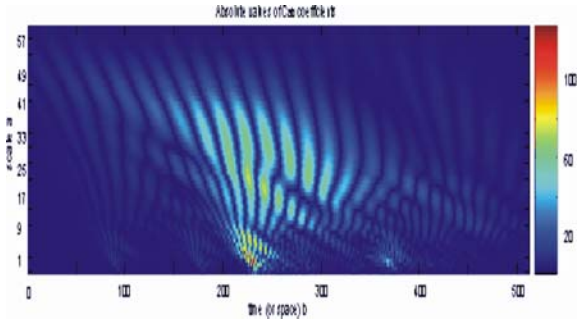

Fig. 12. Wavelet Transform of the sperm whale click. The Wavelet Transform of the click provides a characteristic representation which serves as acoustic signature of the emitted sounds of the sperm whale

2) Results with addition of noise: We are interested in the further testing of the Wavelet Transform when we add noise. To this end, we have purposely degraded the signal-to-noise ratio. Note that the spectrogram which has as a base the Fourier estimator is considerably noise sensitive and performance decreases rapidly when the signal-to-noise ratios are unfavorable. In such an instance, the time-frequency representation obtained is essentially non-exploitable.

Figure 13 is of particular interest while it illustrates that even when the signal-to-noise ratio is entirely unfavorable, as with SNR = -5 dB, we come back to the acoustic signature of the sound emitted by the sperm whale. Certainly, there are many parasites which can interfere with pertinent information. However, it is entirely feasible that a certain number of particular terms be preserved (number of 'continuous and discontinuous lines' of the fanning out, for instance) which lead us to believe that recognition via the 'digital impression' warrants further analysis. Our studies have shown that it is possible to decrease the SNR to -10 dB all the while preserving a desired level of performance. Further results based on more sperm whale studies should be achieved in order to confirm these levels of performance.

 COPYRIGHT © 2005 ENFORMATIKA

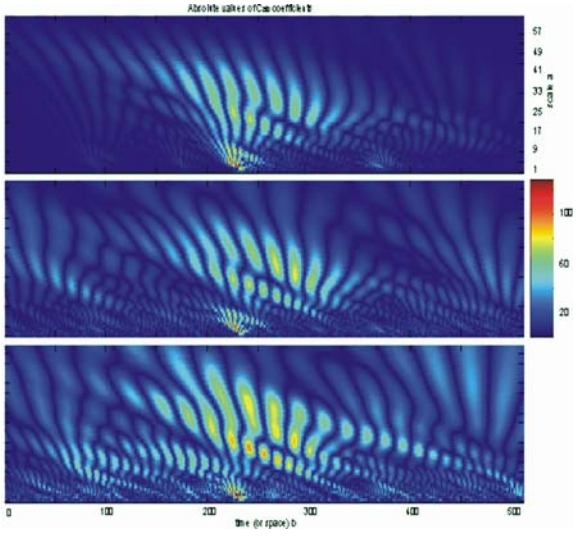

Fig. 13. Resistance to noise. White noise was used. Wavelet Transform results applied to the preceding clicks in the following cases (top to bottom): SNR = 5 dB, SNR = 0 dB, SNR = -5 dB.

IV. Conclusion

With this article, we presented two modern approaches which would avoid the drawbacks of Fourier estimators particularly when the signal-to-noise ratio has deteriorated, as is often the case with marine mammal signals recorded at sea.

Modern technology now allows for other types of real-time representations. We have seen that the parametric model is of interest not only because the algorithms that assure convergence of model coefficients are easily implemented but also because its use is not based on restrictive hypotheses of the Fourier estimators.

We must also mention the potential for calculating frequency representation based on the coefficients, which makes this method easily interpretable even by the novice.

In addition, we have demonstrated other possible uses of coefficients. We can promptly provide a time-evolved representation of these coefficients, as is the practice with the spectrogram.

From the graphic illustration, it is possible to discern the coefficients and to locate an acoustic signature. We can equally see the non-stationary nature for the entire duration of the recording and through this approach classify the ambient noise in a marine zone.

We were also satisfied with the resistance factor of our method when the signal-to-noise ratio changed.

This article also illustrates why the use of the Wavelet Transform for treating sounds emitted from sperm whales is of interest. These emitted sounds are distinctive, notably in their impulsive and brief shape.

Essentially, the Wavelet Transform could have been applied to the signals at several levels. It can serve as (1) a filter which eliminates noise in recorded data, (2) a compression mechanism to optimise the storing of data, and (3) a projection on the basis vectors for extracting relevant information intrinsic to the sperm whale's signals. It is in this latter case that we applied this transform: our aim being to provide a characteristic representation based on the multi-scale projection, from which we could extract sufficient data to confirm the presence of the sperm whale, the discerned shape of which appears as a fanning out.

What is of particular interest is that these terms are to be used in an approach similar to the analysis of digital impressions.

In addition, this article presents the performance of the Wavelet Transform when we add white noise. We have seen that even for entirely unfavorable signal-to-noise ratios, it is nonetheless possible to detect an acoustic signature. This is encouraging for its use under real conditions, keeping in mind that at sea, noise is always present.

It is possible to envision, at present, a real-time application of this approach using the Wavelet Transform thanks to efficient algorithms and ever progressive processors.

This work will need to be carried out through applications on a greater number of sperm whales to prove thoroughly valid. At present, we can say that the approach could be implemented in a complete system of sperm whale recognition and that it is encouraging for three reasons: it allows for a specific form of an acoustic signature, it is noise resistant and its employment does not impede a real-time application.

This expert system could eventually be based on an artificial neural network [26][27][28]. This second objective is to be realized through extraction of intrinsic characteristics of the sounds emitted by each individual marine mammal followed by the treatment of these findings. A database containing the acoustic signature supplements existing information on those mammals we wish to observe which are living in or passing through the marine zone (transient or resident animals).

Acknowledgment

Signals of sperm whales are taken from the data base of M. André, UICMM (*Unidad de Investigación para la Conservación de los Mamiferos Marinos*), University Las Palmas – Gran Canaria.

References

[1] R. J. Urick, *Principles of Underwater Sound*, 3rd ed., Ed. New York: McGraw-Hill, 1983.

[2] R. F. Coates, *Underwater Acoustic Systems.* Ed. London: McMillan, 1990.

[3] W. W. L. Au, R. W. Floyd, R. H. Penner, A. E. Murchison, "Measurements of echolocation signals of the Atlantic bottlenose dolphin," *J. Acoust. Soc. Am.*, vol. 56, pp. 1289–90, 1974.

[4] D. K. Mellinger, Ch W. Clark, "A method for filtering bioacoustic transients by spectrogram image convolution," *Oceans*, 1993.

[5] L. A. Miller, J. Pristed, B. Mohl and A. Surlykke, "The click-sounds of narwhales in Inglefield Bay," *Marine Mammal Sci.*, vol. 11, pp. 491–502, 1995.

 COPYRIGHT © 2005 ENFORMATIKA

[6] I. Tokuda, T. Riede, J. Neubauer and MJ. Owren, "Nonlinear analysis of irregular animal vocalizations," *J. Acoust. Soc. Am.*, vol. 111, pp. 2908–19, 2002.
[7] G. S. Campbell, R. C. Gisiner, D. A. Helweg and L. L. Milette, "Acoustic identification of female steller sea lions," *J. Acoust. Soc. Am.*, vol 111, pp. 2920-28, 2002.
[8] R. Backus, W. E. Schevill, "Physeter clics," *Whales, Porpoises and Dolphins*, Ed. Univ. Calif. by K. S. Norris, pp. 510–28, 1966.
[9] W. Watkins, W. Schevill, "Sperm whale codas," *J. Acoust. Soc. Am.,* vol. 62, 1977.
[10] B. Mohl, M. Wahlberg, P. T. Madsen, "Sperm Whale Clicks: directionality and source level revisited," *J. Acoust. Soc. Am.*, pp. 638–48, 2000.
[11] R. Boileau, "Whale soundings," Zoological Physics, 438, 2002
[12] J. C. Goold, S. Jones, "Time and frequency domain characteristic of sperm whale clicks," *J. Acoust. Soc. Am.*, pp. 1279-91, 1996.
[13] J. C. Goold, J. D. Bennell, S. E. Jones, "Sound velocity measurements in spermaceti oil under the combined influences of temperature and pressure," *Deep-Sea Res*., vol. 43, pp. 1279-1291, 1996.
[14] K. S. Norris, G. W. Harvey, "A theory for the function of the spermaceti organ of the sperm whale," *Physeter Catodon*, Ed. Washington DC: Nasa special publications, vol 262, pp. 397-417, 1972.
[15] W. Lauterborn, U. Parlitz, "Methods of chaos physics and their applications to acoustic," *J. Acoust. Soc. Am.*, vol 84, pp. 1975-93, 1988.
[16] H. Kantz, T. Schreiber, *Nonlinear Times Series Analysis*, Ed. Cambridge PU, 1997.
[17] C. Tiemann, M. Porter, L. Frazer, "Automated Model-Based Localization of Marine Mammals near Hawaï," in Proc. of Oceans 2001 Conference, Hawaï, 2001
[18] J. Ward, K. Fitspatrick, N. DiMarzio, "New algorithms for open ocean marine mammal monitoring," in Proc. of Oceans Conference 2000.
[19] M. Lopatka, "Reconnaissance de signatures acoustiques pour la distinction d'individus dans un groupe de cachalots,", iSnS report, University Paris 12, France, 2002.
[20] S. Haykin, *Signal Processing*, Ed. IEEE Press, 1994.
[21] J. Max, *Méthodes et techniques de traitement du signal et applications aux mesures physiques*, Paris: Ed. Masson, 1982.
[22] ID. Landau, *Identification et commandes des systèmes,* Paris: Ed. Hermes, 1988.
[23] H. Akaïke, "A new look at the statistical model identification," *IEEE Trans. on Auto. Control*, vol.6, pp. 716-23, 1974.
[24] Y. Meyer, *Les ondelettes, Algorithmes et Applications*, Paris: Ed. Armand Colin, 1992.
[25] I. Daubechies, *Ten Lectures on Wavelets*, Philadelphia: Ed. Society for Industrial and Applied Mathematics, 1992
[26] J. Herault, C. Jutten, *Réseaux neuronaux et traitement du signal*, Paris: Ed. Hermes, 1994
[27] D.E. Rumelhart, G.E. Hinton, R.J. Williams, "Learning representations by back-propagation errors," *Nature*, vol. 323, pp. 533-36, 1986.
[28] O. Adam, "Approche compare des techniques connexionnistes et adaptatives pour le traitement des signaux lidar," Thesis, University Paris VI, 1995.

Olivier Adam received the electronic engineering diploma and the PhD degree in signal processing from the University Paris 6, Paris, France in 1991 and 1995, respectiviley. He is Assistant Professor at the University Paris XII since 1995, where he teaches electronics, physics, signal processing and artificial neural network. His research interests include detection, identification and pattern recognition dedicaced to biosignal interpretation.

Maciej Lopatka was born in Wodzislaw Slaski, Poland, on December 25, 1978. He received the M. Sc. Degree in Biomedical Engineering from the University Paris XII, France, in 2002 and the MSEE degree from the Technical University of Wroclaw, Poland, in 2003. He is currently working towards the PhD degree in signal processing of the Signal Theory Section, Institute of Telecommunication and Acoustics, Wroclaw, Poland and of the University Paris XII, France. His research interests are in the general field of signal processing with emphasis on detection and identification of non stationary transitory signals, adaptive and orthogonal filtering.
M. Lopatka is a student member of IEEE.

Christophe Laplanche received the telecommunication engineer diploma from the Ecole Nationale Supérieure des Télécommunications de Bretagne, France, in 2002. He is currently working towards the PhD degree in signal processing of the University Paris XII. His research interests include underwater acoustics, whale sounds signal processing and data analysis.

Jean-François Motsch received the Doctorat d'Etat in 1987 from the University Paris XII, France. He is Professor at the University Paris XII since 1992 where he teaches electronics and signal processing. He is the director of the iSnS research team and the president of the Scientific Committee of the Technology Institute. His research interests signal processing and its application for Biomedical.

 COPYRIGHT © 2005 ENFORMATIKA

Validation of Reverse Engineered Web Application Models

Carlo Bellettini, Alessandro Marchetto, and Andrea Trentini

***Abstract*—**Web applications have become complex and crucial for many firms, especially when combined with areas such as CRM (Customer Relationship Management) and BPR (Business Process Reengineering). The scientific community has focused attention to Web application design, development, analysis, testing, by studying and proposing methodologies and tools.

Static and dynamic techniques may be used to analyze existing Web applications. The use of traditional static source code analysis may be very difficult, for the presence of dynamically generated code, and for the multi-language nature of the Web. Dynamic analysis may be useful, but it has an intrinsic limitation, the low number of program executions used to extract information. Our reverse engineering analysis, used into our WAAT (Web Applications Analysis and Testing) project, applies mutational techniques in order to exploit server side execution engines to accomplish part of the dynamic analysis.

This paper studies the effects of mutation source code analysis applied to Web software to build application models. Mutation-based generated models may contain more information then necessary, so we need a pruning mechanism.

***Keywords*—**Validation, Dynamic Analysis, Mutation Analysis, Reverse Engineering, Web Applications

I. Introduction

WEB applications quality, reliability and functionality are important factors because software glitches could block entire businesses and determine strong embarrassments. These factors have increased the need for methodologies, tools and models to improve Web applications (design, analysis, testing, and so on).

Important factors for Web applications are "speed" (in technology change, content update and fruition), complexity, large dimensions and design/use maturity. Web applications are heterogeneous, distributed, and concurrent: their analysis, understanding, reengineering, and testing are not easy task. Conventional methodologies and tools may not be adequate.

This paper focuses on analysis of legacy Web applications where business logic is embedded into Web pages. Analyzed applications are composed by Web documents (static, active or dynamic) and Web objects. In particular we focus in server side components that dynamically generate Web documents in response to some user inputs and gestures. The structure of these documents is not given *a priori*, but is dynamically constructed based on user interactions.

The analysis of existing Web applications through traditional static source code examination of "highly" dynamic applications is a very difficult task. The most complex problem is to define the structure of the Web documents produced by the server-side component. This problem is traditionally known as an undecidable problem, because it is related to the program execution paths analysis, and in particular to determine if a given execution path is feasible.

Performing static analysis may be difficult and not much effective. Dynamic analysis may be used to integrate static analysis to simplify it, to increase analysis effectiveness, and it may be more "Web-adequate", because it let us perform analysis with some degree of language abstraction.

Dynamic analysis can be: profiling, debugging, user-gestures capture and reply, log files analysis, event-trace recording, and so on. Some of the existing dynamic techniques may define a partial application model. Information is extracted through a limited number of program executions. Dynamic analysis may produce an application model without covering all relevant application behaviors. This is an intrinsic limitation of all types of dynamic analysis.

Static analysis is often preferred, because dynamic analysis is context–based and results are driven by execution cases (potentially infinite). To be useful, dynamic analysis should try to be exhaustive in defining application execution paths.

In this paper we present the validation process for UML models generated by our WAAT system via our reverse engineering technique, which is a mix of static and dynamic techniques. Our approach differs from traditional dynamic ones because it is not focused on user gestures replication but on application evolution. The approach is essentially based on two steps: dynamic analysis to define relevant execution paths and static analysis to analyze them. Static analysis is based on a scanner/parser that analyzes source code, while dynamic analysis is based on mutation analysis and simulation of navigation sessions (i.e., interaction with the Web server hosting the application under analysis). These combined operations can better define relevant execution paths. Mutation analysis tries to bound and simplify user interactions, to lower analysis computational complexity, to increase the accuracy of analysis results and analysis methodology portability. Our proposed technique can be integrated with a set of result

Manuscript received March 12, 2005.

Carlo Bellettini, Alessandro Marchetto, and Andrea Trentini are with Information and Comunication Department, University of Milan. Via Comelico 39, 20135 Milan, Italy.

{Carlo.Bellettini, Alessandro.Marchetto, Andrea.Trentini}@unimi.it

 COPYRIGHT © 2005 ENFORMATIKA

validation techniques, such as bad links analysis, error pages detection, pages similarity analysis, and, finally, user validation analysis. Application execution paths are built through mutation techniques combined with random input values, so that no user interaction is needed.

This paper focuses on validation of WAAT generated models. This phase is very important because the use of source code mutation may solve the intrinsic limitations of traditional dynamic analysis, but it may define a super-set of behaviors, that must be pruned. In the WAAT project we propose two alternative ways for models validation. The first one is integrated in the applications testing phase. We present here the second one: a pruning technique based on log files and partial user intervention.

This paper is organized as follows. Section II details Web reverse engineering used techniques and implications. Section III defines proposed validation model algorithm. Section IV defines a simple case study of Web application validation-model analysis. Finally Section V concludes the presented work and describes future work ways.

II. Model Recovery

The WAAT analysis core is composed by: application **behavior analysis** and application **model building**.

Application **behavior analysis** [2] is performed through static and dynamic analysis. Static and dynamic analyses treat static and dynamic application components using source code and on-line interactions with the Web server. For example, for static pages, we use traditional source code analysis based on a language parser. While, for a single server page generating multiple client pages, we apply dynamic analysis to try to determine a meaningful number of client pages (through mutation analysis and application executions). Then, the dynamically generated client–side pages are analyzed (with traditional source code analysis) to build diagrams. More generally, for every dynamic Web document, we use mutation analysis to define mutants (for example changing the control–flow structure of original source code page) to be fed into session navigation simulations, in which every mutant replaces the original source code and the simulation performs generated interactions. This simulation is used to send input values and page requests to the Web server, and saving responses that are analyzed later.

The aim of mutation is to automatically follow relevant execution paths in the Web application, to cover as many navigation paths as possible.

This approach does not need knowledge about the language, only a simple map of mutant operators, deployable with easy to program parsers and with low computational complexity.

Model building; with the information extracted by the previous phase we build an application OO model (such as described in [3], [6], [4]) using UML class and state diagrams. We have defined a UML meta-model usable to describe applications [2]. Class diagrams are used to describe structure and components of a Web application (e.g., forms, frames, Java applets, input fields, cookies, scripts, and so on), while state diagrams are used to represent behavior and navigational structures (client-server pages, navigation links, frames sets, inputs, scripting code flow control, and so on).

III. Model Validation

The "mutation" generated model may contain more information than what is needed. In particular, it may contain "*Not-Valid*" information, such as not valid dynamically-generated client-side pages. A client-page is "*Valid*" if it is reachable in the original application (without mutants) via an execution path. Since mutation may define a model with a super-set of behaviors we need a pruning technique. Our validation technique follows on these steps:

Model analysis: we extract a list of dynamic server-pages and their related client-side dynamically-generated pages

First-reduction: we prune the list by applying content validation analysis tools (such as HTML Tidy by W3C [5])

Page analysis, subdivided in:

- we analyze client-side dynamically generated pages to extract information, such as: inputs (e.g., GET request parameters sending to call pages); structure, which may be all pages source code (with text), or only the source code structural properties related (such as: form tags, script information, links, and so on). These information are used in the next steps to define page similarity, so a minimal set of information may be composed by navigational system information [8]
- we compute unique hash function with structural extracted information
- finally, we associate a tuple of dynamic-specifics (DS) to every dynamic page. DS={<input page>, <inputs parameters>, <output pages>, <output hash related> }

Second-reduction: pruning using **page analysis** phase results [8]

"**Test cases**" generation; we compute a set of "test cases" with sub-steps:

- for every dynamic page we define specifics such as: TcT={<inputs parameters>, <output pages hash>}
- for every page we define: TcS={<input pages>, <output pages>}

Log files analysis; sub-steps are:

- for every dynamic page we extract requests (URL and related parameter inputs and values)
- for every request in log file [1] we repeat navigation, saving Web response, and re-apply page analysis
- we reduce the set of output pages through structural similarity analysis
- then we fill a table of Tlog tuples: {<dynamic page>, <inputs parameters>, <generated pages>, <output hash related>}

Log based validation sub-steps are:

- we map every TcT with Tlog, by URLs matching. A

 COPYRIGHT © 2005 ENFORMATIKA

match between Tct and Tlof validates the generated client-side page. The set of non-matched pages is labeled "Not-Verified"

Visual navigation; every "Not-Verified" TcT must now be validated with user intervention. We mark "Valid" a path, if the page is reachable in the original application (without mutants). To define page-related path, we use the TcS table.

Model update; finally, we update the model by deleting the remaining "Not Valid" pages

IV. Validation Sample

MiniLogin is a simple Web application we used as validation approach sample. This application is composed by PHP and HTML files. We generated UML models and now we apply our validation technique.

Now we compute DS specifics. For example, the dynamic-specifics for "member.php" dynamic page are: DS={<input page>, <inputs parameters>, <output pages>, <output hash related>}, where:

-*inputs pages*: index.html (via form);

-*inputs parameters*: $login & $pwd (that are: username & password input fields in index.html form);

-*output pages*: defines list of currently dynamic page client-side generated pages

-*output related hash key*: defines a list of hashes related to output pages. Hashes are computed on the output of page structure information extraction analysis ("pages analysis" in previously section).

From DS table we compute TcT (Testcase for Test) and TcS (Testcase for Simulation) tables, defining couples of "<inputs parameters>, <output related hash key>" (e.g., TcT row may be: $login,$pwd & hash(u6.htm)) and defining couples of "<inputs pages>, <output pages>" (e.g.,TcS row may be: index.html & u6.htm).

When specifics are computed, we analyze Web server log files to build the Tlog table. For example, for the "member.php" page we take all Web server requests, replaying navigation and saving the response. In the MiniLogin case we used Apache Web server log- files, so we are able to analyze the GET requests (composed by: URL, parameters and parameter values), while other requests (e.g., POST) can not be used through log-validation.

Then we analyze every saved server response to extract structural information and compute hashes on the defined information.

Then we map Tlog and TcT entries to search pages with hash matches. In our case, we find the maps:

TcT(u6.htm)=Tlog(g16.htm);
TcT(u1.htm)=Tlog(g1.htm);
TcT(u2.htm)=Tlog(g5.htm).

Now we may mark {u6, u1, u2}.htm as "Verified", but also other uX.htm pages as "Not-Verified".

Then we may perform a user based validation asking the user (application developer) to mark "Not-Verified" pages as "Valid" or "Not-Valid". For every dynamic page we define a set of paths in the form of (TcS derived) :

<input pages>→<dynamic page>→<output pages>

In our case one of the paths can be the following:

" index.html → member.php ($login,$pwd) → u3.htm"

After path definition, we ask the user to identify valid paths. When every output page is marked, we may update our model by deleting "Not-Valid" pages.

The process complexity may be summarized by the following: two HTTP-requests for every mutation-generated client-side page, (in particular one GET for every output-page defined in DS and Tlog tables); one hash values computed for every received HTTP-response; and the matching search between HTTP-responses from DS and Tlog.

V. Conclusions

We propose a validation process for our technique for reverse engineering of Web applications. This process reduces the super-set of information extracted with mutation analysis.

The combination of mutation analysis and validation process represents a dynamic reverse engineering technique that bounds and simplifies user interactions.

"Bounds" because it reduces the number of pages that must be examined by a user. The super-set of pages automatically generated by the mutation analysis technique is reduced through cross linking with log files.

"Simplifies" because the user does not need any knowledge about the application language, he only need to choose between reachable and not-reachable paths.

We are currently working on a statistical comparison between our technique and other approaches.

References

[1] Apache Web Server – log file, http://httpd.apache.org/docs/logs.html

[2] C. Bellettini, A. Marchetto, and A. Trentini. *WebUml: Reverse Engineering of Web Applications*. 19th ACM Symposium on Applied Computing (SAC 2004), Nicosia, Cyprus. March 2004.

[3] F. Ricca and P. Tonella. *Building a Tool for the Analysis and Testing of Web Applications: Problems and Solutions*. Tools and Algorithms for the Construction and Analysis of Systems (TACAS'200), Genova, Italy. April 2001.

[4] G.A. Di Lucca, A. R. Fasolino, F. Pace, P. Tramontana, and U. De Carlini. *WARE: A Tool for the Reverse Engineering of Web Applications*. 6th European CSMR 2002, Budapest, Hungary. March 2002.

[5] HTML Tidy, http://www.w3.org/People/Raggett/tidy

[6] J. Conallen. *Building Web Applications with UML*. Addison-Wesley, 2000.

[7] M. A. Friedman and J. M. Voas. *Software Assessment: Reliability, Safety, Testability*. John Wiley & Sons, 1995.

[8] T.H. Haveliwala, A. Gionis, D. Klein, and P. Indyk. *Evaluating Strategies for Similarity Search on the Web Similarity search*. WWW 2002

 COPYRIGHT © 2005 ENFORMATIKA

An Approach to Concerns and Aspects Mining for Web Applications

Carlo Bellettini, Alessandro Marchetto, and Andrea Trentini

***Abstract*—**Web applications have become very complex and crucial, especially when combined with areas such as CRM (Customer Relationship Management) and BPR (Business Process Reengineering), the scientific community has focused attention to Web applications design, development, analysis, and testing, by studying and proposing methodologies and tools. This paper proposes an approach to automatic multi-dimensional concern mining for Web Applications, based on concepts analysis, impact analysis, and token-based concern identification. This approach lets the user to analyse and traverse Web software relevant to a particular concern (concept, goal, purpose, etc.) via multi-dimensional separation of concerns, to document, understand and test Web applications. This technique was developed in the context of WAAT (Web Applications Analysis and Testing) project. A semi-automatic tool to support this technique is currently under development.

***Keywords*—** Aspect Mining, Concepts Analysis, Concerns Mining, Multi-Dimensional Separation of Concerns, Impact Analysis.

I. Introduction

WEB applications quality, reliability and functionality are important factors because software glitches could block entire businesses and determine strong embarrassments. These factors have increased the need for methodologies, tools and models to improve Web applications (design, analysis, testing, and so on).

This paper focuses on legacy Web applications where business logic is embedded into Web pages. Analyzed applications are composed by Web documents (static, active or dynamic) and Web objects [6]. This paper describes an approach to help application developers to document, understand and test Web software. Our goal is to describe a Web application Object-Oriented model, and then define a set of application/design slices ("points of view") to analyze and test the application itself, e.g., to generate a set of test cases specific for these points of view. Several Object-Oriented Web modeling methodologies are presented in literature (see Section II). Web OO diagrams (such as Conallen UML [12]) used to describe applications may be very complex, large, and rich of information. Models (above all generated ones) may be difficult to read and comprehend, so that they may not be much usable as core information to document, analyze and test applications. Our approach may be useful to slice or traverse models for software analysis. For example, it may be very interesting to test or reuse single components or tasks or properties, but it may be very complex to spot the relevant details within the whole design documentation. Software concerns are pieces of software that are responsible for a particular task, concept, goal, etc; while "separation of concerns" refers to the ability to identify, encapsulate and manipulate those software parts relevant to a particular concern.

This paper describes a semi-automatic approach to help the user to document, understand and test Web software by slicing applications diagrams. Application model slicing is based on concerns identification and grouping. Our approach describes a set of guidelines to analyze application evolution under different "points of view" (i.e., slices). In particular we would like to define a concern-mining process to help the user to generate application test cases and/or to verify their coverage measure. Our approach is useful to identify multi-dimensional concerns (MDSOC, [20],[36]) in design applications, it uses the MDSOC "dimensions of Hyperspace" concept to describe application slices in Web software. "Hyperspace" is the concept underlying MDSOC, it provides a powerful composition mechanism that facilitates non-invasive software integration and adaptation. In Hyperspaces, concerns are space dimensions. Our concerns mining approach is based on: concepts analysis[1] [17] (as unit-base to identify concerns); impact analysis [2] (to limit software analysis); and token-based concerns identification (to search identified information relationship). This technique is part of the WAAT (Web Application Analysis and Testing) project [6],[5].

This paper is organized as follows. Section II describes related works. Section III describes applications modeling. Section VI introduces our reverse engineering techniques to model recovery. Section V presents some background. Section VI describes our concerns mining approach. Section VII presents a sample. Section VIII Section IX analyses a case study. Finally, Section X presents conclusions.

Manuscript received March 12, 2005.

Carlo Bellettini, Alessandro Marchetto, and Andrea Trentini are with Information and Comunication Department, University of Milan. Via Comelico 39, 20135 Milan, Italy.

{Carlo.Bellettini, Alessandro.Marchetto, Andrea.Trentini}@unimi.it

[1] Concept analysis is "traditionally" used to show all possible software modularizations in a concise *lattice* structure

 COPYRIGHT © 2005 ENFORMATIKA

II. Related Works

Several Web applications **modeling** approach are presented in literature [6]. RMM [21] is a method based on Entity-Relationship diagrams, and is specialized in applications based on databases. WebML [11] enables the description of a Web site under distinct orthogonal dimensions (such as structural, composition, and so on). [28] introduces a Web application simulation model framework that was designed to be compatible with some existing modeling languages. Often, these web methodologies are extensions of traditional methodologies, such as OOHDM [33] for Object Oriented ones. It uses OO models to define: conceptual, navigational and user interface structure of applications. Moreover, some of these are UML based. WARE [14] and Rational Rose Web Modeler [30] are tools for reverse engineering supporting Conallen's extensions [12]. Both tools perform essentially static analyses to generate model. Our WebUml [6] is tool to reverse engineering Web application through static and dynamic analysis. These OO modeling approach derived are related to our concerns mining technique.

Some currently available Web **testing** tools (e.g. [25]) are usually classifiable as syntax validators, HTML/XML validators, link checkers, load/stress testing tools, and regression testing tools, i.e., they are focused on low-level (implementation bound) or non-functional aspects. Some of these tools are often based on capturing user gestures and replay them through testing scripts. These tools cannot provide structural or behavioral test artifacts. Moreover, they represent a good compromise when a formal model is not available and the only implicit model is the user itself.
Other existing tools, such as xUnit (e.g, [19]), propose a different approach, based on unit/functional testing. Other approaches based on functional, structural and behavioral OO model testing, are: [14], [5], [24], [32]. [14] proposes a strategy to build functional unit-integration testing based on WARE described model. [24] proposes an OO Test Model that captures artifacts representing objects, behaviors, and structure aspects. From this model, structural and behavioural test cases may be derived to support the test process. [32] describes tools: ReWeb, performing several traditional source code analysis, to reverse engineering Web applications into UML model; TestWeb, that uses ReWeb models to test applications through Web site validation paths. [5] describes TestUml tool for XML-based test cases generation derived from WebUml extracted model. [22] defines statistical testing based on usage model described from log files and then analysed with Unified Markov Models.

More details about **Aspect Oriented programming** are in [23], while [3] presents the AspectJ famous software. [20],[36] describe the **MDSOC** and HyperJ tool, while [27] studies the relations between quality factors and MDSOC, while [35] the relations between MDSOC and testing. [32] describes our approach to apply Multi-Dimensional Separation of Concerns (MDSOC) theory at Web applications. [31] describes SOC used to reduce the complexity of Web applications. [18] presents an approach to separate Web navigation concerns and application structure. [9] evaluates AOSD code quality influence and presents an approach for reverse engineering aspects, based on concern verification and aspect construction. [10] evaluates the suitability of clone detection as a technique for the identification of crosscutting concerns via manual concern identification. [13] introduces aspect mining and identification in OO. [8],[37] show an approach to aspect mining based on dynamic analysis technique via program traces investigation, to search recurring execution relations. [26] applies three different separation of concerns (SOC) mechanisms (HyperJ, AspectJ, and a lightweight lexically based approach) to separate features in the two software packages. This paper studies effects that various mechanisms have on code-base structure and on restructuring process required while performing separations.

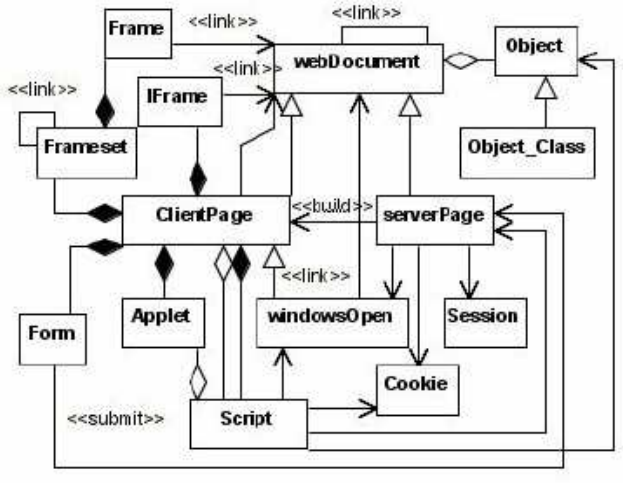

Figure 1: UML Class diagrams Meta-Model

III. Web Applications Modeling

In the WAAT project Web applications are modelled via UML diagrams [6]. The UML model used is based on class and state diagrams. We have defined a UML meta-model (Figure 1: UML Class diagrams Meta-Model), a Web application model is an instance of this meta-model. Class diagrams are used to describe application structure and components (i.e., forms, frames, Java applets, HTML input fields, session elements, cookies, scripts, and embedded objects). State diagrams are used to represent behaviour and navigational structures composed by client-server pages, navigation links, frames sets, form inputs, scripting code flow control, and so on.

The OO application model let us define a mapping between traditional Web application concepts (such as static-dynamic pages, forms, Web objects, and so on) and the MDSOC concepts. This map let us apply separation of concerns methodologies in the Web context, for example to analyse or

 COPYRIGHT © 2005 ENFORMATIKA

test specific assets of existing software. Our approach may be used to "slice" application models by "points of view". This approach is useful with our Web modelling technique, but it may be useful with every other OO modelling techniques applied to Web software (such as presented in previous Section).

IV. MODEL RECOVERY

Our approach ([6],[7]) to model recovery is composed by: **application behavior** analysis, application **model building,** and **model validation**.

Application behavior analysis is performed through static and dynamic analysis. Static and dynamic analyses treat static and dynamic application components using source code and on-line interactions with the Web server. For example, for static pages, we use traditional source code analysis based on a language parser. While, for a single server page generating multiple client pages, we apply dynamic analysis to try to determine a meaningful number of client pages (through mutation analysis and application executions). Then, the dynamically generated client–side pages are analyzed (with traditional source code analysis) to build diagrams. More generally, for every dynamic Web document, we use mutation analysis to define mutants (for example changing the control–flow structure of original source code page) to be fed into session navigation simulations, in which every mutant replaces the original source code and the simulation performs generated interactions. This simulation is used to send input values and page requests to the Web server, and saving responses that are analyzed later.

Mutation analysis is based on mutant operators applied to source code, and in particular to control–flow source code fragments (e.g., "if-then else", "while", etc.), such as logic or Boolean operators, conditions or check operators, and so on. For example, the "=" operator can be mutated into "<>", the ">" operator can be mutated into "≤" or "<", the "AND" operator can be mutated into "OR", etc. The aim of mutation is to automatically follow relevant execution paths in the Web application, to cover as many navigation paths as possible.

This approach does not need knowledge about the language, only a simple map of mutant operators, deployable with easy to program parsers and with low computational complexity.

Model building; with the information extracted by the previous phase we build an application OO model (such as described in the previous section) using UML class and state diagrams.

Model Validation; The "mutation" generated model may contain more information than what is needed. In particular, it may contain "Not-Valid" information, such as not valid dynamically generated client-side pages. A client-page is "Valid" if it is reachable in the original application (without mutants) via an execution path. Since mutation may define a model with a super-set of behaviors we need a pruning technique. Our proposed technique is essentially based on Web server log files analysis validation and "Visual Navigation validation" with the user help.

In particular we analyze the Web server log files (e.g., the Apache Web server log files in Figure 2: Fragment of Web Server Log File) and we replay every Web request (for dynamic pages) to analyse the server response. We match these responses with pages in model (introduced using mutation analysis). The set of matched pages are the "Verified" pages. Every "Verified" page exists in original application. For every "Not-Verified" page we ask the user help to validate it. Via model analysis we may define a set of paths containing the "Not-Verified" pages (every ones for a path). User may mark "Valid" a page in a path, if the page is reachable through that path in the original application (without mutants).

```
127.0.0.1 – [26/May/2004:18:04:02 +0200]
  "GET /website/index.html HTTP/1.1" 200 1560
127.0.0.1 – [26/May/2004:18:05:52 +0200]
  "GET /website/dynamicP.asp?code=056978&name=Alex
  HTTP/1.1" 200 1802
127.0.0.1 – [26/May/2004:18:7:26 +0200]
  "GET /website/clientP2.html HTTP/1.1" 200 1727
127.0.0.1 – [26/May/2004:18:7:59 +0200]
  "GET /website/index.asp HTTP/1.1" 200 700
127.0.0.1 – [26/May/2004:19:02:10 +0200]
  "GET /website/pageResource.html HTTP/1.1" 200 2563
```

Figure 2: Fragment of Web Server Log File

The proposed approach is useful to describe existing Web applications via a UML model built in is semi-automatic way. Model construction is automated via mutation analysis, while model validation is quite user dependent. Traditional ways to analyze existing Web software focus on application source code analysis of control-flow expressions to identify representative inputs values. Inputs values are used to define feasible application behaviors. In this conventional approach user must know the application language and must know the control-flow concepts and condition control analysis. The use of mutation analysis decreases user interactions needed to build application models, because mutation changes the analysis perspective, from source code analysis to application analysis. The model may contain spurious information and must be pruned and validate (via Model Validation approach).

 COPYRIGHT © 2005 ENFORMATIKA

V. BACKGROUND

Our proposed approach to concerns mining is based on MDSOC and Concept Analysis. In this section we briefly introduce these theories.

MDSOC (Multi-Dimensional Separation of Concerns) is an approach to implement separation of concerns (SOC) by IBM. IBM implemented a tool named Hyper/J to support MDSOC in Java software. Separation of concerns refers to the ability to identify, encapsulate, and manipulate software fragments relevant to a particular concern (concept, goal, purpose, etc.). Concerns are the primary motivation for organizing and decomposing software into manageable and comprehensible parts. MDSOC lets the user model applications via multi-dimensional structure (named Hyperspace), instead of other OO-techniques that model applications by only one dimension ("tyranny of the dominant decomposition").

Moreover MDSOC encapsulates many kinds of concerns at the same time, and models overlapped concerns and concerns interaction. MDSOC is very useful for on-demand software re-modularization.

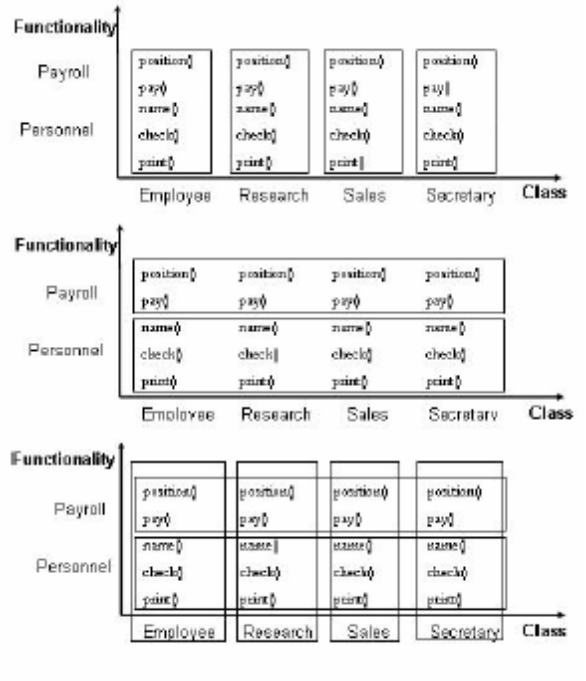

Figure 3: MDSOC hyperspaces sample

Figure 3: MDSOC hyperspaces sample, (see [29]) shows Hyperspace samples for an example personnel software system, these Hyperspaces are composed by two dimensions, axes are software *Class* (e.g., *Employee*, *Research*, *Sales*, *Secretary*) and interesting *Functionality* (*Payroll*, *Personnel*), while points in space are software units, such as class methods (or statements). In case of "dominant tyranny" (OO or Aspect modelling [36]) only one concern type is encapsulated (e.g., first/second plane in Figure 3, where only class or functionality are encapsulated). Instead, MDSOC supports clean separation of multiple, overlapped and interlaced concerns simultaneously, and on-demand re-modularization (e.g., the third plane in Figure 3: MDSOC hyperspaces sample, shows the on-demand re-modularization for system class and functionality).

Formal Concept Analysis (FCA, [15]) is a theory of data analysis which identifies conceptual structures among data sets. Concept Analysis is applied to many fields, such as medicine and psychology, musicology, linguistic databases, library and information science, software re-engineering, civil engineering, ecology, and others.

Concept Analysis important capability is the graphical visualization of these structures among data via the *concept-lattice*. Lattices may be interpreted as classification systems. For example in software engineering, FCA may be useful to show all possible software modularizations in only one concept-lattice or to re-modularize software (e.g., introducing "aspects" in OO existing software).

Concept analysis provides a way to identify grouping of *objects* that have common *attributes.* Given a context=(O, A, R), where: O=objects, A=attributes, R=binary relation between O and A, we may use the concept-analysis grouping algorithm to define concepts. Concepts are "the maximal collection of objects sharing common attributes".

Default Name

A	B	C	D	E	F
Default Name	Four-legged	Hair-covered	intelligent	marine	thumbed
cats	X	X	0	0	0
chimpanzees	0	X	X	0	X
dogs	X	X	0	0	0
dolphins	0	0	X	X	0
humans	0	0	X	0	X
whales	0	0	X	X	0

Figure 4: Context-Matrix sample

For example (see [34]), Figure 4: Context-Matrix sample, shows a generic context matrix with couples of "object-attribute", where objects are living beings, while attributes are five of their possible characteristics. For example:

({cats, dogs}, {four-legged, hair covered}) is a concept.
({cats, chimpanzees}, {hair covered}) is not a concept.

Then, we may define a relationship via hierarchical organization of the defined concepts by describing the relative concept-lattice (in Figure 5: Concept-Lattice sample). Last, by applying the FCA algorithm the concepts table (see Table I) is built.

 COPYRIGHT © 2005 ENFORMATIKA

TABLE I
CONCEPTS SAMPLE

top	({*..all..*}, ø)
C5	({chimpanzees, dolphins, humans, whales}, {intelligent})
C4	({cats, chimpanzees, dogs}, {hair-covered})
C3	({chimpanzees, humans}, {intelligent, thumbed})
C2	({dolphins, whales}, {intelligent, marine})
C1	({chimpanzees}, {hair-covered, intelligent, thumbed})
C0	({cats, dogs}, {hair-covered, four-legged})
bot	(ø, {*..all..* })

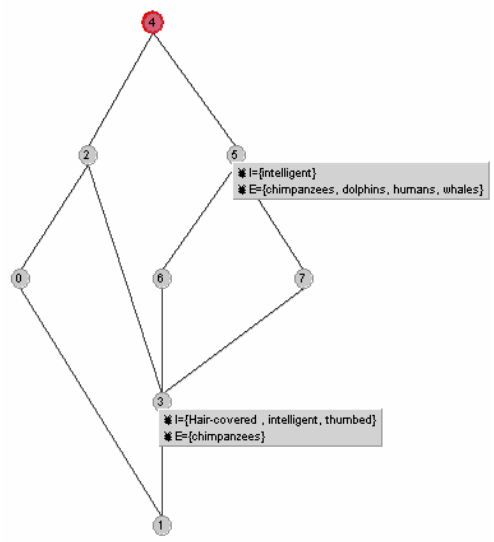

Figure 5: Concept-Lattice sample

VI. CONCERNS DEFINITION ALGORITHM

MDSOC technique is used to build application slices, where every concern (or concerns composition) may be used to define a software code/design slice. MDSOC is realized through Hyperspaces: concerns space organized in multi-dimensional structure. In this structure every dimension is a set of disjoint concerns (i.e., they have no software units in common). We define a semi-automatic concerns mining approach, so concerns identification must be limited to information extracted from applications models or source code. For example, software functionality identification is a semi-automatic task, because the user helps to identify software components. We may lower user interactions by only applying MDSOC to concerns that are automatically extracted. When functionality information cannot be automatically identified, than we use: variables, functions, class, Web documents/objects, links, input-variables, and so on.

Our approach is composed by: **Application Modeling** (**AM**, model definition), **Concerns Elaboration** (**CE**, Hyperspaces definition through model and source code analysis, and Hyperspaces use to reduce models and code taken into account), **Testing** (**T**, the extracted and reduced information may be used to define/refine test cases).

In next Section we present a simple case study useful to describe the proposed approach.

Application Modeling (**AM**) consists in application model definition. We use reverse engineering techniques to define UML diagrams for existing applications. Moreover, diagrams may also be manually refined by the user.

Concerns Elaboration (**CE**) to identify, define, and extract concerns based on application model or source code analysis, subdivide in:

- *Artifacts extraction*: from application model we extract some interesting artifacts such as class, association, variables, methods, links, Web pages (e.g., static, dynamic, dynamically generated), objects (e.g., database, files, reused code), and so on. We use this knowledge to identify concerns (it may be a limitation, i.e., concerns about functionality cannot be completely defined without user know-how).
- *Objects-attributes selection*: from the selected artifacts we define "object-attribute"[2][17] couples to use in concept analysis. We may limit the number of couples by asking user help. Generally speaking, example of couples may be: variables-classes, instance variables-classes, variables-functions, instance variables-functions, and so on. To select couples user may know concerns/aspects theory, and how define a concern/aspect using classes, variables, functions, and so on. To this task we have defined a set of rules, such as, to define aspect in OO software we may analyze the instance variables used by software functions, and if a function uses variables defined in more than one class, this is a candidate to define a crosscutting concerns (aspects).
- *Impact matrix definition*: from the application model we define a matrix I = [class x class]. $\forall\ i_{k,m} \in I = 1$ if $\exists$ class relationship (i.e. association in class diagram between $class_k$ and $class_m$), 0 otherwise. The matrix is then used to decrease analysis computational cost.
- *Context matrix definition*: for every couple defined we build an objects-attributes matrix C = [object x attribute]. $\forall\ c_{k,m} \in C = 1$ if there is a def-use relationship between $object_k$ and $attribute_m$, 0 otherwise.
- *Concept definition/visualization*: we define concepts through the C contexts matrix. We analyze this matrix grouping the maximum number of objects that have common attributes (by concept definition in concept-analysis). To visualize the defined concepts we may use the concept-lattice [17] structure, and in particular we may use existing software such as ToscanaJ [38] or Galicia [16].
- *Concerns identification:* we may identify concerns by the concepts defined in the concept lattice structure. Every node in the lattice is a concept (by concept-analysis definition). Every concept is a concern. Concept is groups of "objects" (in concept analysis sense) that sharing "attributes" (in concept

[2] where "objects-attributes" is defined in concept-analysis theory

 COPYRIGHT © 2005 ENFORMATIKA

analysis sense). For example, for the "object-attribute" couple "variables-classes", concepts are groups of variables that sharing definition or use classes. In this example, code fragments using a common set of variables may represent candidates to identify software behaviors.

• *Concerns composition*: to compose the defined concerns we must analyze the concerns dependencies to define autonomous behaviors or behaviors dependencies. To this aim we propose two ways. With the first approach we may traverse the lattice structure based on the concepts dependencies (associations between nodes in lattice structure). For example, from bottom to top nodes to identify concept-objects dependencies, while from top to bottom nodes to identify concept-attributes dependencies. With the other approach we identify concerns by iteratively grouping previously defined concepts. We define a new "attributes-concepts" matrix[3] A = [attribute x concept]. $\forall$ $a_{k,m} \in A = 1$ if *attribute* is contained in *concept*. By recursively applying the "attributes-concepts" matrix, at each step we build supersets of concepts (grouping concepts that share attributes) that are used as concepts as well in the next step. In this way is possible to group concept/concerns that share information. Every information-sharing between concerns represents a concerns dependencies candidate.

Testing (**T**): to slice applications. The defined concerns may be useful to slice application models or source code. In particular, we would like to use extracted information (application slices) to define test cases and to compute application coverage level for a set of already available test-cases. For example, we may define test cases from a UML model (e.g., from a statechart, see Section III) via traditional OO techniques and then use the reduced diagram to verify test-cases coverage (e.g., uniformly coverage or specialized one). Otherwise we may define test-cases directly from the reduced diagrams, because they represent sets of application features (software fragment with potentially independent behavior).

VII. SAMPLE

"*MiniLogin*" (Figure 6: MiniLogin application Home-Page) is a simple Web application composed by some PHP/HTML files, and its main functionality is to control reserved login-password Web area.

Figure 6: MiniLogin application Home-Page

[3] where *attribute* is from the C matrix, and *concept* was defined in the previous "Concept definition" step

Application Modeling (**AM**): we reverse engineer (through the approach proposed in Section IV) the application UML model, composed by class and state diagrams. Figure 7: MiniLogin UML Class diagram, shows the generated application class diagram (meta-model instance).

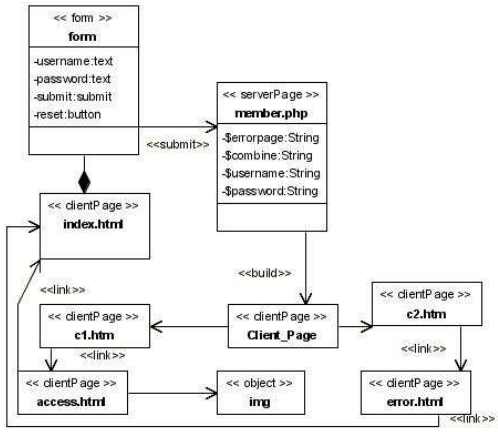

Figure 7: MiniLogin UML Class diagram

Concerns Elaboration (**CE**): defines MiniLogin concerns.

Artifacts extraction: we extract MiniLogin artifacts, lists of: classes, variables, functions, links, and so on.

• *Objects-attributes selection:* we manually select couples of objects attributes to use in concept analysis. E.g., variables-class (named "case-A"), variables-functions, and so on.

• *Impact matrix definition* (due to lack of space we exemplify only a couple of entries): "form" is related to "member.php", while "form" is not related to "C2.html".

• *Context matrix definition* (due to lack of space we exemplify only a couple of entries): for "case-A" "$errorpage" is related to "member.php", "username" is related to "form", and "username" is related to "member.php", while "username" is not related to "C2.htm"

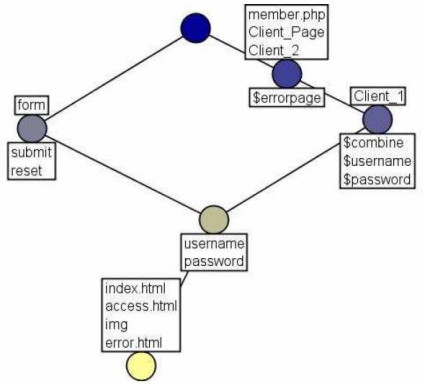

Figure 8: Concept Lattice for MiniLogin "case-A"

• *Concept definition/visualization:* we use the context matrix to define concepts as defined in formal concept-analysis [9]. We may use existing tools (such as ToscanaJ [38], to define and visualize concepts through concept-lattice). Figure

 COPYRIGHT © 2005 ENFORMATIKA

8: Concept Lattice for MiniLogin "case-A", shows the concept lattice for "case-A" concepts.

TABLE II
CASE-A, CONCEPTS

Concept	Object	Attribute
Top	*...all...*	-
C3	{username, password, $errorpage, $combine, $username, $password}	{member.php, Client_Page, c1.htm}
C2	{username, password, $combine, $username, $password }	{member.php, Client_Page, c2.htm, c1.htm }
C1	{username, password, submit, reset}	{form }
C0	{username, password}	{form, member.php, Client_Page, c1.htm, c2.htm }
Bottom	-	*...all...*

• *Concerns identification:* we identify concerns based on the defined concepts. Every lattice node is a concern. Table II shows "case-A" concepts.

TABLE III
CASE-A, "ATTRIBUTES-CONCEPTS" MATRIX

Attribute	C0	C1	C2	C3
index.html				
form	1	1		
member.php	1		1	1
Client_Page	1		1	1
c1.htm	1		1	1
c2.htm	1		1	
img				
access.html				
error.html				

• *Concerns composition:* we compose concerns via concepts grouping. We build the attributes-concepts matrix, with attributes used (rows) and concepts (columns). A cell is = 1 if the attribute is related to the concept (see Table III). Then we group concepts by looking for attributes sharing (in our "case-A", variables). E.g., for "case-A" we group concepts into Z0-to-Z4 groups. Where Z0={C0,C1}; Z1={C0,C2}; and Z2/3/4={C0,C2,C3}.

Now we repeat the attributes-concepts matrix definition, using the same attributes list, but with the newly-grouped concepts (Z0-to-Z4) and then we group these concepts attributes-based defining other new concepts (called ZZ0-to-ZZ4). Then we stop because these concepts are completely overlapped. Finally, we may define the set of composed concerns, where every Cx, Zx and ZZx is a good candidate (usable for our testing task). To reduce the number of candidates we delete overlapped concerns (see Table IV). Every defined concern represents a clearly defined software behavior candidate. We use these concerns to describe the Hyperspace slicing our application, and define the reduced diagrams.

TABLE IV
CASE-A, CONCERNS ENCAPSULATION, ALL ITERATIONS

Concept	Object	Attribute
C3	{username, password, $errorpage, $combine, $username, $password}	{member.php, Client_Page, c1.htm}
C2	{username, password, $combine, $username, $password }	{member.php, Client_Page, c2.htm, c1.htm }
C1	{username, password, submit, reset}	{form }
C0	{username, password}	{form, member.php, Client_Page, c1.htm, c2.htm }
Z2	{username, password, $errorpage, $combine, $username, $password}	{form, member.php, Client_Page, c1.htm, c2.htm}
Z0	{username, password, submit, reset}	{form, member.php, Client_Page, c1.htm, c2.htm }
ZZ0	{username, password, $errorpage, $combine, $username, $password, submit, reset}	{form, member.php, Client_Page, c1.htm, c2.htm }

Testing (**T**): from the reduced diagrams we may automatically define test cases or we may use these diagrams to verify coverage measures of already available test cases (such as in the user metrics driven test cases definition process [5]).

VIII. CASE STUDY

Figure 9: Home-Page *MiniWP* application, shows the case study application selected for experiment. This application is mini Web portal that functions as news reader, images viewer, and Web reserved area control. MiniWP is composed by twenty PHP/HTML files and few TXT "database" files. Figure 10: *MiniWP* Class Diagram, shows the MiniWP class diagram recoved by our reverse engineering approach.

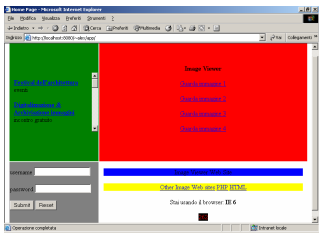

Figure 9: Home-Page *MiniWP* application

 COPYRIGHT © 2005 ENFORMATIKA

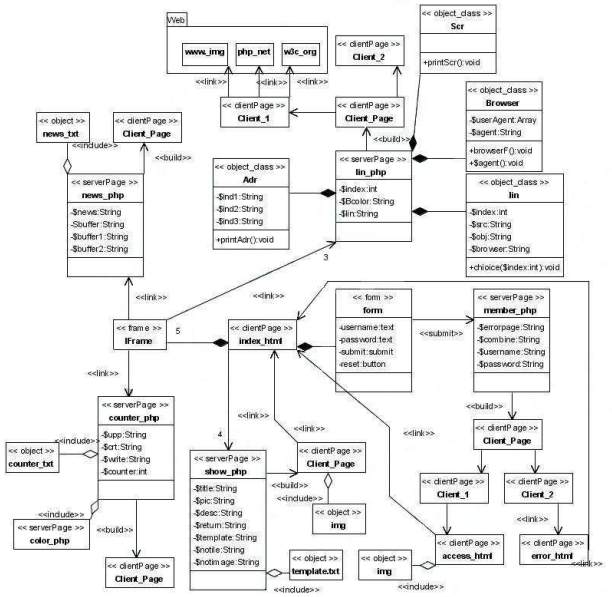

Figure 10: *MiniWP* Class Diagram

To apply our concerns mining technique, we select variable-class as *meta*-couple of "object-attribute" (to define class level crosscutting concerns). Then we complete the related *Context-Matrix,* where every cell is 1 if exist def-use relationship between variable and class (if variable is used or defined in class). Then we build the concept lattice related to our selected *meta*-couple. Figure 11: *MiniWP* "variable-class" concept lattice, shows the concept-lattice for MiniWP variable-class *meta*-couple. Based on this lattice we may define the 27 concepts for variable-class *meta*-couple. These are the defined concerns for our case. Then we may define the "attributes-concepts" matrix, we may perform attribute grouping operation, and we may iterate these two last steps to compose the concerns (Figure 12: MiniWP "variable-class" concerns, shows the iteration last step).

Figure 13: MiniWP Class Diagram "marked", shows the class diagram marked with concerns composed that let us slice the model.

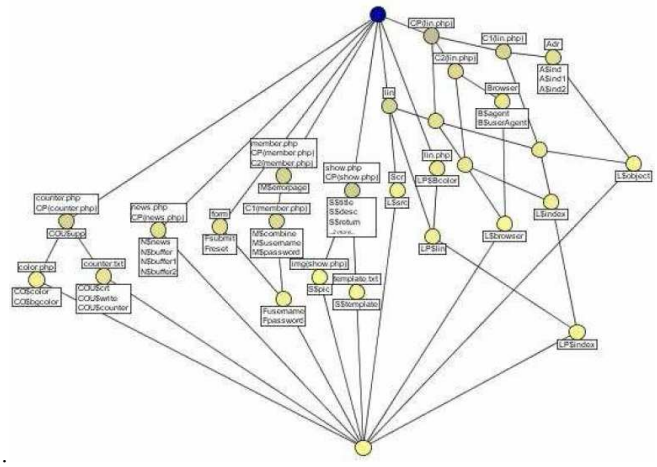

Figure 11: *MiniWP* "variables-class" concept lattice

	0	1	2	3	4
index.html					
show_php				1	
Client_Page(show_php)				1	
img(show_php)				1	
template.txt				1	
Web (Package)					
Iframe					
counter_php	1				
counter_txt	1				
color_php	1				
Client_Page(counter_php)	1				
news_php			1		
Client_Page(news_php)			1		
news_txt					
form		1			
member_php		1			
Client_Page(member_php)		1			
Client_1(member_php)		1			
Client_2(member_php)		1			
access_html					
img(access_html)					
error_html					
lin.php					1
Lin					1
Adr					1
Scr					1
Browser					1
Client_Page(lin_php)					1
Client_1(lin_php)					1
Client_2(lin_php)					1

Figure 12: *MiniWP* "variables-class" iterated concerns

IX. CONCLUSIONS

We proposed a semi-automatic multi dimensional concerns mining approach based on: concept analysis combined with a grouping technique. This approach may help the user in slicing applications via model analysis, and it may be used to semi-automatically define application test cases, or test coverage measures or also to understand software evolution. We are currently investigating efficient pruning techniques to reduce the number of concerns generated by our approach. We are also working on a tool to integrate our approach in the WAAT project.

 COPYRIGHT © 2005 ENFORMATIKA

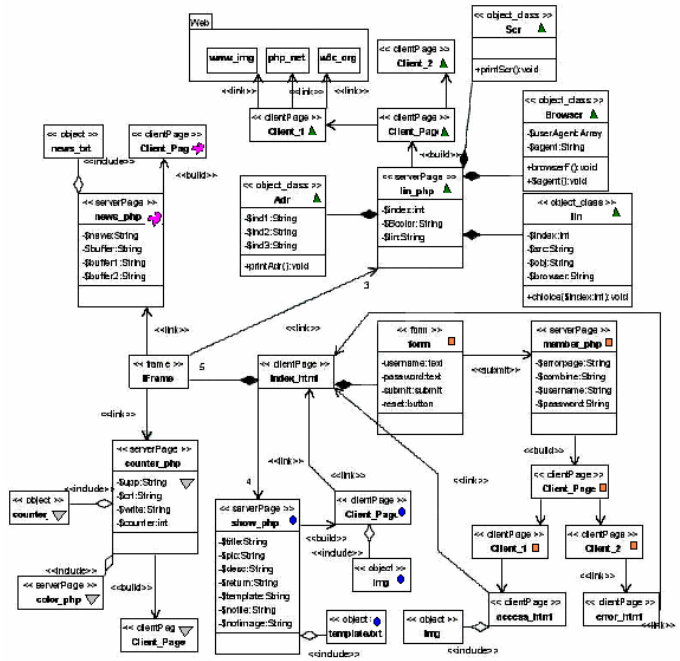

Figure 13: MiniWP Class Diagram "marked"

REFERENCES

[1] Apache Web Server – log file, http://httpd.apache.org/docs/logs.html

[2] T. Apiwattanapong, A. Orso and M.J. Harrold, "Efficient and Precise Dynamic Impact Analysis Using Execute-After Sequences*" 27th IEEE and ACM SIGSOFT International Conference on Software Engineering (ICSE 2005)*. USA. 2005

[3] Aspectj. http://eclipse.org/aspectj

[4] C. Bellettini, A. Marchetto, and A. Trentini, "Applying MDSOC to Web Applications" Accepted for publication – *9th World Multi-Conference on Systemics, Cybernetics and Informatics*. Orlando, Florida, USA. July 2005

[5] C. Bellettini C., A. Marchetto, and A. Trentini, "TestUml: User-Metrics Driven Web Applications Testing" 20th ACM Symposium on Applied Computing. USA 2005

[6] C. Bellettini C., A. Marchetto, and A. Trentini, "WebUml: Reverse Engineering of Web Applications". *19th ACM Symposium on Applied Computing (SAC 2004)*, Nicosia, Cyprus. March 2004

[7] C. Bellettini C., A.Marchetto, and A. Trentini, "Validation of Reverse Engineeered Web Application Model." *2th World Enformatika Conference* (WEC 2005). Istanbul, Turkey. February 2005

[8] S. Breu and J. Krinke. "Aspect Mining Using Event Traces". *19th. Conference on Automated Software Engineering 2004 (ASE 04)*, Linz, Austria. September 2004

[9] M. Bruntink, A. van Deursen, and T. Tourwè "An Initial Experiment in Reverse Engineering Aspects from Existing Applications". *11th IEEE Working Conference on Reverse Engineering (WCRE 04)*, Netherlands. November 2004

[10] M. Bruntink, A. van Deursen, R. van Engelen, and T. Tourwè, "An Evaluation of Clone Detection Techniques for Identifying Cross-Cutting Concerns". *IEEE International Conference on Software Maintenance (ICSM 04)*, 2004

[11] S. Ceri, P. Fraternali, and A. Bongio. "Web Modeling Language (WebML): a modeling language for designing Web sites." *Ninth International World Wide Web Conference (WWW9)*, Amsterdam, Netherlands. May, 2000

[12] J. Conallen. *Building Web Applications with UML*. Addison-Wesley, 2000

[13] A. Deursen, M. Marin, and L. Moonen, "Aspect Mining and Refactoring". *First International Workshop on REFactoring: Achievements, Challenges, Effects (REFACE03)*, Canada. November 2003

[14] G. A. Di Lucca, A. Fasolino, F. Faralli, and U. De Carlini, "Testing web applications". *International Conference on Software aintenance (ICSM'02)*, Montreal, Canada. October 2002

[15] Formal Concept Analysis, http://www.upriss.org.uk/fca/fca.html

[16] Galicia, http://www.iro.umontreal.ca/~galicia

[17] B. Ganter and R.Wille, "Formal Concept Analysis*". Springer-Verlag*, Berlin, Heidelberg, New York, 1996

[18] M. Han and C. Hofmeister, "Separating and Representing Navigation Concerns in Web Applications". *Lehigh University, Technical Reports*, 2004

[19] Httpunit. http://httpunit.sourceforge.net

[20] Hyperj. http://www.research.ibm.com/hyperspace

[21] T. Isakowitz, E. A. Stohr, and P. Balasubranian. "RMM: A Methodology for Structured Hypermedia Design." *Communications of the ACM*, August 1995

[22] C. Kallepalli and J. Tian. "Measuring and Modeling Usage and Reliability for Statistical Web Testing." *Ieee Transactions on Software Engineering*, November 2001

[23] G. Kiczales, J. Lamping, A. Mendhekar, C. Maeda, C. Lopes, J. Loingtier, and J. Irwin, "Aspect-Oriented Programming". *11th Europeen Conf. Object-Oriented Programming*, Springer Verlag. 1997.

[24] D. C. Kung, P. Hsia, and J. Gao. "Testing Object-Oriented." *Software. Wiley-IEEE Press*, 2002

[25] Mercury interactive. http://www.merc–int.com

[26] G. Murphy, A. Lai, R. Walker, and M. Robillard,. "Separating Features in Source Code: An Exploratory Study". *23rd International Conference on Software Engineering*, Toronto, Canada. May, 2001

[27] N. Noda and T. Kishi, "On Aspect-Oriented Design Applying Multi-Dimensional Separation of Concerns on Designing Quality Attributes". *First Workshop on Multi-Dimensional Separation of Concerns in Object-oriented Systems (OOPSLA'99)*, November 1999

[28] P. Peixoto, K. Fung, and D. Lowe. "A Framework for the Simulation of Web Applications." *Fourth International Conference on Web Engineering (ICWE 2004)*, M¨unchen, Germany. July 2004

[29] B. Pekilis. "Multi-Dimensional Separation of Concerns and IBM Hyper/J." *Technical Research Report*, University of Waterloo, Canada. January 2002

[30] Rational Rose Web Modeler, http://www.rational.com

[31] A. Reina, J. Torres, and M. Toro, "Aspect-Oriented Web Development vs. Non Aspect-Oriented Web Development". *Workshop of nalysis of Aspect-Oriented Software (AAOS 2003)*, University of Darmstadt, Germany. July 2003

[32] F. Ricca and P. Tonella, "Building a Tool for the Analysis and Testing of Web Applications: Problems and Solutions". *Tools and Algorithms for the Construction and Analysis of Systems (TACAS'200)*, Genova, Italy. April 2001

[33] D. Schwabe, R. Pontes, and I. Moura. "OOHDM-Web: An Environment for Implementation of Hypermedia Applications in the WWW." *SigWEB Newsletter*, 8, June 1999

[34] M. Siff and T. Reps, "Identifying modules via concept analysis." In M. J. Harrold and G. Visaggio, editors, Proc. IEEE Intl. Conf. on Software Maintenance, Bari, Italy, 1997. IEEE Comp. Soc. Press.

[35] J. Stanley and M. Sutton "Multiple Dimensions of Concern in Software Testing". *First Workshop on Multi-Dimensional Separation of Concerns in Object-oriented Systems (OOPSLA'99)*, November 1999

[36] P. Tarr, H. Ossher, W. Harrison, J. Stanley, and M. Sutton, "N-degrees of separation: Multi-Dimensional Separation of Concerns". *21st International Conference on SoftwareEngineering*, IEEE Computer Society Press, 1999

[37] P. Tonella and M. Ceccato, "Aspect Mining through the Formal Concept Analysis of Execution Traces". *11th IEEE Working Conference on Reverse Engineering (WCRE 04)*, Netherlands. November 2004

[38] ToscanaJ, http://toscanaj.sourceforge.net

 COPYRIGHT © 2005 ENFORMATIKA

Training Radial Basis Function Networks Using Reduced Sets as Center Points

Rana Yousef [1] and Khalil el Hindi[2]

Abstract—The behavior of Radial Basis Function (RBF) Networks greatly depends on how the center points of the basis functions are selected. In this work we investigate the use of instance reduction techniques, originally developed to reduce the storage requirements of instance based learners, for this purpose. Five Instance-Based Reduction Techniques were used to determine the set of center points, and RBF networks were trained using these sets of centers. The performance of the RBF networks is studied in terms of classification accuracy and training time. The results obtained were compared with two Radial Basis Function Networks: RBF networks that use all instances of the training set as center points (RBF-ALL) and Probabilistic Neural Networks (PNN). The former achieves high classification accuracies and the latter requires smaller training time. Results showed that RBF networks trained using sets of centers located by noise-filtering techniques (ALLKNN and ENN) rather than pure reduction techniques produce the best results in terms of classification accuracy. The results show that these networks require smaller training time than that of RBF-ALL and higher classification accuracy than that of PNN. Thus, using ALLKNN and ENN to select center points gives better combination of classification accuracy and training time. Our experiments also show that using the reduced sets to train the networks is beneficial especially in the presence of noise in the original training sets.

Keywords—Radial basis function networks, Instance-based reduction, PNN.

1. INTRODUCTION

Lazy learners such as k-nearest neighbor (KNN) and locally weighted linear regression are known to have good classification accuracies when the target function is very complex, but can be approximated using a combination of several local functions. However, lazy learners take long classification time which may limit their use in real life applications. Eager learners, on the other hand, have lower classification time because they produce global optimization of the target function (e.g. decision trees and neural networks). Radial basis function networks (RBF networks) are regarded as eager learners that can approximate the target function using a combination of several local functions [11]. Therefore, RBF networks combine the advantages of lazy and eager learners.

[1]KASIT, University of Jordan, Amman, Jordan. Email: rana@ju.edu.jo
[2]KASIT, University of Jordan, Amman, Jordan. Email: hindi@ju.edu.jo

An RBF network can be regarded as a two-layer feed-forward neural network. Such a network is characterized by a set of inputs and a set of outputs. In between the inputs and outputs there is a layer of processing units called hidden units. Each of them implements a radial basis function, which is normally chosen to be the Gaussian function. The adjustable parameters of RBF networks are the parameters of the basis functions and the linear weights. The parameters of the basis functions include the centers and the widths of the basis functions [1].

Given a set of training examples of the target function, RBF networks are typically trained in a two-stage process. First, the number of hidden units is determined and each hidden unit is defined by choosing the values of the centers and the widths that define its basis function. Second, the weights are trained to maximize the fit of the network to the training data [11].

Radial Basis Function Networks have been successfully applied to a large diversity of applications including Image Processing [3, 8], Speech Recognition [9], Medical Diagnosis [14] and Pattern Recognition [10].

As can be seen, the RBF networks are employed mostly in classification problems. Many experiments show that RBF networks are superior over other neural networks approaches because of the following reasons [15]:

- RBF networks are capable of approximating nonlinear mappings effectively.
- The training time of the RBF networks is quite low compared to that of other neural network approaches. This follows from the fact that the input layer and the output layer of an RBF network are trained separately and sequentially.
- RBF networks are quite successful for identifying regions of sample data not in any class because it uses a nonmonotonic transfer function based on the Gaussian density function.

RBF networks are superior to many other learning algorithms, in terms of classification accuracy and training time, only when the center points of the RBF network are properly located [7, 8, 13, 18].

Despite the advantages of RBF networks, they are not widely used. The main reason for this is that it is not straight forward to design an optimal RBF network to solve a given problem. Choosing too many center points has a negative effect on

 COPYRIGHT © 2005 ENFORMATIKA

training time, and choosing a few center points has a negative effect on the classification accuracy.

The goal of this work is to study the effect of using instance reduction techniques to choose center points for RBF networks with respect to classification accuracy and training time. Several Instance-Based Reduction Techniques [19] will be used to find a subset of the training set to act as center points of this RBF network. By doing so, we hope to achieve good classification accuracies and, at the same time, reduce the time needed to train RBF networks. We compared our technique with two RBF networks; RBF networks that use all instances of the training set as center points and Probabilistic Neural Network. We also used the same reduced sets to train the networks besides using then to locate the center points.

Instance Reduction techniques are usually applied to the training sets to reduce the storage requirements in instance based learning algorithms. See [19] for an excellent survey of such techniques

Section 2 gives an overview of the Radial Basis Function Networks. Section 3 presents reduction techniques that are used in this research.. Section 4 describes the experiments that were performed, the results obtained and an analysis of these results. Section 5 is the conclusion section.

2. RADIAL BASIS FUNCTION NETWORKS

Learning with Radial Basis Functions is one approach to function approximation that is closely related to distance-weighted regression and also to artificial neural networks [11]. The learned hypothesis is a function of the form:

$$f^*(x) = w_0 + \sum_{u=1}^{k} w_u K_u(d(x_u, x)) \qquad (2.1)$$

where w_0 is the bias unit, each x_u is an instance from X, the set of training instances, and where the kernel function $K_u(d(x_u, x))$ is defined so that it decreases as the distance $d(x_u, x)$ increases. k is a user-provided constant that specifies the number of kernel functions to be included. Even though *f*(x)* is a global approximation of *f(x)*, the contribution from each of the $K_u(d(x_u, x))$ terms is localized to a region surrounding the point x_u. It is common to choose each function $K_u(d(x_u, x))$ to be a Gaussian function centered at point x_u with some variance σ_u^2:

$$K_u(d(x_u, x)) = e^{\frac{1}{2\sigma_u^2} d^2(x_u, x)} \qquad (2.2)$$

Distance $d(x_u, x)$ is usually the Euclidian distance defined by the following equation for continuous attribute values:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2.3)$$

where *x* and *y* are vectors of n attribute values.

For discrete attribute values, the distance was assigned zero if the attributes have the same values and 1 otherwise.

The function given by Equation 2.1 can be viewed as describing a two-layer network where the first layer of units computes the values of the various $K_u(d(x_u, x))$ and each of the hidden units produces an activation determined by a Gaussian function centered at some instance x_u. Therefore, its activation will be close to zero unless the input *x* is near x_u. Then the second layer computes a linear combination of these first-layer unit values. An example of an RBF network is given in Figure 1, the network shown in this figure has just one output, multiple output units can also be included [11].

Radial Basis Function Networks are typically trained in a two-stage process. First, the number k of hidden units, which equals the number of center points, is determined and each hidden unit *u* is defined by choosing the values of x_u and σ_u^2 that define its kernel function K_u. Second, the weights w_u are trained to maximize the fit of the network to the training data.

2.1 Locating Center Points of Radial Basis Functions

One of the most important parameters that should be properly determined is the set of centers of the basis functions. The key to determining the locations of these centers is to view them as representing the input data density. Several alternative methods have been proposed for choosing such center points.

 COPYRIGHT © 2005 ENFORMATIKA

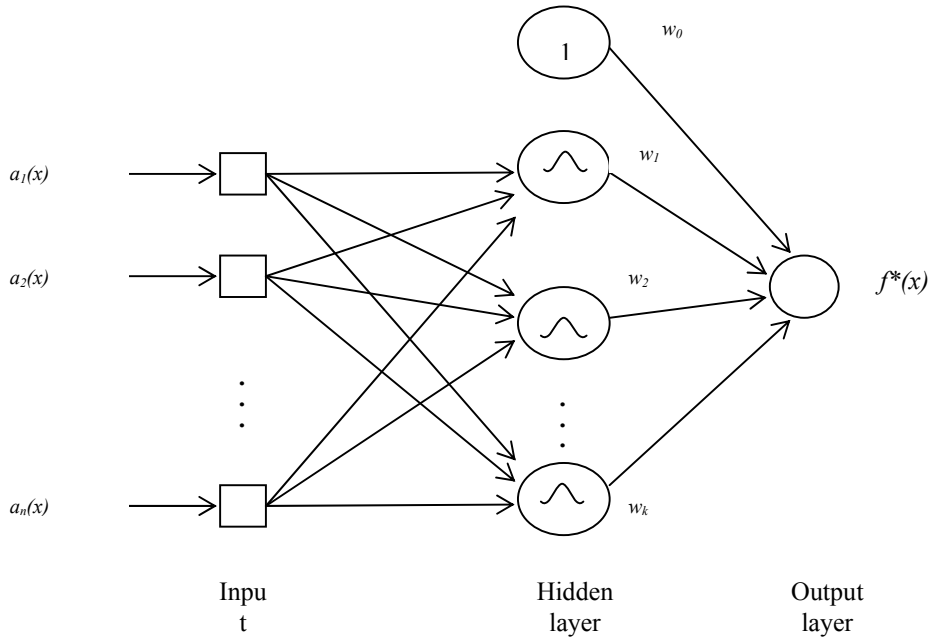

Figure 1: Radial Basis Function Network

One of the original methods used to choose a set of centers of an RBF network is to allocate a Gaussian kernel function for each training example $(x_i, f(x_i))$. Using this approach, RBF networks learn a global approximation to the target function in which each training example $(x_i, f(x_i))$ can influence the value of *f** only in the neighborhood of x_i. One advantage of this approach is that it allows the RBF network to fit the training data exactly. That is, for any set of *p* training examples the weights $w_0 \ldots w_p$ for combining the *p* Gaussian kernel functions can be set so that $f^*(x_i) = f(x_i)$ for each training example $(x_i, f(x_i))$. This is ok as long as the training set is noise free. However, most datasets suffer from noisy instances. Hence, the advantage of the exact fit will turn out to be one of the major disadvantages of this approach, a problem known as overfitting that may cause a bad generalization. This method also proved to be expensive in terms of memory requirements and training time [1, 4].

Another simple way to choose the RBF centers is to randomly locate a subset of the training data set. This choice is sensitive to how representative the selected data are of the overall population [1, 4].

Another approach is to cluster the input vectors and then locate the basis functions at the cluster centers. A cluster is a collection of objects which are similar to one another within the same cluster. A variety of clustering techniques can be used, such as the K-means algorithm [5]. There are a number of problems with clustering. Among them:

- Current clustering techniques do not address all the requirements adequately (and concurrently).
- Dealing with large number of dimensions and large number of data items can be problematic because of time complexity.
- The effectiveness of the method depends on the definition of "distance" (for distance-based clustering).
- If an obvious distance measure doesn't exist we must "define" it, which is not always easy, especially in multi-dimensional spaces.
- The result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

Many other techniques were also suggested, some are based on genetic algorithms [10], others are based on decision trees [7] and others are based on supervised learning [18].

2.2 Determining the Widths of Radial Basis Functions

The width of each Gaussian in the hidden layer is frequently determined by heuristics. The objective of these schemes is to achieve a certain amount of overlapping between the responses of neighboring units so that the radial basis functions form a smooth and contiguous interpolation over the regions of the input space that they represent.

 COPYRIGHT © 2005 ENFORMATIKA

One of the heuristics that can be used is to set the width for each basis function as the average Euclidean distance of its P nearest neighboring functions [6]. Calculating the width in this way is computationally expensive especially when the number of center points is large.

Another way is to let all radial basis functions have the same width given by:

$$\sigma_a = \frac{d}{\sqrt{2k}} \qquad (2.4)$$

where d is the maximum distance between the chosen centers and k is the number of center points. Such a choice for the width does not guarantee that the Gaussian functions are too peaked or too flat. These extremes have to be avoided [15].
In this work widths are calculated as the standard deviation of the centers' values of that attribute. The value of the width for each attribute is given by:

$$\sigma_{a_j} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(a_j(x_i) - \mu_a)^2} \qquad (2.5)$$

where a_j is attribute j of the instance, j = 1..n, where n is the number of attributes. k is the number of center points of the basis function, μ_a is the mean of the values of attribute a_j in the set of center points and is given by the equation:

$$\mu_a = \frac{1}{k}\sum_{u=1}^{k} a_j(x_u) \qquad (2.6)$$

2.3 Calculating the output layer weights

Weights of the output layer can be updated using the gradient decent rule [11]:

$$\Delta w_j = \eta \sum_{x \in D} (f(x) - f^*(x)) a_j(x) \qquad (2.7)$$

where η is the learning rate, D is the training set.

However, because the transformation between the inputs and the corresponding outputs of the hidden units is fixed once the basis parameters are determined; the network can be viewed as equivalent to a single-layer network with linear output units [12]. Minimization of the sum-squared error function yields the well-known least-squares solution for the weights:

$$\mathbf{A}\,\mathbf{w} = \mathbf{H}^{\mathrm{T}}\,\mathbf{F} \qquad (2.8)$$

From this equation one can find the weight vector as the solution of the following so-called normal equation:

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{H}^{\mathrm{T}}\,\mathbf{F} \qquad (2.9)$$

where,

F is the vector of training set outputs. For p instances in the training set:

$$\mathbf{F} = \begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_p) \end{bmatrix} \qquad (2.10)$$

H is the design matrix where each element of this matrix is the value resulting from applying the Gaussian function, given by equation 2.2, over instance **x**. Suppose $h_i(\mathbf{x}_j)$ denotes the value of applying the Gaussian function with center i to instance j, the design matrix **H** will be given by:

$$\mathbf{H} = \begin{bmatrix} h_1(\mathbf{x}_1) & h_2(\mathbf{x}_1) & \dots & h_k(\mathbf{x}_1) \\ h_1(\mathbf{x}_2) & h_2(\mathbf{x}_2) & \dots & h_k(\mathbf{x}_2) \\ \vdots & \vdots & & \vdots \\ h_1(\mathbf{x}_p) & h_2(\mathbf{x}_p) & \dots & h_k(\mathbf{x}_p) \end{bmatrix} \qquad (2.11)$$

x is a vector corresponding to an instance from the training set and consists of n of attributes.

$$\mathbf{x} = \begin{bmatrix} a_1(x) \\ a_2(x) \\ \vdots \\ a_n(x) \end{bmatrix} \qquad (2.12)$$

$\mathbf{A}^{-1}$, the variance matrix, is

$$\mathbf{A}^{-1} = (\mathbf{H}^{\mathrm{T}}\,\mathbf{H})^{-1} \qquad (2.13)$$

And the result is the vector of weights **w**:

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix} \qquad (2.14)$$

this weight vector minimizes the following cost function:

$$C = \sum_{i=1}^{p} (f(x_i) - f^*(x_i))^2 \qquad (2.15)$$

 COPYRIGHT © 2005 ENFORMATIKA

As can be seen, the network weights can be computed by fast linear matrix inversion techniques. In practice, it is tended to use singular value decomposition (SVD) to avoid possible ill-conditioning of $\mathbf{H}$, i.e. ($\mathbf{H}^T \mathbf{H}$) being singular (the matrix ($\mathbf{H}^T \mathbf{H}$)$^{-1}$ can't be found). Such methods provide an approximate inverse by diagonalising the matrix, inverting only the eigenvalues which exceed zero by a parameter-specified margin, and transforming back to the original coordinates. This provides an optimal minimum-norm approximation to the inverse in the least-mean-squares sense [12].

2.4 Probabilistic Neural Network Architecture

Probabilistic neural networks can be used for classification problems. When an input is presented, the first layer computes distances from the input vector to the training input vectors, and produces a vector whose elements indicate how close the input is to a training input. The second layer sums these contributions for each class of inputs to produce as its net output a vector of probabilities. Finally, a "compete" transfer function on the output of the second layer picks the maximum of these probabilities, and produces a 1 for that class and a 0 for the other classes [17].

3. Instance Reduction Techniques

Some reduction techniques were designed for instance-based learning algorithms to decide what instances to store for use during generalization in order to avoid excessive storage and time complexity, and possibly to improve classification accuracy by avoiding noise and overfitting. In this section we review instance reduction techniques that we used to locate center points for RBF networks. These techniques were chosen carefully to cover the wide spectrum of reduction techniques from noise filtering to pure reduction techniques. See [19] for an excellent survey of other techniques. Most of the algorithms use a subset S of the original instances in the training set T as their representation.

3.1) Noise filtering techniques

3.1.1) Edited Nearest Neighbor (ENN) - (Wilson, 1972):

In this algorithm S starts out the same as T, and then each instance in S is removed if it does not agree with the majority of its k nearest neighbors (with k = 3, typically). This edits out noisy instances as well as close border cases, leaving smoother decision boundaries. It also retains all internal points, which keeps it from reducing the storage requirements as much as most other reduction algorithms.

3.1.2) All k-NN - (Tomek, 1976):

This algorithm is an extension of the ENN. It works as follows: for i = 1 to k, flag as bad any instance not classified correctly by its i nearest neighbors. After completing the loop all k times, remove any instances from S flagged as bad. This method retains internal points thus limiting the amount of reduction that it can accomplish. The above three methods, namely ENN, RENN and ALL k-NN serve more as noise filters than serious reduction algorithms.

3.2) Pure Reduction Algorithms

3.2.1) EXPLORE - (Cameron-Jones, 1995):

The encoding length heuristic is used to determine how good the subset S is in describing T. The basic algorithm begins with a growing phase that takes each instance i in T and adds it to S if that results in a lower cost than not adding it. The growing phase can be affected by the order of presentation of the instances.
After all instances are seen, pruning is done, where each instance i in S is removed if doing so lowers the cost of the classifier.

The Explore method then performs 1000 mutations to try to improve the classifier. Each mutation tries adding an instance to S, removing one from S, or swapping one in S with one in T-S, and keeps the change if it does not increase the cost of the classifier. The generalization accuracy of the Explore method is quite good empirically, and its storage reduction is much better than most other algorithms.

3.3) Ordered Removal (Reduction and noise filtering algorithm)

3.3.1) DROP2 - (Wilson and Martinez, 2000)

DROP2 removes an instance, P, only if at least as many of its associates in the original training set (including those that may have been pruned) would be classified correctly without it. Also, DROP2 attempts to remove center instances (that are far from enemy instances) before boarder instances (that are close to enemy instances). This is done by sorting the instances in the reduced set by the distance to their nearest enemy.

3.3.2) DROP5 - (Wilson and Martinez, 2000)

DROP5 modifies DROP2 by adding a pre-pass to it. During this new pass instances that are near their enemy are removed first and proceeds outwards. After this pass DROP2 is applied until no further improvement can be made. The first pass serves to remove noisy instances in addition to many internal instances.

 COPYRIGHT © 2005 ENFORMATIKA

4. USING THE REDUCED SETS AS CENTER POINTS

In this section we present the results we obtained using the reduced sets as center points. The intuition is that if the discussed reduction techniques really retain the most relevant instances, then these instances should serve as good center points for RBF networks.

Twenty datasets obtained from the Machine Learning Repository at University of California Irvine (UCI) were used to compare the different methods of locating center points. A summary of the datasets appears in Table 1.

Some of the datasets that were used in the experiments contained missing attribute values (e.g. breast-cancer-wisconsin, heart.cleveland.2), which is the case in many real world applications. A heuristic method was used to deal with those missing values before training begins: If the attribute value is continuous, the average value of that attribute vector is used; if it is discrete, the most common value is used.

Table 1 Datasets used in the experiments

Datasets	Instances	Attributes	Classes
australian	690	14	2
bcw	683	9	2
breast-cancer-wisconsin	699	9	2
crx	55	15	2
echocardiogram	132	9	2
glass	214	9	7
Heart	270	13	2
heart.cleveland.2	303	13	5
heart.hungarian.2	294	13	5
heart.long-beach-va.2	200	13	5
heart.swiss.2	123	13	5
horse-colic	301	23	2
ionosphere	351	34	2
iris	150	4	3
liver.bupa	345	6	2
pima-indians-diabetes	768	8	2
tic-tac-toe	958	9	2
vehicle	846	18	4
Voting	435	16	2
wine	178	13	3

Ten-fold cross-validation was used in all experiments.

In this work, the centers of the Gaussian functions were located five times for each dataset, each time using one of the reduction techniques described above. Locating the centers this way is considered to be a supervised learning that takes the advantages of the available class values to find a reduced subset of the training set. This helps reduce storage requirement and/or remove noisy instances from the training set.

Table 2 presents the size ratio of the reduced sets obtained using the 5 reduction techniques discussed previously, the reduction algorithms ENN and ALLKNN have the highest storage requirements. The size of the centers generated using ENN ranges between 59.16% and 87.48% of the training set. The average size of the reduced sets is 76.21%. The set of centers generated using ALLKNN is smaller than that generated by ENN, which ranges between 46.30% (in the dataset liver.bupa) and 85.85% (in the dataset bcw), with an overall average equals to 70.02%.

ENN and ALLKNN are considered to be noise filtering techniques rather than pure reduction techniques [19], so they only remove noisy instances from the training set and keeps the rest of the instances, that's why the storage requirement for these two algorithms is the greatest.

The Reduction Algorithm EXPLORE causes a huge reduction in the datasets that reaches on average to only 1.06% of the training set. This of course will degrade the classification accuracy, as will be seen later.

DROP2 and DROP5 produced reduced sets that had an average ratio equal to 17.84% and 13.60% of the original training set, respectively.

After training the network, the testing is done using the test set. To perform the test, we need to have the set of centers and the widths that were used in training, and the weights that were generated as a result of the training. All these will be used to test an unseen instance as follows:

An indication of the distance between the unseen instance and each one of the centers used in training is calculated using the Gaussian function. In other words, the design matrix **Ht** is found were each value in this matrix is the result of applying the Gaussian function on the unseen example. Of course, the centers and widths that were used in training are used here. These values are then multiplied by the corresponding weight. The class of this instance is calculated as the summation of these values rounded to an integer value.

In Matrix notation, the predicted class of any unseen example x is the value resulting from the approximation function $f^*(x)$:

$$f^*(x) = \text{round}\ (\mathbf{Ht} * \mathbf{w})$$

where **Ht** the design matrix, **w** the weight vector.

The predicted class value is compared to the actual class value and the classification accuracy is found for this fold.

Another important factor is the training time. The training time was measured for each fold, and then the average time was determined.

 COPYRIGHT © 2005 ENFORMATIKA

Table 2 Training sets' sizes and the ratio of the original training set that is used as a set of centers for the RBF network

Datasets	Training Set size	Centers (%)				
		ENN	ALLKNN	Explore	DROP2	DROP5
australian	621	77.78	70.69	0.32	14.01	9.02
bcw	615	87.48	85.85	0.33	5.04	3.25
breast-cancer	629	87.28	85.69	0.32	5.56	3.66
crx	50	70.00	56.00	2.00	26.00	22.00
echocardiogram	119	80.67	76.47	0.84	12.61	12.61
glass	193	63.73	58.55	4.15	27.98	23.83
Heart	243	75.72	66.26	0.82	21.40	15.64
heart.cleveland.2	273	75.09	65.57	0.73	19.41	14.29
heart.hungarian.2	265	74.34	68.30	0.75	15.47	9.81
heart.long-beach-va.2	180	67.22	57.22	1.11	24.44	15.56
heart.swiss.2	111	82.88	79.28	0.90	5.41	2.70
horse-colic	271	72.69	63.84	0.74	19.93	15.13
ionosphere	316	75.32	73.73	0.63	11.39	9.18
iris	135	86.67	85.19	2.22	15.56	11.85
liver.bupa	311	59.16	46.30	0.64	35.05	26.69
pima-indians-diabetes	691	68.89	58.18	0.29	25.33	19.54
tic-tac-toe	862	80.63	75.99	0.23	22.51	18.79
vehicle	761	66.49	57.82	1.84	29.17	23.78
Voting	392	85.97	84.44	0.51	6.89	4.59
wine	160	86.25	85.00	1.88	13.75	10.00
Average		76.21	70.02	1.06	17.84	13.60

Table 3 shows the classification accuracy (Acc) and the training time (Time) for the 20 datasets. Each column has a title which is the reduction technique name prefixed with RBF. For the sake of comparison, table 3 also shows the classification accuracy and training time obtained using the whole training set as center points. These results are shown in the RBF-ALL column.

Each value marked with an asterisk indicates an improvement in accuracy compared to RBF-ALL no matter how small.

As can be seen from the table, among all reduction techniques, RBF-ENN and RBF-ALLKNN have the highest classification accuracies. Despite the fact that the average classification accuracies for RBF-ENN and RBF-ALLKNN is about the same as RBF-ALL, some improvements can be seen in classification accuracies for about half of the datasets (those values marked with an asterisk). The average classification accuracies for RBF-ENN and RBF-ALLKNN equals to 81.17% and 81.21%, respectively, which is about the same as RBF-ALL that have an average classification accuracy of 81.47%. This is expected since ENN and ALLKNN are noise filters, and retain most of the instances in the training set, they only edit out noisy instances and retain the center points. A good RBF network is the one which uses a set of centers that is noise-free and is a good representative of the training set. The two algorithms can be used to locate centers having these two important characteristics.

The differences in classification accuracies are statistically insignificant as will be seen later. The most importance achievement is the reduction in training time. In all cases, training time was reduced compared to the time needed to train RBF-ALL. On average, RBF-ALLKNN needs only half the training time needed by RBF-ALL, and RBF-ENN needs about 60% of the training time required by RBF-ALL, i.e. there is a substantial reduction in time.

 COPYRIGHT © 2005 ENFORMATIKA

Table 3: Classification accuracy (Acc) and training time (Time) for RBF networks generated using (1) all instances as center points (2) centers located using the 5 reduction techniques.

Datasets	RBF-ALL		RBF-ENN		RBF-ALLKNN		RBF-EXPLORE		RBF-DROP2		RBF-DROP5	
	Acc	Time	Acc	Time	A cc	Time	A cc	Time	A cc	Time	A cc	Time
australian	83.91	37.22	84.06*	23.61	83.77	19.50	58.55	0.14	79.13	3.84	81.30	2.59
bcw	96.63	21.37	96.93*	18.06	96.93*	16.51	69.11	0.15	91.80	1.41	91.51	0.96
breast	96.85	25.93	97.00*	18.98	97.00*	18.10	68.81	0.15	90.56	1.68	91.70	1.12
crx	54.55	0.04	50.91	0.03	50.91	0.02	38.18	0.03	52.73	0.02	45.46	0.01
echo	87.88	0.20	90.15*	0.14	89.39*	0.11	81.82	0.03	85.61	0.02	87.88*	0.02
glass	46.73	0.98	47.20*	0.35	47.20*	0.26	35.98	0.06	42.06	0.12	43.46	0.10
Heart	79.63	2.51	80.00*	1.34	79.26	1.00	60.74	0.03	75.93	0.32	78.89	0.22
heart.cleve	80.53	3.86	81.85*	2.19	82.18*	1.65	58.42	0.04	75.58	0.47	77.56	0.32
heart.hung	79.25	3.18	79.59*	1.81	79.93*	1.47	67.35	0.04	76.53	0.35	77.89	0.20
heart.long	70.50	0.85	70.00	0.32	72.00*	0.20	29.00	0.03	65.50	0.07	64.50	0.06
heart.swiss	88.62	0.16	87.81	0.12	87.81	0.10	45.53	0.02	47.15	0.02	30.89	0.04
horse-colic	69.44	3.81	68.44	2.29	67.77	1.56	33.22	0.04	60.80	0.44	58.47	0.35
ionosphere	94.59	6.78	94.30	4.01	94.59*	3.84	35.90	0.05	76.92	0.51	91.45	0.42
iris	98.00	0.21	97.33	0.15	97.33	0.14	85.33	0.01	96.67	0.03	98.00*	0.03
liver.bupa	69.28	3.84	70.15*	1.66	69.86*	1.16	51.59	0.03	69.28*	0.97	69.57*	0.76
pima	75.65	47.30	75.13	24.88	75.78*	16.74	65.23	0.18	73.83	6.31	75.00	6.12
tic-tac-toe	99.79	92.96	97.70	60.40	97.60	53.74	41.75	0.25	90.61	10.51	85.28	9.40
vehicle	67.26	70.99	65.25	33.59	65.25	22.17	42.44	1.10	66.19	9.94	66.79	8.20
Voting	95.40	8.14	95.86*	6.00	95.86*	5.53	66.67	0.06	92.87	0.50	92.87	0.38
wine	94.94	0.65	93.82	0.43	93.82	0.37	56.18	0.01	79.21	0.04	84.27	0.03
Average	81.47	16.55	81.17	10.02	81.21	8.21	54.59	0.12	74.45	1.88	74.64	1.57

* Improvement in classification accuracy, or have same classification accuracy.

RBF-EXPLORE have the worst average classification accuracies of 54.59%. This is expected since EXPLORE drastically reduces the training set, so the number of centers is not sufficient to train the RBF network. RBF-DROP2 and RBF-DROP5 have moderate classification accuracies equal to 74.45% and 74.64%, respectively. These two algorithms use a noise filer phase but also cause a relatively large reduction in the training set, because they attempt to retain only non-noisy border points. This will remove the mass of the training set located in the centers.

Comparing Results with Probabilistic Neural Networks (PNN)

In this section, RBF-ALLKNN and RBF-ENN are compared to PNN. We used the Matlab implementation of PNN[3]. Table 4 compares the classification accuracies and training times for PNN, RBF-ENN and RBF-ALLKNN.

As can be seen from the table 4, the classification accuracy for RBF-ENN and RBF-ALLKNN is much better than that of PNN, but the training time is greater. PNN uses all instances in the training set as a weight vector for the first layer and tries to learn the center points by adjusting these weights. From the results, it is apparent that learning the center points is fast but this is on the expense of the classification accuracy.

[3] The function is called "newpnn" and it is present in the neural network toolbox.

 COPYRIGHT © 2005 ENFORMATIKA

Table 4: The Classification Accuracies and Training Times for PNN, RBF-ENN and RBF-ALLKNN.

Datasets	PNN		RBF-ENN		RBF-ALLKNN	
	Acc	Time	Acc	Time	Acc	Time
australian	65.80	0.34	84.06	23.61	83.77	19.50
bcw	96.78	0.30	96.93	18.06	96.93	16.51
breast	96.28	0.31	97.00	18.98	97.00	18.10
crx	67.27	0.19	50.91	0.03	50.91	0.02
echoc	87.12	0.19	90.15	0.14	89.39	0.11
glass	68.22	0.21	47.20	0.35	47.20	0.26
Heart	62.22	0.21	80.00	1.34	79.26	1.00
heart.cleve	62.38	0.22	81.85	2.19	82.18	1.65
heart.hung	62.93	0.22	79.59	1.81	79.93	1.47
heart.long	73.00	0.20	70.00	0.32	72.00	0.20
heart.swiss	90.24	0.28	87.81	0.12	87.81	0.10
horse-colic	66.45	0.21	68.44	2.29	67.77	1.56
ionosphere	86.90	0.23	94.30	4.01	94.59	3.84
iris	92.67	0.19	97.33	0.15	97.33	0.14
liver.bupa	66.09	0.20	70.15	1.66	69.86	1.16
pima	74.74	0.27	75.13	24.88	75.78	16.74
tic-tac-toe	77.45	0.33	97.70	60.40	97.60	53.74
vehicle	59.81	0.41	65.25	33.59	65.25	22.17
Voting	92.41	0.22	95.86	6.00	95.86	5.53
wine	71.35	0.19	93.82	0.43	93.82	0.37
Average	76.01	0.25	81.17	10.02	81.21	8.21

Statistical Analysis

Table 5 shows the 95% confidence intervals for the time needed by RBF-ENN and RBF-ALLKNN as a percentage of the time needed by RBF-ALL.

As can be seen from the table, using RBF-ENN and RBF-ALLKNN reduces the training time needed by RBF-ALL. The average lower bound is 47.17% and 39.29% of the time needed by RBF-ALL when using RBF-ENN and RBF-ALLKNN, respectively. The upper bounds are 75.54% and 60.34%, respectively. This means that with 95% confidence (probability), the time needed by RBF-ENN is lower than that required by RBF-ALL by at least 47.17% and by at most 75.54%.

Table 6 shows the confidence intervals for the difference in classification accuracy between RBF-ALL and each of RBF-ENN and RBF-ALLKNN. From the table we can see that the classification accuracies of RBF-ALLKNN and RBF-ENN are comparable to that of RBF-ALL, i.e. the differences in classification accuracies are not statistically significant.

Table 5: 95% Confidence Intervals for the percentage of the time needed by RBF-ALL

Datasets	RBF-ENN		RBF-ALLKNN	
	Lower bound	Upper bound	Lower bound	Upper bound
australian	55.65	71.43	46.47	58.44
bcw	74.98	94.66	68.85	86.14
breast	58.01	89.87	58.12	82.68
crx	25.67	135.8	2.02	102.1
echoc	40.85	102.97	34.56	77.57
glass	27.66	43.25	19.63	32.93
Heart	45.44	61.59	33.34	46.20
heart.cleve	47.94	65.83	34.74	50.77
heart.hung	47.62	66.98	40.02	52.63
heart.long	29.72	44.73	19.77	27.14
heart.swiss	54.64	90.77	51.73	78.96
horse-colic	53.16	67.00	29.57	52.20
ionosphere	53.38	65.04	49.78	63.47
iris	58.20	82.12	58.86	74.74
liver.bupa	20.93	65.31	26.08	34.53
pima	44.73	60.82	26.08	44.41
tic-tac-toe	56.92	73.21	51.55	64.18
vehicle	41.96	52.90	28.72	33.85
Voting	63.94	84.12	59.58	76.83
wine	42.09	92.28	46.31	67.03
Average	47.17	75.54	39.29	60.34

Table 6: 95% Confidence Intervals for the difference in classification accuracy between RBF-ALL and each of RBF-ENN and RBF-ALLKNN

Datasets	RBF-ENN		RBF-ALLKNN	
	Lower bound	Upper bound	Lower bound	Upper bound
australian	-1.36	1.65	-2.47	2.18
bcw	-1.41	2.00	-1.41	2.00
breast	-1.34	1.63	-0.69	0.97
crx	-16.21	9.55	-16.21	9.55
echoc	-12.14	16.31	-11.05	13.80
glass	-21.97	23.01	-23.97	25.01
Heart	-1.78	2.52	-2.52	1.78
heart.cleve	-2.93	5.58	-2.65	5.96
heart.hung	-1.65	2.34	-1.91	3.24
heart.long	-5.70	4.70	-6.05	9.05
heart.swiss	-5.66	4.00	-5.66	4.00
horse-colic	-7.47	5.47	-6.86	3.53
ionosphere	-4.87	4.30	-4.28	4.28
iris	-4.53	3.20	-4.53	3.20
liver.bupa	-5.27	6.95	-7.93	9.07
pima	-2.82	1.79	-3.67	3.93
tic-tac-toe	-13.50	9.34	-14.21	9.84
vehicle	-6.09	2.07	-8.14	4.11
Voting	-3.42	4.34	-3.42	4.34
wine	-10.62	8.39	-10.62	8.39
Average	-6.54	5.96	-6.91	6.41

 COPYRIGHT © 2005 ENFORMATIKA

Training RBF networks using reduced set of the training sets

In the previous experiments, the reduction techniques were used to locate the set of center points of an RBF network. As we saw from the results, using these reduced sets to locate the center points caused a substantial reduction in the training time, but concerning the classification accuracy, there was a reduction in classification accuracy with a minimal reduction when ENN and ALLKNN algorithms were used. The reason for this is that these two algorithms are noise-filtering techniques that remove noisy instances. In the following experiments we used the reduction techniques not only to locate center points but also to produce a reduced set to train the network. The following tables give the results obtained from such experiments. First, the experiments were performed on the original training sets, then they were performed on noisy training sets, noise was added artificially with percentages equal to 5%, 10% and 15%. For short, RBF networks trained using a reduced set of the training set is given the reduction technique name prefixed with RBF-T.

Table 7: Classification accuracy (Acc) and time (T%) for RBF networks trained using a reduced set of training set. Time was calculated as a percentage of the time needed to train the RBF network using all instances of the training set.

Datasets	RBF-All	RBF-T-ENN		RBF-T-ALLKNN		RBF-T-Explore		RBF-T-DROP2		RBF-T-DROP5	
	Acc	Acc	T%	Acc	T%	Acc	T%	Acc	T%	Acc	T%
australian	83.913	85.94	56.34	86.38	38.14	56.67	0.18	74.20	0.47	75.362	0.35
bcw	96.633	96.78	77.82	97.07	71.47	67.50	0.25	88.73	0.14	88.29	0.12
breast	96.853	96.28	71.79	96.85	81.87	67.24	0.18	86.12	0.18	90.42	0.31
crx	54.545	60.00	91.40	60.00	71.04	40.00	21.27	49.09	85.52	52.73	77.83
echoc	87.879	89.39	79.45	86.36	72.38	81.82	6.96	85.61	12.27	86.36	22.76
glass	46.729	52.80	29.99	51.87	24.85	28.97	4.55	38.79	7.12	44.39	7.63
Heart	79.63	82.59	50.37	81.85	30.75	60.00	2.06	75.93	3.69	76.30	4.30
heart.cleve	80.528	83.17	24.62	82.84	15.61	57.43	0.60	72.61	1.21	75.25	1.20
heart.hung	79.252	81.63	51.31	80.95	36.92	68.37	1.72	72.11	2.01	71.43	1.35
heart.long	70.5	73.50	32.28	74.50	24.16	28.00	3.20	62.50	5.88	57.00	3.82
heart.swiss	88.618	90.24	102.8	89.43	95.16	53.50	7.19	42.28	24.53	20.33	31.88
horse-colic	69.435	74.75	43.40	72.76	26.69	33.22	1.08	50.17	2.25	49.50	1.89
ionosphere	94.587	94.59	44.67	95.16	38.88	35.90	2.75	57.27	0.85	82.91	0.85
iris	98	97.33	59.24	97.33	57.35	79.33	13.03	92.67	7.46	93.33	1.78
liver.bupa	69.275	66.96	23.42	66.96	11.58	52.17	0.28	64.35	5.59	63.48	3.27
pima	75.651	75.65	42.97	76.30	21.50	64.71	0.11	74.09	2.40	75.39	0.84
tic-tac-toe	99.791	85.80	67.18	81.11	50.73	39.46	0.04	66.49	1.55	65.45	0.85
vehicle	67.258	65.25	28.85	66.55	19.53	38.77	0.07	54.85	2.76	60.64	1.77
Voting	95.402	95.63	67.51	95.63	65.20	65.75	0.73	89.43	0.65	86.90	0.72
wine	94.944	94.94	64.26	94.94	60.07	46.63	6.13	74.16	11.53	79.78	7.26
Average	81.47	82.16	55.48	81.74	45.69	53.27	3.62	68.57	8.90	69.76	8.54

As can be seen from table 7, there's an improvement in classification accuracy in almost all cases when RBF networks are trained using a reduced set generated by ENN and ALLKNN, particularly the classification accuracy was improved for 12 datasets from the overall 20 datasets in the case of RBF-T-ENN and 14 datasets from the overall 20 datasets in the case of RBF-T-ALLKNN. Using all instances in the training set to train the RBF network is vulnerable to overfitting. Hence, the misclassification occurs, especially in the presence of noise. Since ENN and ALLKNN are noise filtering techniques, RBF networks trained using reduced sets generated by these reduction techniques have the highest classification accuracies compared to RBF networks trained using reduced set generated by other reduction techniques. The average classification accuracies for these networks were increased from 81.47% (using all instances of the training set) to 82.16% and 81.74% (using RBF-T-ENN and RNF-T-ALLKNN, respectively).

RBF-T-EXPLORE has the worst classification accuracies, which is 53.27% on average. RBF-T-DROP2 and RBF-T-DROP5 have moderate classification accuracies equals to 68.57% and 69.76%, respectively, which is expected because both of these techniques have a noise filtering phase.

In all cases, the training time was considerably reduced compared to the time needed to train RBF-ALL. This is because there is a reduction in storage requirement. It is obvious from the table that RBF-T-ENN and RBF-T-ALLKNN achieve the best combination of classification accuracy and training time.

Table 8 shows the results obtained from the same experiments but with 5% noise added to the training set.

 COPYRIGHT © 2005 ENFORMATIKA

Table 8: Classification accuracy (Acc) and time (T%) for RBF networks trained using a reduced set of training set in presence of 5% noise.

Datasets	RBF-ALL	RBF-T-ENN		RBF-T-ALLKNN		RBF-T-Explore		RBF-T-DROP2		RBF-T-DROP5	
	Acc	Acc	T%	Acc	T%	Acc	Acc	T%	Acc	T%	Acc
australian	82.174	85.362	44.798	84.928	35.379	55.797	0.1324	74.928	0.618	69.13	0.2921
bcw	96.05	97.07	68.56	96.78	48.48	68.23	0.20	92.97	0.24	88.73	0.15
breast-	96.57	96.28	52.53	96.57	44.10	66.38	0.11	92.28	0.25	88.13	0.21
crx	50.91	54.55	114.71	65.46	92.16	43.64	29.90	47.27	38.24	49.09	76.96
echoc	88.64	86.36	90.27	87.88	73.75	81.82	7.81	86.36	16.13	86.36	15.88
glass	42.06	48.60	24.94	50.94	16.75	35.51	5.60	36.92	4.70	37.38	4.70
Heart	80.00	81.85	34.61	79.26	21.60	56.30	1.55	75.56	2.70	77.04	2.73
heart.cleve	79.21	82.18	20.03	81.52	11.26	54.79	0.57	68.98	1.35	67.66	0.84
heart.hung	78.91	81.97	40.76	81.63	27.49	65.65	1.48	70.07	1.63	69.39	1.80
heart.long-	72.00	76.50	30.45	76.00	21.93	29.50	3.18	59.00	3.22	51.00	4.79
heart.swiss.	80.49	91.06	84.03	91.06	73.82	93.50	1.96	68.29	28.40	30.89	18.46
horse-colic	67.44	74.09	37.92	73.09	21.26	33.22	1.32	54.82	2.79	53.49	2.35
ionosphere	93.16	94.02	41.29	94.30	34.80	36.18	0.68	62.39	0.84	74.36	0.97
iris	96.67	97.33	60.39	96.67	47.22	82.67	14.86	85.33	9.54	90.00	14.98
liver.bupa	68.41	64.06	23.70	65.22	9.66	53.33	0.72	64.93	5.84	62.90	3.11
pima	74.87	75.65	31.83	75.13	22.30	65.10	0.12	73.83	2.70	72.40	0.77
tic-tac-toe	80.69	82.78	48.84	80.06	37.24	40.08	0.06	48.64	1.93	62.00	0.94
vehicle	60.99	65.25	26.73	65.13	14.60	40.07	0.06	55.32	2.69	58.51	1.55
Voting	95.17	96.09	53.53	95.86	48.99	64.37	0.58	84.60	0.94	89.43	0.92
wine	88.76	92.70	62.08	93.82	55.19	46.63	5.52	70.79	8.96	76.40	8.34
Average	78.66	81.19	49.60	81.56	37.90	55.64	3.82	68.66	6.69	67.71	8.04

As can be seen from table 8, adding noise to the training set causes a degradation in classification accuracy when using all instances of the training set but gives a better improvement in classification accuracy in almost all cases when RBF networks are trained using a reduced set generated by ENN and ALLKNN, particularly the classification accuracy was improved for 17 datasets from the overall 20 datasets in the case of ENN and 16 datasets from the overall 20 datasets in the case of ALLKNN. Using all instances in the training set to train the RBF network is vulnerable to overfitting. Hence, the misclassification occurs, especially in the presence of noise. In Many cases, removing noisy instances from the training set to obtain a clean set tends to improve classification accuracies more than the improvement that was achieved using noise free training sets. Since ENN and ALLKNN are noise filtering techniques, RBF networks trained using reduced sets generated by these reduction techniques have the highest classification accuracies compared to RBF networks trained using reduced set generated by other reduction techniques. The average classification accuracies for these networks were increased from 78.66% (using all instances of the training set) to 81.19% and 81.56% (using RBF-T-ENN and RBF-T-ALLKNN, respectively).

RBF-T-EXPLORE also have the worst classification accuracies, which is 55.64% on average. RBF-T-DROP2 and RBF-T-DROP5 have moderate classification accuracies equals to 68.66% and 67.71%, respectively, which is expected because both of these techniques have a noise filtering phase.
In all cases, the training time was considerably reduced compared to the time needed to train RBF-ALL, and here the reduction in time is even more than that when using noise-free training set, especially in the case of RBF-T-ENN and RBF-T-ALLKNN. This is because there is more reduction in storage requirement in the presence of 5% noise than in the case of noise free training sets. It is obvious from the table that ENN and ALLKNN achieve the best combination of classification accuracy and training time.

Table 9 shows the results obtained from the same experiments but with 10% noise added to the training set.

 COPYRIGHT © 2005 ENFORMATIKA

Table 9: Classification accuracy (Acc) and time (T%) for RBF networks trained using a reduced set of training set in presence of 10% noise.

Datasets	RBF-ALL	RBF-T-ENN		RBF-T-ALLKNN		RBF-T-Explore		RBF-T-DROP2		RBF-T-DROP5	
	Acc	Acc	T%	Acc	T%	Acc	Acc	T%	Acc	T%	Acc
australian	82.754	85.51	35.798	83.913	22.43	55.652	0.1767	75.362	1.0576	67.246	0.3215
bcw	96.05	96.78	68.69	96.49	39.85	67.94	0.25	94.58	0.46	90.04	0.19
breast	95.71	96.57	50.97	97.00	40.27	66.24	0.25	93.71	0.42	88.13	0.20
crx	47.27	54.55	77.34	67.27	100.00	40.00	15.76	43.64	68.97	50.91	76.85
echoc	80.30	87.12	72.19	88.64	59.41	81.82	7.22	82.58	12.54	85.61	20.00
glass	35.98	46.73	20.78	50.47	17.44	29.91	4.36	36.45	5.32	37.85	4.33
Heart	76.67	79.63	33.77	81.48	17.93	57.78	1.53	70.37	3.60	71.48	3.43
heart.cleve	70.96	83.17	15.35	82.18	9.07	59.08	0.64	68.98	1.05	65.35	0.85
heart.hung	74.15	83.33	30.80	80.95	19.45	68.37	1.47	67.69	1.97	64.63	1.78
heart.long	71.00	74.50	23.30	76.50	15.70	42.50	2.58	59.50	4.56	51.50	4.49
heart.swiss	82.93	91.06	64.91	89.43	55.01	93.50	5.17	54.47	15.18	21.14	15.51
horse-colic	60.80	72.43	32.74	71.76	17.01	36.21	1.15	51.50	2.70	54.82	2.57
ionosphere	92.88	92.88	32.51	93.45	23.53	35.90	0.66	65.81	1.02	69.80	0.66
iris	94.67	97.33	52.84	96.67	39.54	75.33	16.69	88.67	13.54	94.00	18.74
liver.bupa	64.93	64.93	21.36	66.38	7.67	49.86	0.42	63.19	4.87	63.77	3.06
pima	73.96	75.78	28.02	76.82	14.69	64.58	0.11	73.05	3.62	72.92	1.21
tic-tac-toe	77.56	80.48	41.67	78.08	28.16	37.68	0.06	54.59	2.96	59.92	1.04
vehicle	55.20	63.59	23.29	64.07	10.59	40.19	0.06	51.42	3.04	51.66	1.36
Voting	92.18	94.94	50.48	95.17	44.06	65.98	0.49	89.43	1.47	83.68	0.64
wine	75.84	92.14	55.80	91.57	47.17	44.94	5.75	72.47	7.98	76.40	9.42
Average	75.09	80.67	41.63	81.41	31.45	55.67	3.24	67.87	7.82	66.04	8.33

As can be seen from table 9, In the presence of 10% noise in the training sets, RBF-T-ENN improves the classification accuracies for 17 datasets from the 20 datasets and RBF-T-ALLKNN improves the classification accuracy for all the 20 datasets. The average classification accuracies for these networks were increased from 75.09% (using RBF-ALL) to 80.67% and 81.41% (using RBF-T-ENN and RBF-T-ALLKNN).

In all cases, the training time was considerably reduced compared to the time needed to train RBF-ALL, and here the reduction in time is even more than that when using 5% noisy training sets, especially in the case of RBF-ENN and RBF-ALLKNN. This is because there is more reduction in storage requirement in the presence of 10% noise. It is obvious from the table that ENN and ALLKNN achieve the best combination of classification accuracy and training time.

Table 10 shows the results obtained from the same experiments but with 15% noise added to the training set.

As can be seen from table 10, In the presence of 15% noise in the training sets, RBF-T-ENN improves the classification accuracies for 18 datasets from the 20 datasets and RBF-T-ALLKNN improves the classification accuracy for all the 20 datasets. The average classification accuracies for these networks were increased from 72.11% (using RBF-ALL) to 79.56% and 80.08% (using RBF-T-ENN and RBF-T-ALLKNN).

In all cases, the training time was considerably reduced compared to the time needed to train RBF-ALL, and here the reduction in time is even more than that when using 10% noisy training sets, especially in the case of RBF-ENN and RBF-ALLKNN. This is because there is more reduction in storage requirement in the presence of 15% noise. It is obvious from the table that ENN and ALLKNN achieve the best combination of classification accuracy and training time.

Tables 11(a) and 11(b) show the confidence intervals for the difference in classification accuracy between RBF-ALL and each of RBF-T-ENN and RBF_T_ALLKNN.

From these tables we can see that the classification accuracies of RBF-T-ALLKNN and RBF-T-ENN are comparable to that of RBF-ALL, i.e. the differences in classification accuracies are not statistically significant.

 COPYRIGHT © 2005 ENFORMATIKA

Table 10: Classification accuracy (Acc) and time (T%) for RBF networks trained using a reduced set of training set in presence of 15% noise.

Datasets	RBF-ALL	RBF-T-ENN		RBF-T-ALLKNN		RBF-T-Explore		RBF-T-DROP2		RBF-T-DROP5	
	Acc	Acc	T%	Acc	T%	Acc	Acc	T%	Acc	T%	Acc
australian	81.304	84.64	32.537	84.638	17.611	55.942	0.4604	76.667	1.4108	68.406	0.5051
bcw	94.729	96.05	45.49	96.05	32.26	65.30	0.28	92.09	0.84	89.17	0.36
breast	94.85	96.14	45.30	96.42	31.98	67.81	0.26	93.42	0.70	88.27	0.29
crx	49.091	47.27	92.31	50.91	78.28	43.64	35.75	43.64	78.73	36.36	70.59
echoc	83.333	84.85	76.57	87.12	59.40	81.82	2.04	80.30	17.17	85.61	21.25
glass	33.178	48.60	20.39	49.53	15.68	27.57	5.21	36.45	6.77	35.05	5.77
Heart	74.444	80.00	30.16	80.74	15.17	55.19	1.45	71.48	3.34	69.26	2.48
heart.cleve	62.376	84.16	13.73	82.18	7.38	57.10	0.52	67.66	1.27	73.60	1.20
heart.hung	57.823	78.91	26.98	79.59	14.33	67.01	1.70	68.37	2.25	60.54	1.88
heart.long	71.5	74.50	25.70	75.50	19.26	33.00	3.16	62.00	5.40	58.00	4.33
heart.swiss	77.236	90.24	76.92	87.81	71.36	93.50	2.62	70.73	20.46	27.64	23.40
horse-colic	56.811	71.43	27.13	72.09	12.29	33.22	0.74	57.81	3.03	56.15	2.41
ionosphere	92.877	94.30	30.75	94.87	18.78	37.04	0.73	68.66	1.20	70.09	0.89
iris	85.333	94.67	42.79	96.00	34.21	76.67	14.07	87.33	10.76	84.00	10.53
liver.bupa	63.478	61.45	20.03	64.35	7.79	45.80	0.72	60.87	5.33	57.97	3.68
pima	73.828	75.26	26.27	75.00	12.08	64.32	0.15	74.09	3.76	75.13	1.15
tic-tac-toe	72.756	77.45	36.42	78.60	20.81	37.79	0.06	56.16	3.69	46.45	1.51
vehicle	49.645	62.29	16.18	61.47	8.19	43.26	0.08	46.57	2.83	52.36	1.38
Voting	91.264	96.32	40.98	96.55	39.05	63.22	0.57	88.05	1.14	82.76	0.64
wine	76.404	92.70	52.89	92.14	39.52	43.82	6.78	71.35	8.93	76.40	11.12
Average	72.11	79.56	38.98	80.08	27.77	54.65	3.87	68.68	8.95	64.66	8.27

Table 11 (a): 90% Confidence Intervals for the difference in classification accuracy between RBF-ALL and each of RBF-T-ENN and RBF-T-ALLKNN in the case of noise free datasets and in the presence of 5% noise

Datasets	Noise free				5% noise			
	RBF-T-ENN		RBF-T-ALLKNN		RBF-T-ENN		RBF-T-ALLKNN	
	low	high	low	high	low	high	low	high
australian	-4.23	8.287	-3.58	8.503	-3.61	9.985	-2.31	7.82
bcw	-3.34	3.64	-2.78	3.667	-1.53	3.581	-2.72	4.19
breast-cancer-wisconsin	-2.55	1.404	-1.57	1.57	-2.71	2.135	-2.21	2.209
crx	-19.3	29.94	-26.6	38.58	-42.9	49.58	-31.1	59.06
echocardiogram	-10.7	6.162	-13.7	3.159	-15.2	10.85	-11.7	10.23
glass	-8.15	23.22	-10.6	23.74	-9.19	25.16	-0.56	21.16
Heart	-3.93	9.859	-7.78	12.22	-8.17	11.87	-9.7	8.223
heart.cleveland.2	-8.44	13.75	-6.58	11.2	-12.3	18.34	-13.1	17.77
heart.hungarian.2	-7.49	12.31	-5.88	9.301	-9.22	15.4	-8.78	14.3
heart.long-beach-va.2	-3.91	9.915	-5.31	13.31	-3.65	12.65	-8.7	16.7
heart.swiss.2	-4.1	9.102	-4.1	7.429	-7.46	28.87	-7.46	28.87
horse-colic	-6.82	17.45	-9.81	16.44	-13.8	27.04	-12.7	23.98
ionosphere	-3.8	3.816	-3.71	4.867	-2.97	4.7	-2.13	4.419
iris	-4.12	2.791	-4.12	2.791	-7.4	8.734	-7.29	7.289
liver.bupa	-16.6	11.92	-12	7.363	-16.4	8.233	-15.1	9.335
pima-indians-diabetes	-5.51	5.509	-3.4	4.705	-3.96	5.526	-5.25	5.78
tic-tac-toe	-21.1	0.031	-19.7	-10.8	-5.95	10.11	-4.52	3.256
vehicle	-5.99	2.208	-9.25	8.039	0.373	8.131	0.331	7.935
Voting	-2.56	3.021	-2.56	3.021	-2.73	4.584	-2.9	4.283
wine	-3.44	2.326	-3.44	2.326	-9.11	12.51	-6.2	11.89
Average	-7.3	8.8	-7.8	8.6	-8.9	13.9	-7.7	13.4

 COPYRIGHT © 2005 ENFORMATIKA

Table 11 (b): 90% Confidence Intervals for the difference in classification accuracy between RBF-ALL and each of RBF-T-ENN and RBF-T-ALLKNN in the presence of 10% and 15% noise

Datasets	10% Noise				15% Noise			
	RBF-T-ENN		RNF-T-ALLKNN		RBF-T-ENN		RNF-T-ALLKNN	
	low	high	low	high	low	high	low	high
australian	-3.73	9.23	-3.15	5.47	-1.68	8.35	-1.56	8.22
bcw	-2.09	3.56	-1.82	2.70	-3.00	5.63	-2.69	5.33
breast-cancer-wisconsin	-2.11	3.82	-1.52	4.09	-2.61	5.18	-3.04	6.19
crx	-25.14	39.81	-12.67	52.00	-27.51	24.85	-42.41	48.41
echocardiogram	-8.80	22.76	-9.62	26.65	-12.77	15.74	-9.79	17.37
glass	-11.65	33.03	-0.39	29.40	-0.26	30.95	-4.16	36.67
Heart	-6.00	11.93	-4.70	14.33	-3.15	14.26	-7.45	20.04
heart.cleveland.2	-6.65	31.25	-3.81	26.39	2.11	41.59	-0.96	40.70
heart.hungarian.2	-4.10	22.63	-9.73	23.50	-8.08	50.57	-5.88	49.67
heart.long-beach-va.2	-9.35	16.35	-7.58	18.58	-8.72	14.72	-8.70	16.70
heart.swiss.2	-4.45	20.86	-7.05	20.26	-3.84	29.87	-2.07	23.22
horse-colic	-2.46	25.67	-3.35	25.25	-1.47	30.68	1.95	28.57
ionosphere	-6.93	6.96	-5.96	7.12	-4.89	7.77	-3.86	7.86
iris	-6.55	11.89	-7.00	11.00	-5.43	24.09	-6.58	27.92
liver.bupa	-13.81	13.83	-12.79	15.73	-12.26	8.24	-12.14	13.87
pima-indians-diabetes	-4.09	7.73	-4.08	9.81	-2.13	4.99	-3.28	5.61
tic-tac-toe	-4.68	10.53	-4.96	6.01	-6.64	16.04	-5.55	17.24
vehicle	-2.29	19.06	-2.11	19.80	1.31	24.02	2.35	21.29
Voting	-7.98	13.49	-7.78	13.77	-5.06	15.19	-5.17	15.75
wine	2.98	29.64	-0.54	32.04	0.69	32.12	-0.10	31.80
Avgerage	-6.5	17.7	-5.5	18.2	-5.3	20.2	-6.1	22.1

5. CONCLUSION

In this work, several techniques for locating the centers of Radial Basis Function Networks were examined. These techniques are ENN, ALLKNN, EXPLORE, DROP2 and DROP5 which represent a sample of instance reduction techniques. This sample of techniques was chosen carefully to represent the wide spectrum of techniques.

The performance, in terms of classification accuracy and training time, for RBF networks trained using these reduction techniques was compared with two extremes: RBF-ALL and PNN. The former achieves high classification accuracies and the latter requires smaller training time. Results showed that RBF networks trained using sets of centers located by noise-filtering techniques (ALLKNN and ENN) rather than pure reduction techniques produce the best results in terms of classification accuracy.

Our results also show that using noise filtering techniques to determine the training set as well as the center points substantially improves the classification accuracy.

REFERENCES

[1] Bors, A.G. (2001). Introduction of the Radial Basis Function (RBF) Networks. *Online Symposium for Electronics Engineers*, 1(1), 1 - 7.

[2] Cameron-Jones, R.M. (1995). Instance Selection by Encoding Length Heuristic with Random Mutation Hill Climbing. Proceedings of the *Eighth Australian Joint Conference on Artificial Intelligence*, 99-106.

[3] Feitosa, R.Q., Vellasco, M.M., Oliveira, D.T., Andrade, D.V. and Maffra, S.A. (2002). Facial Expression Classification Using RBF and Back-Propagation Neural Networks, Proceedings of the 4th World Multiconference on Systemics, Cybernetics and Informatics (*SCI'2000*) and the 6th International Conference on Information Systems Analysis and Synthesis (*ISAS'2000*), Orlando, USA, August 2000, 73-77.

[4] Ghosh, J. and Nag, A., (2002). Knowledge enhancement and reuse with radial basis function networks, IJCNN '02. Proceedings of the 2002 *International Joint Conference on Neural Networks*, Vol. 2, 1322 -1327.

[5] Hwang Y.S. and Bang S.Y. (1997). An Efficient Method to Construct a Radial Basis Function Neural Network Classifier. *Neural Networks*, 10(8), 1495-1503.

[6] Karayiannis, N.B. and Mi, G.W. (1997). Growing radial basis neural networks: merging supervised and unsupervised learning with network growth techniques, *IEEE Transactions on Neural Networks*, 8(6),1492-1506.

[7] Kubat, M. (1998). Decision Trees Can Initialize Radial-Basis-Function Networks. *IEEE Transactions on Neural Networks*, 1(9), 813-821.

[8] LeCun, Y., Jackel, L.D., Bottou, L., Brunot, A., Cortes, C., Denker, J.S., Drucker, H., Guyon, I., Muller, U.A., Sackinger, E., Simard, P. and Vapnik, V. (1995). Comparison of learning algorithms for handwritten digit recognition, Proceedings of *International Conference on Artificial Neural Networks*, Paris, 53-60.

[9] Low, R. and Togneri, R. (1998). Speech Recognition Using the Probabilistic Neural Network, Proceedings of *ICSLP98*, Sydney, Australia, Paper No. 645.

 COPYRIGHT © 2005 ENFORMATIKA

[10] Mak, M.W. and Cho, K.W. (1998). Genetic Evolution of Radial Basis Function Centers for Pattern Classification, Proceedings of International Conference on Neural Networks (*ICNN'98*), USA, 669-673.
[11] Mitchell, T. (1997). *Machine Learning*. New York: McGraw-Hill.
[12] Orr, M.J. (1996). Introduction to radial basis function networks. Technical report, Institute for Adaptive Neural Computation, Division of Informatics, University of Edinburgh. Retrieved April, 2004, from: *http:www.anc.ed.ac.uk/mjo/intro/intro.html*.
[13] Panchapakesan, C., Palaniswami, M., Ralph, D. and Manzie, C. (2002). Effects of moving the centers in an RBF network. *IEEE Transactions on Neural Networks*, 13(6), 1299-1307.
[14] Rüdiger, W. B., (2001). Medical Analysis and Diagnosis by Neural Networks. *ISMDA*, Spain, 1-13.
[15] Sahin, F. (1997). *A Radial Basis Function Approach to a Color Image Classification Problem in a Real Time Industrial Application*, Master's thesis, Virginia polytechnic institute, Blacksburg.
[16] Tomek, I. (1976). An Experiment with the Edited Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(6), 448-452.
[17] Wasserman, P.D. (1993). *Advanced Methods in Neural Computing*, New York: Van Nostrand Reinhold, 35-55.
[18] Wettscherick, D. and Dietterich, T., (1992). Improving the performance of radial basis function networks by learning center locations, *Advances in Neural Information Processing Systems*. 4 (1), 1133-1140.
[19] Wilson, D. and Martinez, T., (2000). Reduction Techniques for Instance-Based Learning Algorithms, *Machine Learning*, 38(3), 257-286.
[20] Wlison, D.L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, 2(3), 408-421.

 COPYRIGHT © 2005 ENFORMATIKA

Novelty as a Measure of Interestingness in Knowledge Discovery

Vasudha Bhatnagar, Ahmed Sultan Al-Hegami, and Naveen Kumar

***Abstract*—** Rule Discovery is an important technique for mining knowledge from large databases. Use of objective measures for discovering interesting rules leads to another data mining problem, although of reduced complexity. Data mining researchers have studied subjective measures of interestingness to reduce the volume of discovered rules to ultimately improve the overall efficiency of KDD process.

In this paper we study *novelty* of the discovered rules as a subjective measure of interestingness. We propose a hybrid approach based on both objective and subjective measures to quantify *novelty* of the discovered rules in terms of their deviations from the known rules (knowledge). We analyze the types of deviation that can arise between two rules and categorize the discovered rules according to the user specified threshold. We implement the proposed framework and experiment with some public datasets. The experimental results are promising.

***Keywords*—** Knowledge Discovery in Databases (KDD), Interestingness, Subjective Measures, Novelty Index.

I. INTRODUCTION

THE vast search space of hidden patterns in the massive databases is a challenge for the KDD community. For example, in a database with n distinct items, the number of potential frequent item sets is exponential in n. In a database with n records, the potential number of clusters is $\frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}\binom{k}{i}(i)^n$ [6]. However, a vast majority of these patterns are pruned by the score functions engaged in the mining algorithm. To avoid computing the score function for the entire search space, optimization strategies are used. For example, in association rule mining, confidence is the commonly used score function and the anti-monotonic property of frequent itemsets is the optimization strategy.

Despite massive reduction of search space by employing suitable score function and optimization strategies, all of the discovered patterns are not useful for the users. Consequently, researchers have been strongly motivated to further restrict the search space, by supplying constraints to the data mining algorithm [5] and providing good measures of interestingness [9,18].

Constraints based mining techniques allow the users to specify the rules to be discovered according to their background knowledge, thereby making the KDD process more effective [5,10]. A complicated mining query can be used to express constraints specified by the user in order to make the mining process more efficient [22,23,24,25,26].

There are two types of interestingness measures that have been studied in data mining literature viz. Objective and Subjective measures. Objective measures are based on the statistical significance (certainty, coverage, etc.) or structure (simplicity) of the patterns [7,10]. Subjective measures are based on the end-user who evaluates the patterns on the basis of novelty, actionability and unexpectedness, etc. [2,3,4,9,12,13].

In real life KDD endeavors, it is often required to compare the rules mined from datasets generated under different contexts (for example, at different points in time or in two different locations). Unless the underlying data generation process has changed dramatically, it is expected that the rules discovered from one set are likely to be similar (in varying degrees) to those discovered from another set. Some of the discovered rules may be identical to the known rules, some may be generalization/ specialization of the known rules and some others may be same or different with varying degrees of sameness/difference.

As the number of rules discovered by data mining algorithms becomes huge, the time consumed and the space required for maintaining and understanding these rules becomes vast. *Novelty* of a rule can be used as an effective way to filter the rule set discovered from the target data set thereby, reducing the volume of the output.

Novelty is defined as the extent to which the discovered rules contribute to new knowledge [1,2,3]. In this paper we focus on the quantification of *novelty* and use this measure for categorization of discovered rules. Though *novelty* is a subjective measure, we propose a strategy to quantify objectively the *novelty index* of each discovered rule, and facilitate categorization of rules based on the degree of novelty desired by the user. Asking the user to specify a threshold to filter rules of desired degree of novelty, captures user subjectivity.

Vasudha Bhatnagar is lecturer at the Department of Computer Science, University of Delhi, Delhi, INDIA; (e-mail: vbhatnagar@cs.du.ac.in).

Ahmed Sultan Al-Hegami is Ph.D scholar in the Department of Computer Science, University of Delhi, Delhi, INDIA; (e-mail: ahmed_s_gamil@yahoo.com)

Naveen Kumar is reader at the Department of Computer Science, University of Delhi, Delhi, INDIA; (e-mail: nk@cs.du.ac.in).

 COPYRIGHT © 2005 ENFORMATIKA

II. Related Work

There are many proposals that have studied the novelty in disciplines such as robotics, machine learning and statistical outliers detection [14,15,16,17]. Generally, these methods build a model of training set that is selected to contain no examples of the important (i.e. novel) class [11]. Subsequently, the mechanisms detect the deviation from this model by some way. For instance, Kohonen and Oja proposed a novelty filter, which is based on computing the bit-wise difference between the current input and the closest match in the training set [11]. In [21], a sample application of applying association rule learning is presented. By monitoring the variance of the confidence of particular rules inferred from the association rule learning on training data, it provides information on the difference of such parameters before and after the testing data entering the system. Hence, with some pre-defined threshold, abnormalities can be fairly detected.

The techniques that have been proposed in statistical literature are focused on modeling the support of the dataset and then detecting inputs that don't belong to that support. The choice of whether to use statistical methods or machine learning methods is based on the data, the application, and the domain knowledge [14].

To our knowledge no concrete work has been conducted to tackle the novelty measure in data mining. The work proposed in [8] detects the novelty of rules mined from text [8]. In this work, the novelty is estimated based on the lexical knowledge in WordNet. The proposed approach defines a measure of semantic distance between two words in WordNet and determined by the length of the shortest path between the two words (w_i,w_j). The novelty is defined as the average of this distance across all pairs of the words (w_i,w_j), where w_i is a word in the antecedent and w_j is a word in the consequent.

In [2], a framework has been proposed to quantify novelty in terms of computing the deviation of currently discovered knowledge with respect to domain knowledge and previously discovered knowledge. The approach presented in [2] is intuitive in nature and lays more emphasis on user involvement in quantification process by way of parameter specification. In the present work, the quantification is performed objectively and user involvement is sought for categorization of rules based on *novelty index*.

III. Novelty Index

Let D_i denote the database extension at time t_i, and K_i denote the knowledge discovered from D_i. Figure 1 shows the knowledge discovered at two time instances. Major volume of K_{i+1} would be the overlapping region that represents previously discovered knowledge. The shaded portion denotes the novel knowledge. Thus the rules falling in the shaded area are assigned high degree of *novelty* compared to those in the overlapping regions.

The proposed framework assigns a *novelty index* to each discovered rule that indicates its proximity/deviation from some existing rule in the rule base of previously discovered knowledge.

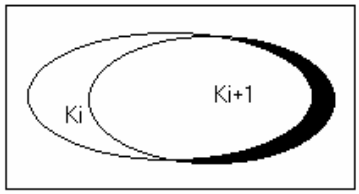

Figure 1. Regions of Discovered Rules

Novelty index of a rule is the deviation with respect to a given rule set. It is a pair (Ă, Č) that indicates the deviation of the antecedent and consequent of the rule with those of the closest rule in the previously discovered knowledge. To compute the *novelty index* of a rule, the deviation is measured for the antecedent and the consequent at conjunct level and subsequently the conjunct level deviation is combined to compute rule level deviation.

A. Definitions and Notations

A rule *R* has the form: $A \rightarrow C$ where A denotes an antecedent and C denotes a consequent. Both A and C are in CNF ($c_1 \Lambda c_2 \Lambda \ldots\ldots \Lambda c_k$). The conjunct c_j is of the form <Å, O, V>. Where Å is an attribute, *Dom* (Å) is the domain of Å, and V $\in$ *Dom* (Å), O $\in$ {=,<,>,≥,≤}. Without loss of generality, we consider both A and C as sets of conjuncts for further processing.

B. Deviation at Conjunct Level

In order to quantify deviation between any two conjuncts, the attributes, operators, and attribute values of the two conjuncts in question need to be taken into account.

Definition 3.1 Two conjuncts c_i and c_j <$Å_iO_iV_i$> and <$Å_jO_jV_j$> respectively) are compatible if and only if $Å_i = Å_j$. Otherwise, we consider c_i and c_j as non-compatibles.

Definition 3.2 Let c_i and c_j be two non-compatible conjuncts. The deviation δ (c_i,c_j) between them is defined to be 1.

We capture the following four types of deviations between two compatible conjuncts.

- Z-deviation: This type signifies identical conjuncts and is quantified by numeric 0.
- V-deviation**:** This type of deviation signifies the magnitude of change in the value of the attribute in two conjuncts. In order to normalize, we quantify this type of deviation as the ratio of the change to the range of the attribute value.

  This method of computation of V-deviation is suitable for only numeric and ordinal attributes. In case of nominal attributes, the change in value can be quantified in terms of probabilities. Since ordinal domains generally have small and manageable cardinality, prior domain knowledge can be used to assign probabilities to domain values. In case it is not feasible to assign probabilities in the above-mentioned way (e.g. color of car), the dataset itself can be used to

 COPYRIGHT © 2005 ENFORMATIKA

compute probabilities corresponding to each domain value.

- C-deviation**:** This type of deviation signifies the deviation in the conditional operators in the two conjuncts. In order to quantify C-deviation, we take into account the type of change in the condition. The operators are formatted on a number line as shown in Figure 2. The deviation between the operators is quantified by the distance between the operators on the numberline.

  We define a function *opdist* (O_1, O_2) → {0,1,2,3,4}, which denotes the distance between the two distinct operators (O_1, O_2) on the numberline. We define four possible values of deviations (0, 1/5, 2/5, 3/5, 4/5) between any two operators, ranking the extent of deviation between condition operators in two conjuncts.

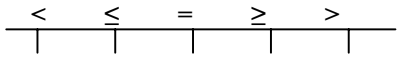

Figure 2. Operators on Numberline

- CV-deviation**:** Quantifies V-deviation in presence of C-deviation. It captures the co-occurrence of change in both conditions and attribute values in two conjuncts.

  We compute the C-deviation (*c*) and V-deviation (*v*) independently of each other, in the two given conjuncts. The user defines a real valued function $f(c,v) \rightarrow [0,1]$ to combine the two types of deviations. Depending on the importance of the type of deviations for a specific application in a domain, different functions can be used for computing deviations on different attributes.

  Typically, $f(c,v)$ is of the form w_1c+w_2v, where $w_1+w_2=1$.

Note that, the computation of deviation between two conjuncts is objective in all types of deviation, except CV-deviation, where the user subjectivity is captured. Depending on the importance of either C or V deviation, the user assigns appropriate weights w_1 and w_2.

The following definition formalizes the quantification of Conjunct level deviation.

Definition 3.3 Let c_1 and c_2 be two compatible conjuncts $<A_1O_1V_1>$ and $<A_2O_2V_2>$ respectively). The deviation of c_1 with respect to c_2 is defined as follows:

$$\delta(c_1,c_2)=\begin{cases} 0 & \text{if } O_1=O_2 \text{ and } V_1=V_2 \quad \text{Z-deviation.} \\ \dfrac{|V_1-V_2|}{Range(Dom(A_1))} & \text{if } O_1=O_2 \text{ and } V_1\neq V_2 \quad \text{V-deviation} \\ opdist(O_1,O_2)/5 & \text{if } O_1\neq O_2 \text{ and } V_1=V_2 \quad \text{C-deviation.} \\ f(c,v) & \text{if } O_1\neq O_2 \text{ and } V_1\neq V_2 \quad \text{CV-deviation.} \end{cases}$$

Lemma 3.1. The conjunct level deviation lies between [0,1]
Proof 3.1. By definitions 3.2 and 3.3.

It is easy to see that:

i) $\delta(c_i,c_j) \geq 0$,
ii) $\delta(c_i,c_i) = 0$,
iii) $\delta(c_i,c_j) = \delta(c_j,c_i)$.

However, $\delta(c_i,c_j)$ does not satisfy triangular inequality in case of CV-deviation, where we capture user subjectivity.

C. Conjunct Set Deviation

In order to compute novelty index of a rule, it is necessary to define the deviation Ψ between two conjunct sets, since both antecedents and consequents are considered to be sets of conjuncts. The deviation $\Psi(S_1,S_2)$ between two conjuncts sets is quantified based on the analysis of the possible types of differences between two sets of conjuncts S_1 and S_2. Without loss of generalization, we assume that an attribute occurs at most once in a conjunct set S. Computation of deviation at this level is based on counting incompatible conjuncts among the two sets and quantifying total deviation among the compatible conjuncts. Intuitively, it is the number of incompatible conjuncts that contribute most towards the value of the deviation. While comparing two sets of conjuncts namely S_1 and S_2, three possibilities arise.

1 S_1 and S_2 are identical.
2 S_1 is a generalization / specialization of S_2.
3 S_1 and S_2 are different.

In case 1, the deviation must be nil, while in case 2, it is desirable to quantify the degree of generalization / specialization. In case 3, the degree of novelty is decided on the basis of deviation between the two sets needs to be quantified.

We compute the deviation between two conjunct sets as follows.

Definition 3.4 Let S_1 and S_2 be two conjunct sets with cardinalities $|S_1|$ and $|S_2|$ respectively. Let *k* be the pairs of compatible conjuncts between S_1 and S_2. The deviation between S_1 and S_2 is computed as:

$$\Psi(S_1,S_2)= \frac{\{|S_1|+|S_2|-2*k\}+\sum_{i=1}^{k}\delta(c_1^i,c_2^i)}{|S_1|+|S_2|}$$

where (c_1^i,c_2^i) is the i^{th} pair of compatible conjuncts.

Theorem 3.1 For any two conjunct sets S_1 and S_2,
$0 \leq \Psi(S_1,S_2) \leq 1$.

Proof 3.1 The proof follows by simple reasoning. We consider two extreme cases where there are no compatible conjuncts and another with all equal conjuncts.

In case there are no compatible conjuncts, $k=0$ and the second component of the numerator vanishes. With all non-compatible conjuncts, $\Psi(S_1,S_2)=1$. In case the two conjunct sets are equal, $k=\frac{|S_1|+|S_2|}{2}$ and the second component in the numerator reduces to zero. Thus $\Psi(S_1,S_2) = 0$, which captures Z-deviation.

Note that

 COPYRIGHT © 2005 ENFORMATIKA

i) $\Psi(S_1,S_2) \geq 0$,
ii) $\Psi(S_1,S_1) = 0$, and
iii) $\Psi(S_1,S_2) = \Psi(S_2,S_1)$.

We do not expect Ψ to satisfy triangular inequality in view of its violation by underlying conjunct level deviation function. Therefore, Ψ can't be used as a distance metric.

Example 1
Given two conjunct sets S_1 and S_2 as follows:
$S_1 = \{C_1, C_2, C_3, C_4, C_5\}$
$S_2 = \{C_1, C_2\}$

The deviation of S_2 with respect to S_1 is computed according to definition 3.4 as follows:

$$\Psi(S_1,S_2) = \frac{|5| + |2| - 2*2 + 0}{|5| + |2|} = \frac{3}{7} = 0.42$$

Note that S_1 is specialization of S_2, and hence, the deviation of S_1 with respect to S_2 also indicates the degree of specialization. Similarly, $\Psi(S_2,S_1)$=0.42 also indicates the degree of generalization of S_2 with respect to S_1.

Although the deviation between two conjunct sets in case of specialization / generalization can be interpreted as degree of specialization / generalization as shown in Example 1, the interpretation is not natural. Intuitively, the degree of specialization (generalization) of S_1 (S_2) with respect to S_2 (S_1) should reflect the ratio of extra (i.e. newly added) conjuncts to the total number of conjuncts in the specialized (generalized) rule.
For this reason, we propose an alternative approach which computes the deviation of specialization / generalization in the next section. However, use of Definition 3.4 for computation of degree of imparts uniformity and mathematical elegance.

D. Generalization and Specialization

Before computing $\Psi(S_1,S_2)$ for two conjunct sets S_1 and S_2, it is necessary to check for generalization or specialization between them. If one of the sets is completely subsumed by the other, then this is a case for *GS* deviation. In this case it is necessary to compute the degree of generalization or specialization to enable the end-user judge the degree of novelty.

Definition 3.5 Let S_1 and S_2 be two conjunct sets such that S_1 is completely subsumed by the S_2. Then S_1 is generalization of S_2 and the deviation (degree of generalization) is given by:

$$\Psi(S_1,S_2) = \frac{|S_2| - |C|}{|S_2|}$$

where C is the set of subsumed conjuncts.

Definition 3.6 Let S_1 and S_2 be two conjunct sets such that S_2 is completely subsumed by S_1. Then S_1 is specialization of S_2 and the deviation (degree of specialization) is given by:

$$\Psi(S_1,S_2) = \frac{|S_1| - |C|}{|S_1|}$$

where C is the set of subsumed conjuncts.

Note that high magnitude of generalization / specialization indicates high degree of novelty.

Example 2
Given two conjunct sets S_1 and S_2 as follows:
$S_1=\{C_1, C_2, C_3, C_4, C_5\}$
$S_2=\{C_1, C_2\}$

The deviation (degree of *specialization*) of S_1 with respect to S_2 is computed according to Definition 3.6 as follows:

$$\Psi(S_1,S_2) = \frac{|5| - |2|}{5} = \frac{3}{5} = 0.6$$

Similarly, in case of deviation (degree of generalization) of S_1 with respect to S_2 is computed according to Definition 3.5 is 0.6.
Note that the computations in Example 2 are closer to the intuitive notion of the deviation compared to those of Example 1.

IV. COMPUTING NOVELTY INDEX

Novelty index of a rule *r* is defined with respect to a given rule set *R*. It is computed as paired deviation of antecedent and consequent of *r* relative to the closest rule in *R*. The rule $s \in R$, from whose antecedent the deviation of *r* is minimum is considered to be closest. The novelty index is defined as follows.

Definition 3.7 Let r: $A_r \rightarrow C_r$ be a rule whose *novelty index* is to be computed with respect to the rule set R. Then

$N_r^R = (\min_{s \in R} (\Psi(A_r,A_s)), \Psi(C_r,C_s))$ gives the *novelty index.*

Having computed the *novelty index* for all the rules in the currently discovered rule set with respect to previously discovered rule set, the task of rule reduction can be performed in several ways. Some of the suggested ways are:

i) select the top *K* novel rules,
ii) select rules with novelty index exceeding a threshold,
iii) categorize the indexed rules as per Table I.

V. ALGORITHM

We give below the algorithm for categorizing rules in a give rule set R_{CDK} (currently discovered knowledge) with respect to the rule base R_{PDK} (previously discovered knowledge)
Input : Two rulesets R_{PDK} & R_{CDK}
Output: Updated R_{PDK} and a Tag for each rule in R_{CDK}
Process:
For each R_i ($A_i \rightarrow C_i$) in R_{CDK}
Find R_j ($A_j \rightarrow C$)$_j$ from R_{PDK}, such that $\Psi(A_i,A_j)$ is minimum.
Compute $\Psi(C_i,C_j)$
Categorize R_i as per Table I.
If category is **Novel,** add to R_{PDK}.

 COPYRIGHT © 2005 ENFORMATIKA

TABLE I
CATEGORIZATION OF DISCOVERED RULES

Categorization	Semantics	Condition
Conforming Rules	Rules that have been discovered earlier.	$\check{A} \leq \Phi$ & $\check{C} \leq \Phi$
Generalized (Specialized) Rules	Rules that are generalization (specialization) of some earlier discovered rules.	$A_r(A_s)$ subsumes $A_s(A_r)$ & $\check{C}=0$
Unexpected Rules	Rules that are unexpectedly different from the previously discovered rules.	$\check{A} \leq \Phi$ & $\check{C} \geq \Phi$ OR $\check{A} \geq \Phi$ & $\check{C} \leq \Phi$
Novel Rules	Rules that add to knowledge. Such rules do not fall into any of the earlier specified categories.	$\check{A} \geq \Phi$ & $\check{C} \geq \Phi$

Where ($\check{A} = \Psi(A_i,A_j)$ & $\check{C} = \Psi(C_i,C_j)$) & Φ is a user specified threshold).

VI. EXAMPLE

For better understanding of our framework, we present an example from the 'supmart' dataset available in CBA [20]. The following set of rules was discovered by CBA at time T_1, and we designate this as currently discovered knowledge.
We consider the previously discovered knowledge to be null at T_1.

R_1: potato chips = Y ^ ketchup = Y → beer = Y ^ orange juice = Y
R_2: orange juice = Y ^ ketchup = Y → beer = Y ^ potato chips = Y
R_3: sugar = Y ^ potato chips = Y ^ orange juice = Y → ketchup = Y
R_4: sugar = Y ^ potato chips = Y ^ ketchup = Y → orange juice = Y
R_5: sugar = Y ^ orange juice = Y ^ ketchup = Y → potato chips = Y
R_6:potato chips = Y ^ orange juice = Y ^ ketchup = Y → sugar = Y
R_7:sugar = Y ^ potato chips = Y → orange juice = Y ^ ketchup = Y
R_8:sugar = Y ^ orange juice = Y →potato chips = Y ^ ketchup = Y
R_9:potato chips = Y ^ orange juice = Y → sugar = Y ^ ketchup = Y
R_{10}: beer = Y ^ orange juice = Y ^ ketchup = Y → sugar = Y

If the user specifies $w_1 = 0.4$, $w_2 = 0.6$ for CV-deviation and $\Phi = 0.5$, the *novelty index* assigned to rules discovered at time T_1 given bellow in the format:
Rule: {Closest Rule, [$\check{A}$, $\check{C}$], Category}.

R_1: {-, [1,1], Novel}, R_2: {R_1, [0.5,0.5], Conformed}
R_3: {R_1, [0.6,1], Novel}, R_4: {R_1, [0.2,0.3], Conformed}
R_5: {R_3, [0.3,1], Unexpected}, R_6: {R_1, [0.2,1], Unexpected}
R_7: {R_3, [0.2,0.3],Conformed},R_8:{R_3, [0.2,0.3], Conformed}
R_9: {R_3,[0.2,0.3],Conformed},R_{10}: {R_1, [0.6,1], Novel}
Now, previously discovered knowledge consists of ruleset R_{PDK}={ R_1, R_3, R_{10}}

The following rules are discovered latter at time T_2 (R_{CDK}):

R_{11}: orange juice = Y ^ ketchup = Y → sugar = Y
R_{12}: ketchup = Y → sugar = Y ^ orange juice = Y
R_{13}: ketchup = Y → tomato sauce = Y
R_{14}: potato chips = Y ^ tomato sauce = Y → ketchup = Y
R_{15}: potato chips = Y ^ ketchup = Y → tomato sauce = Y
R_{16}: tomato sauce = Y ^ ketchup = Y → potato chips = Y
R_{17}: potato chips = Y → tomato sauce = Y ^ ketchup = Y
R_{18}: ketchup = Y → potato chips = Y ^ tomato sauce = Y
R_{19}: beer = Y ^ potato chips = Y ^ tomato sauce = Y→ketchup = Y
R_{20}: beer = Y ^ potato chips =Y ^ ketchup = Y→ tomato sauce = Y

Categorizing this ruleset with respect to R, we get:
R_{11}:{R_{10}, [0.3,0],Generalized},R_{12}:{R_1,[0.3,0.5],Conformed}
R_{13}: {R_1, [0.3,1], Unexpected}, R_{14}:{R_1, [0.5,1],Unexpected}
R_{15}: {R_1, [0,1], Unexpected}, R_{16}: {R_1, [0.5,1], Unexpected}
R_{17}: {R_1, [0.3,1], Unexpected},R_{18}: {R_1, [0.3,1],Unexpected}
R_{19}: {R_1, [0.6,1], Novel}, R_{20}: {R_1, [0.2,1], Unexpected}

R_{PDK} is now updated to {R_1, R_3, R_{10}, R_{19}}

VII. EXPERIMENTAL STUDY

The proposed approach is implemented in C language and tested using public datasets available in [19]. Since, there are no other approaches available, which objectively quantify novelty and yet take user subjectivity into account; we could not perform any comparison against our approach. The following experiments were conducted to show the effectiveness of the framework:

A. Experiment One

We worked with five public datasets available at [19]. We considered each of these datasets as evolving with time, and partitioned them into 3 increments: D_1, D_2 and D_3 mined at times T_1, T_2 and T_3 respectively. We took each of these partitions to be equal for purpose of generating rules.

The datasets were mined, using CBA [20], with 0.1% and 1% as minimum confidence and support respectively, uniformly for all datasets. The low thresholds enable generation of large number of rules; thereby demonstrating the efficiency of the framework. The discovered rules were categorized as in Table I, with Φ=0.5 and $f(c,v) \rightarrow$ [0.4,0.6] for CV-deviation. Table II summarizes the result.

TABLE II
DISCOVERED RULES AT TIME T_1, T_2 AND T_3 FOR DIFFERENT DATASETS WITH Φ=0.5

Dataset	Time	Instances	Discovered rules	Novel	Unexpected	Specialized	Generalized	Conformed
Census	T_1	12000	942	29	239	4	19	652
	T_2	12000	1061	6	189	20	21	825
	T_3	8561	636	3	58	8	7	560
Supmart	T_1	40	2775	25	1576	62	75	1025
	T_2	40	1875	0	1026	49	103	661
	T_3	48	1570	0	717	40	116	697
German	T_1	333	117	13	66	0	0	38
	T_2	333	118	9	43	0	0	66
	T_3	334	133	4	56	3	1	69
Sick	T_1	933	29	4	18	0	1	6
	T_2	933	33	2	17	7	0	7
	T_3	934	32	2	16	5	1	8
Heart	T_1	90	38	7	5	24	0	2
	T_2	90	95	6	6	71	2	10
	T_3	90	40	2	4	19	2	13

 COPYRIGHT © 2005 ENFORMATIKA

B. Experiment Two

The second experiment was performed using 'census' dataset to study the effect of novelty threshold Φ on the number of rules of different categories. This dataset contains 48842 instances, mix of continuous and discrete attributes, and 2 class values. With same partitions (12000,12000,8561) and support and confidence thresholds as in the previous experiment. The number of rules varied as per our expectation. The result is shown in Table III.

TABLE III
DISCOVERED RULES AT TIME T_1, T_2 AND T_3 FOR DIFFERENT (Φ)

Novelty Degree (Φ)	Time	Discovered rules	Novel	Unexpected	Specialized	Generalized	Conformed
Φ=0.9	T_1	942	4	318	0	2	618
	T_2	1061	0	451	5	2	603
	T_3	636	0	241	6	1	388
Φ=0.8	T_1	942	6	241	1	4	690
	T_2	1061	1	235	1	2	822
	T_3	636	0	130	4	1	501
Φ=0.7	T_1	942	10	325	1	5	601
	T_2	1061	2	314	4	6	735
	T_3	636	0	164	5	2	465
Φ=0.6	T_1	942	16	227	15	11	673
	T_2	1061	7	135	16	16	887
	T_3	636	1	79	16	9	531
Φ=0.5	T_1	942	29	239	4	19	652
	T_2	1061	6	189	20	21	825
	T_3	636	3	58	8	7	560
Φ=0.4	T_1	942	49	439	7	27	420
	T_2	1061	20	306	36	36	663
	T_3	636	8	114	29	15	470
Φ=0.3	T_1	942	62	593	5	27	255
	T_2	1061	25	629	39	34	334
	T_3	636	15	268	25	24	304

VIII. Conclusion and Future Work

In this paper, we proposed a strategy for rule set reduction based on the *Novelty index* of the rule. *Novelty index* of a newly discovered rule is the quantification of its deviation with respect to the known rule set. The quantification is objective and based on assumption of additive nature of newness. User subjectivity is captured by specification of threshold(s) for rule categorization.

The framework is implemented and evaluated using synthetic and real-life datasets and has shown positive results. The generated rules were categorized as conforming, generalized/specialized, unexpected and novel rules.

References

[1] A. S. Al-Hegami, " Subjective Measures and their Role in Data Mining Process ", In Proceedings of the 6th International Conference on Cognitive Systems, New Delhi, India, 2004.

[2] A. S. Al-Hegami, V. Bhatnagar, and N. Kumar, " Novelty Framework for Knowledge Discovery in Databases ", In Proceedings of the 6th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2004), Zaragoza, Spain, 2004, pp 48-55.

[3] A. S. Al-Hegami, " Interestingness Measures of KDD : A Comparative Analysis ",In Proceedings of the 11th International Conference on Concurrent Engineering: Research and Applications, Beijing, China, 2004, pp 321-326.

[4] B. Padmanabhan and A. Tuzhilin, " Unexpectedness as a Measure of Interestingness in Knowledge Discovery ", Working paper # IS-97-6, Dept. of Information Systems, Stern School of Business, NYU, 1997.

[5] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques", 1st Edition, Harcourt India Private Limited. 2001.

[6] M. H. Dunham, " Data Mining: Introductory and Advanced Topics ",1st Edition, Pearson Education (Singaphore) Pte. Ltd., 2003.

[7] G. Piateskey-Shapiro, and C. J. Matheus, "The Interestingness of Deviations", In Proceedings of AAAI Workshop on Knowledge Discovery in Databases, 1994.

[8] S. Basu, R. J. Mooney, K. V. Pasupuleti, and J. Ghosh, "Using Lexical Knowledge to Evaluate the Novelty of Rules Mined from Text ", In Proceedings of the NAACL workshop and other Lexical Resources: Applications, Extensions and Customizations, 2001.

[9] A. Silberschatz and A.Tuzhilin, "On Subjective Measures of Interestingness in Knowledge Discovery", In Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining. 1995.

[10] B. Liu, W. Hsu, and S. Chen, " Using General Impressions to Analyse Discovered Classification Rules ", In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD 97), 1997.

[11] T. Kohonen, " Self-Organization and Associative Memory ", 3rd Edition, Springer, Berlin. 1993.

[12] A. Silberschatz and A. Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems ", IEEE Transactions on Knowledge and Data Engineering. V.5, No.6. 1996.

[13] B. Liu and W. Hsu, " Post Analysis of Learned Rules ", In Proceedings of the 13th National Conférence on AI(AAAI'96), 1996.

[14] S. Marsland, " On-Line Novelty Detection Through Self-Organization, with Application to Robotics ", Ph.D. Thesis, Department of Computer Science, University of Manchester, 2001.

[15] N. Japkowicz , C. Myers, and M. Gluck, " A Novelty Detection Approach to Classification", In Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995.

[16] S. Roberts, and L. Tarassenko, "A Probabilistic Resource Allocation Network for Novelty Detection", In Neural Computation, 6(2), 1994

[17] A. Ypma, and R. Duin, "Novelty Detection Using Self-Organizing Maps", In Progress in Connectionist-Based Information Systems. Volume 2, 1997.

[18] Uthurusamy, R., "From Data Mining to Knowledge Discovery", In Advances in Knowledge Discovery and Data mining. Edited by U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, CA:AAAI/MIT Press, 1996.

[19] http://kdd.ics.uci.edu/

[20] http://www.comp.nus.edu.sg/~dm2/p_download.html

[21] T. Yairi, Y. Kato and K. Hori, " Fault Detection by Mining Association Rules from House-keeping Data ", In Proceedings of International Symposium on Artificial Intelligence, Robotics and Automation in Space (SAIRAS 2001), 2000.

[22] G. Psaila, "Discovery of Association Rule Meta-Patterns", In Proceedings of 2nd International Conference on Data Warehousing and Knowledge Discovery (DaWaK99), 1999.

[23] J. Pei and J. Han, "Can We Push More Constraints into Frequent Pattern Mining", In Proceeding of the 6th ACM SIGKDD, 2000.

[24] F. Bronchi, F. Giannotti, A. Mazzanti and D. Pedreschi, "Adaptive Constraint Pushing in Frequent Pattern Mining", In Proceedings of the 7th PKDD'03, 2003, pp 47-58.

[25] F. Bronchi, F. Giannotti, A. Mazzanti and D. Pedreschi, "ExAMiner: Optimized Level-wise Frequent Pattern Mining with Monotone Constraints", In Proceedings of the 3rd International Conference on Data Mining (ICDM03), 2003, pp 11-18.

[26] F. Bronchi, F. Giannotti, A. Mazzanti and D. Pedreschi, "Exante: Anticipated Data Reduction in Constrained Pattern Mining", In Proceedings of the 7th PKDD'03, 2003, 59-70.

 COPYRIGHT © 2005 ENFORMATIKA

A New Technique for Multi Resolution Characterization of Epileptic Spikes in EEG

H. N. Suresh, Dr. V. Udaya Shankara

***Abstract*—** A technique proposed for the automatic detection of spikes in electroencephalograms (EEG). A multi-resolution approach and a non-linear energy operator are exploited. The signal on each EEG channel is decomposed into three sub bands using a non-decimated wavelet transform (WT). The WT is a powerful tool for multi-resolution analysis of non-stationary signal as well as for signal compression, recognition and restoration. Each sub band is analyzed by using a non-linear energy operator, in order to detect spikes. A decision rule detects the presence of spikes in the EEG, relying upon the energy of the three sub-bands. The effectiveness of the proposed technique was confirmed by analyzing both test signals and EEG layouts.

***Keywords*—**EEG, Spike, SNEO, Wavelet Transform

I. INTRODUCTION

THE EEG is an important clinical tool for diagnosing, monitoring and managing neurological disorder related to epilepsy. This disorder is characterized by sudden recurrent and transient disturbances of mental function and / or movement of the body that results from excessive discharge of groups of brain cells.

The presence of Epileptiform activity in the EEG confirms the diagnosis of epilepsy, which some times can be confused with other disorders producing similar seizures like activity. During seizures the scalp EEG of patients with epilepsy is characterized by high amplitude, synchronized periodic EEG waveforms, reflecting abnormal discharge of large group of neurons. Between seizures, epileptiform transient waveform, which includes spikes and sharp waves, are typically observed on the scalp EEG of such patients. Detecting and classifying sharp transient waveforms by visual screening of the EEG record is a complex and time-consuming operation. Also such EEG records requires highly trained professional who are in generally short supply. Hence a requirement for automatic detection of EEG spikes and seizures. In addition the use of EEG monitoring, which produces 24 hours or longer continuous EEG recording, is becoming more common thus further increasing the need for automated detection methods. In the past many methods have been investigated to detect the EEG spikes. Mimetic techniques have been widely used to detect spikes, but difficulties arise with artifacts. These problems increase the number of false detection's, which commonly plague all automatic systems.

Manuscript received January 25, 2005

H. N. Suresh is with Bangalore Institute of Technology, Bangalore–04, Karnataka India.(Tel::91-80-26422726, M:9448685021, e-mail: hnssuresh@yahoo.com).

V. Udaya Shankara is with Sri Jayachamarajendra College of Engineering working as a Professor in the Department of Instrumentation Technology, Mysore, Karnataka. (e-mail: udaya@sjce.ac.in).

Although fairly successful this approach becomes increasingly difficult due to the proliferation of the rules and the need for computers with large memories and large processing power. In addition, EEGers cannot agree on a complete set of rules acceptable to all, limiting the success of this method. If we used Fourier Transform (F.T) to detect spikes, but that gives only frequency information of the signal. The Short time Fourier Transform gives time and frequency information simultaneously, but it suffers from resolution problems. In this research work, the smoothed non-linear energy operator (SNEO) has been proposed for the analysis of EEG signals.

I.I METHODOLOGY

Spike detection in EEG is an important task for the diagnosis of epilepsy. The shape and size of epileptic spikes essentially change from one patient to the other. They appear in the EEG as isolated events, as well as quasi-periodic oscillations of spike-and-wave. Epileptic spike detection is a very difficult task, since normal brain activity, non-pathological events that resemble pathological ones, noise and instrumental artifacts can be misinterpreted as epileptic spikes. our approach to spike detection relies on the observation that the impulse-like shape of spike would result in a broad-band signal, displaying large energy at all frequencies. Indeed when analyzed with a filter bank like the one provided by the wavelet multi-resolution decomposition, a spike generates events in all the sub-bands. On the contrary, normal brain activity and non-pathological events likely have low frequency contents and appear only in low-resolution sub-bands. In the presence of broadband noise, on the other hand, the mid-range frequency sub-bands have a large spike signal-to-noise-ratio, thus allowing for an easier detection. Our scheme does not decimate the EEG sub-bands, as in non-redundant representations, avoiding the problems arising from the shift-variant property of the wavelet transform.

The wavelet representation is a powerful technique that has been successfully exploited in the analysis of non-stationary signals, like biomedical signal processing [1, 2]. Unlike classical Fourier analysis, the wavelet representation allows for trading frequency resolution and time resolution. In its discrete implementation, the wavelet transform can be viewed as a filter bank, which provides a multi-resolution decomposition of the signal [3]. The signal is decomposed into a series of sub-bands, each relative to a peculiar spectral region, whose bandwidth linearly increases with frequency [4]. The simplest approaches that can be devised for spike detection in a multi-resolution analysis framework consist of energy estimates in number of sub-bands [5]. Indeed, although very fast, a single-resolution approach like that in [6] has some limitations. In [7] a nonlinear energy operator

 COPYRIGHT © 2005 ENFORMATIKA

(SNEO) is proposed for the direct analysis of the EEG signal. We show that multi-resolution analysis combined with SNEO give some advantages and provide a useful tool for EEG analysis.

II. SUBBAND DECOMPOSTION PRINCIPLES

In this section we briefly review the discrete-time wavelet transform and its relations with subband decomposition.

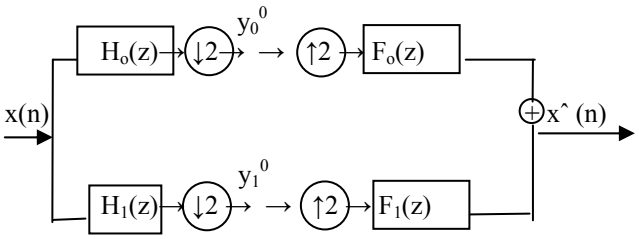

Figure 1: Two- channel subband system.

Consider the two-channel filter bank Fig. 1. The input signal x(n) is decomposed into two sub-bands by filtering with the low-pass filter $H_o(z)$ and the high pass filter $H_1(z)$. The output of the filters is decimated by a factor two. It is well know that it is possible to design the analysis filter $H_o(z)$, $H_1(z)$ and the synthesis filter pair $F_o(z)$,$F_1(z)$ in order to have perfect reconstruction of x(n) at the output of the synthesis stage. One possible way to achieve perfect reconstruction is to design the analysis filter impulse response ho(n) such that its z-transform satisfies.

$$H_o(z)H_o(z^{-1}) + H_o(-z)\,H_o(-z^{-1}) = 2, \qquad (1)$$

and choose $f_o(n) = h_o(-n)$, $f_1(n) = h_1(-n)$, $h_1(n) = (-1)^{1-n}$ Hospet (1-n). Note that the above equations imply that the filter impulse response $h_o(n)$ is orthogonal to its even-translates, namely

$$\langle h_o, n, h_o, n+2k \rangle \triangleq \sum_n h_o(n)h_o(n+2k) = \delta(k),$$

and that $< h_1, n, h_o.n+2k > = 0$, for all k. It is easy to see that the synthesis filters satisfy similar orthogonality conditions.

If we explicitly write the synthesis stage output as a function of the sub-band signal $y_0^0(n)$ $y_1^0(n)$, we have for an orthogonal perfect reconstruction system,

$$x(n) = \sum_k y_o^0(k)fo(n-2k) + \sum_k y_1^o(k)f1(n-2k) \qquad (2)$$

Thus equation (2) can be interpreted as the series expansion of the input over the orthogonal family of function {fo(n-2k), k∈z}.

In an octave filter bank, or discrete time wavelet transform, the low-pass signal $y_o^0(n)$ is further split by low-pass filtering and sub-sampling with the analysis filter. Fig. 2 shows the equivalent scheme for a two-stage sub-band scheme, where $y_o^0(n)$ is split into $y_o^1(n)$ and $y_1^1(n)$, and $Ho,o(z = Ho(z), Ho(z^2), Ho,_1(z) = H_0(z) H_1(z^2)$. The equivalent scheme is obtained by applying the Noble Identities, which allow to exchange the role of decimators and filters in the iterated sub-band scheme (3). Note that, for an analysis filter $H_o(z)$ with approximate bandwidth [0,fc/4], the equivalent filters Ho,o (z), $Ho,_1(z)$, and $H_1(z)$, have bandwidth [0, Fc/8], [fc/8, Fc/4], [F_c/4, Fc/2] respectively, where Fc is the input signal sampling frequency. Thus, the sub-bands $y_1^j(n)$provide a multi resolution representation of the input, each relative to a different frequency band. In particular, $y_o^1(n)$ is a decimated smooth version of x(n), while $y_1^1(n)$ and $y_1^0(n)$ are detailed signals to be added in the synthesis stage. Note that the decimators in Fig. 2 give rise to a ship-variant analysis stage. This is not a desirable feature when our goal is performing time localization of events, rather that providing a compact representation of the signal. To perform spike detection, we consider the signals $z_1^j(n)$ before decimation in Fig. 2, where j denotes the multi resolution level, and i ∈{0,1}.

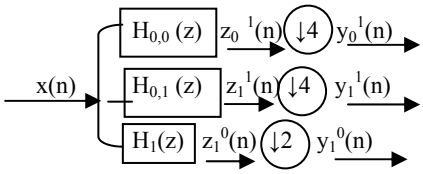

Fig. 2 Equivalent scheme for two levels of multi-resolution analysis.

III. THE SNEO OPERATOR IN THE FRAME WORK OF MULTI RESOLUTION ANALYSIS

The smoothed Nonlinear Energy Operator (SNEO) has been proposed in [7] for the analysis of EEG signals. SNEO is a smoothed version of the nonlinear energy operator.

$$\psi[x(n)] = x^2(n) - x(n+1)\,x(n-1) \qquad (3)$$

smoothing is achieved by low-pass filtering $\psi[x(n)]$, in order to obtain an estimate SNEO $[x(n)]$ of the expectation $E[\psi(n)]$. Indeed, taking the expectation of (3), for a stationary zero mean process $x(n)$ we obtain.

$$E[\psi[x(n)]] = r_x(0) - r_x(2) = \int_0^{2\pi} R_x(e^{j\omega})\,(1 - \cos 2\omega)\,d\omega/2\pi \qquad (4)$$

Where $r_x(k) = E[x(n)x(n+k)]$ is the input process autocorrelation function and $R_x(e^{j\omega})$ is the spectral density of $x(n)$. From equation (4) one can see that SNEO $[x(n)]$ is an approximation of the power of a band pass filtered version of the input process. For non-stationary process, a similar interpretation can be given in terms of the evolutionary spectrum [8]. More-over, if the smoothing low-pass filter has a short compact support,, the information provided by SNEO $[x(n)]$ is relative to the local characteristics of $x(n)$ around time n.

Beside its good properties for spike detection, the SNEO operator has some disadvantages, pointed out in the sequel, with respect to interference immunity, which our multi-resolution approach should overcome. Assume first that a constant value K is added to the EEG signal $x(n)$, during a given time interval. Such a phenomenon is produced, as an example, by patient movements, which produce an offset in the EEG measurement. We have

$$\psi[x(n) + K] = \psi[x(n)] + K\,(2x(n) - x(n-1) - x(n+1)).$$

Although low pass filtering attenuates the interference term, it is apparent that SNEO $[x(n)]$ depends on the local DC

 COPYRIGHT © 2005 ENFORMATIKA

value of the signal, and this is not a desirable effect in spike detection.

A more important drawback is the SNEO response to sinusoidal interference. Remarkably enough, when $x(n)= \cos\omega_0 n$, we have

$$\psi[x(n)] = ½ - ½ \cos 2\omega_0 = \text{const.}$$

Indeed, due to the additive property of the SNEO operator [7], when a sinusoidal interference is added to the EEG signal in a given time interval, it increases the SNEO output, which is misunderstood by a threshold based detector.

Our scheme exploits the SNEO operator in the framework of multi-resolution analysis. The signal is analyzed using three level discrete-time wavelet decomposition. The 5-tap almost orthogonal linear phase filters of [8] are used in the experiments. The detail signals $z_1^0(n)$, $z_1^1(n)$ and $z_1^2(n)$ are then processed using the SNEO operator. Note that, when the EEG signal is sampled by an F_s Hz frequency, the three details signals pertain to the Frequency bands $[F_s/4, F_s/2]$ Hz, $[F_s/8, F_s/4]$ Hz and $[F_s/16, F_s/8]$ Hz, respectively. An impulse-like signal, as a spike, generates a significant output in all the three sub-bands. On the other hand, sinusoidal, band pass and low pass interference is present in some or none of the sub-bands. Our idea is to devise a spike detector based upon the values $\text{SNEO}[z_l^j(n)]$, $j = 0, 1, 2$, $i = 1$. Given a specific threshold on each of the three levels, we say that a spike is detected at time n what at that time $\text{SNEO}[z_l^j(n)]$, is above the level threshold, for all $j = 0, 1, 2$, $i = 1$. A specific threshold value is used in each subband, to take into account the peculiar subband amplitudes corresponding to a spike.

IV. Data selection

The EEG data used in the study were obtained from 10 patients who were diagnosed with epilepsy and were under evaluation in the centre, National Institute of Mental Health and Neuro Sciences, Bangalore.

V. Experimental results

In this section, we show some results obtained applying our technique. Following [7], we consider a 250 sample test signal.

$$x(n)= \sin(2\pi n/75) – \sin(4\pi n/75 +\pi/2) + \sin(8\pi n/75) \quad (5)$$

adding synthetic spikes in random positions. A synthetic spike is a triangular symmetric pulse having a 5-sample support, random amplitude which is uniformly distributed in the range [2.5, 5], and a random sign. To avoid simultaneous spikes, we force a time separation of at least 11 samples between them. Fig. 3a shows a typical test signal including 4 spikes. In order to assess the immunity of our method to

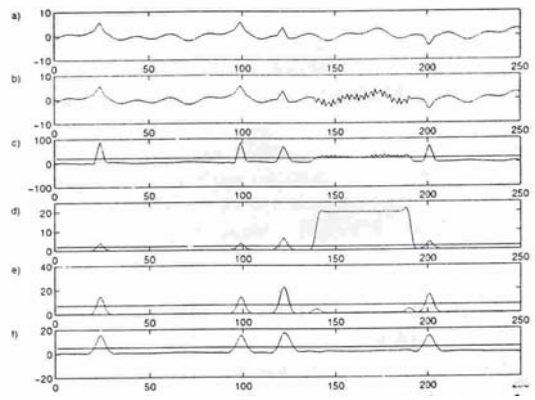

Figure 3 : a) Test signal, b) Test signal plus interference, c) SNEO and d)-e)-f) M-SNEO output, when applied to the signal in frame b).

Sinusoidal interference, we consider in Fig.3b the same test signal with a bursty 50 Hz sinusoidal interference around time sample $n = 170$.

We assume a sampling frequency F_s = 128 Hz. Fig. 3c shows the output of the SNEO operator, while Figs 3d, 3e, and 3f, show the results obtained by applying the SNEO operator to the three subbands $z_1^0(n)$, $z_1^1(n)$ and $z_1^2(n)$ (such three-fold out-put is hereafter called M-SNEO). Figure 3 also shows the thresholds, represented by solid lines, on each resolution level. On each level, the threshold was set to 80% of the M-SNEO magnitude corresponding to spike having minimum amplitude (i.e.2.5). Inspecting Figure 3 one can see that the use of detection criterion that takes into account the simultaneous presence of energy in the three subbands can be beneficial, providing a better immunity to interference than single resolution SNEO. In order to assess the noise sensitivity of the proposed procedure, we performed a set of 100 simulations where gaussian noise is added to the test signal (5). In these simulations, 8 spikes with random positions and random amplitude, are added to $x(n)$. The signal to noise ratio is calculated on the basis of the signal variance before the addition of the spikes. For a given spike-detector, let the false-negative ratio be FN=(Number of spikes missed) / (Actual number of spikes), and the false-positive ratio be FP=(Number of False spikes detected) / (Actual number of spikes). Table 1 reports FN and FP, together with the standard deviation of the results obtained applying the SNEO and M-SNEO detectors to both 100 synthetic signals with 30 dB SNR, and 100 synthetic signals with 5 dB SNR. Note that M-SNEO, with respect to single resolution SNEO, almost halves the number of missed spikes, simultaneously displaying a comparable or lower number of erroneous identifications.

 COPYRIGHT © 2005 ENFORMATIKA

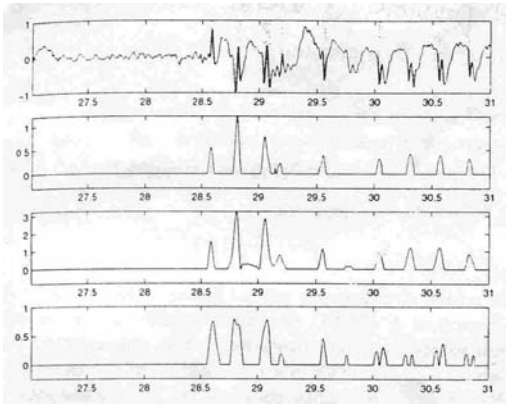

Figure 4: a) EEG signal, and b)-c)-d) M-SNEO

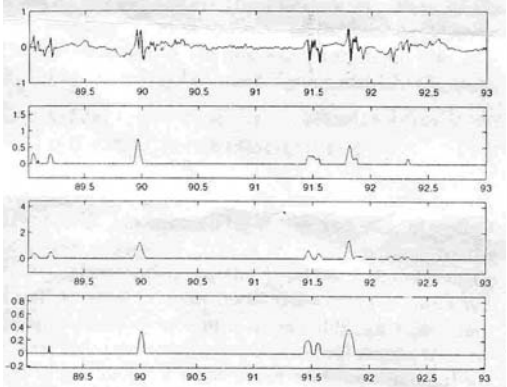

Figure 5: a) EEG signal, and b)-c)-d) M-SNEO

TABLE I
DETECTION RATIOS FOR THE TEST SIGNAL

Type	SNR	SNEO	M-SNEO
FN	30dB	0.05±0.07	0.02±0.06
FP	30dB	0.003±0.02	0.003±0.02
FN	5dB	0.11±0.09	0.06±0.09
FP	5dB	0.17±0.14	0.14±0.13

Eventually, let us consider EEG tracings. A set of 36 long-term EEG layouts, recorded by an 8 channel MEDILOG 9000 system, was processed. Analog signals were converted into digital ones by an A/D converter, having a 128 Hz sampling rate on each channel. Performing statistical computations on a distinguished subset encompassing 15 of our EEG tracings identified the threshold value on each resolution level. These tracings are representative of clinically relevant activity, in terms of morphology, spatial distribution, and discharge duration. They include also several artifacts. The computed thresholds were sub-sequentially exploited for analyzing the remaining EEG layout. Figures 4 and 5 show the M-SNEO output obtained analyzing one channel of two EEG layouts. On each resolution level, the threshold is shifted to zero, and the (shifted) M-SNEO output values larger than the threshold are plotted. The marked time intervals are shadowed in the frame showing the EEG signal. Inspecting Fig.4, we see that epileptic spike-and-waves are correctly detected. Figure 5, on the other hand, shows that non-epileptic spikes can be marked (in this case, chew artifacts). Note, however, that the number of marked time intervals was a small fraction of the overall recording time, thus reducing the cost for subsequent human analysis. Such reduction was our main goal.

CONCLUSION

Results found that, the multi-resolution analysis combined with SNEO give some advantages and provide a useful tool for EEG analysis.

ACKNOWLEDGMENT

Author wishes to express deep gratitude to Professor V. Udayashankara, for his advice & support throughout this research.

REFERENCES

[1] M. Unser and A. Aldroubi, "A review wavelets in biomedical applications," *Proceedings of the IEEE*, vol.84.pp. 626-638. April 1996.

[2] I. Clark. R. Biscay. M. Echeverria. And T. Virues, "Multiresolution decomposition of non-stationary EEG signals: a preliminary study", *Comput. Biol. Med*., vol.25, no.4 pp. 373-382, 1995.

[3] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs NJ: Prentice-Hall, 1995.

[4] S.G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intel*., vol.11, pp. 674-693, July 1989.

[5] C. E. D' Attellis, S. I. Isaacson, and R. O. Sirne, "Detection of epileptic events in electroencephalograms using wavelet analysis", *Annals of Biomed. Engng*, Vol 25, pp. 286-293, 1997.

[6] F. Sartoretto and M. Ermani, "Automatic detection of epileptiform activity by single-level wavelet analysis", *Clinical Neurophysiology*, vol.110, pp. 239-249, 1999.

[7] S. Mukhopadhyay and G. C. Ray. "A new interpretation of nonlinear energy operator and its efficacy in spike detection", *IEEE Trans. On Biomed. Engng*, vol.45, no.2, 1998.

[8] E. P. Simoncelli and E. H. Adelson. "Subband transforms" in subband Image Coding (J. W. Woods. ed.) pp. 143-192, Kluwer Academic Publisher. 1991.

[9] Miroslaw latka & Ziemowit "*Wavelet analysis of epileptic Spikes*". Wroclaw university of Technology, Poland, Dec. 22, 2002.

Prof. H.N. Suresh received his BE (E&C) and M Tech degree in Bio-medical engineering from the University of Mysore, Mysore in 1989 and 1994 respectively. Since 1989, he has been actively engaged in teaching and research at Bangalore & Mysore University, Karnataka , India. Currently Assistant Professor of Electronic and instrumentation Technology, Bangalore Institute of Technology, Bangalore. His main professional interest are in Bio-medical signal processing, artificial neural network, & Adaptive systems. He is a member of IEEE, BME & ISTE.

 COPYRIGHT © 2005 ENFORMATIKA

Considerations of Public Key Infrastructure (PKI), Functioning as a Chain of Trust in Electronic Payments Systems

Theodosios Tsiakis, George Stephanides, and George Pekos

***Abstract*—** The growth of open networks created the interest to commercialise it. The establishment of an electronic business mechanism must be accompanied by a digital – electronic payment system to transfer the value of transactions. Financial organizations are requested to offer a secure e-payment synthesis with equivalent level of security served in conventional paper-based payment transactions. PKI, which is functioning as a chain of trust in security architecture, can enable security services of cryptography to e-payments, in order to take advantage of the wider base either of customer or of trading partners and the reduction of cost transaction achieved by the use of Internet channels. The paper addresses the possibilities and the implementation suggestions of PKI in relevance to electronic payments by suggesting a framework that should be followed.

***Keywords*—**Electronic Payment, Security, Trust

I. Introduction

COMMERCE partners (customers, merchants and financial organizations) are no longer interacting by direct physical experience. Instead their experience is mediated through multidimensional interactive environments. Consequently, it is an uppermost issue for the transaction process of exchange of information over open heterogeneous environments (as the Internet) to create trust. The formal procurator, the World Wide Web can be thought as an untrusted environment with no trust affiliations. In contradistinction, a desired trusted environment is the one that the entities constitute it, are unique, unquestionably identifiable and ruled by a set of priorities and conditions.

Trust has a vital influence on consumer activities and thereby on e-commerce success. To address the role of trust in e-commerce, we need to answer a number of questions such as [1]:

- What factors influence the level of trust in the Internet?
- How does trust influence participation in e-commerce?

Internet and particular the services of WWW must constitute an image of life that reflects both human knowledge and human relationships.

Manuscript received May 5, 2005.
Email: {tsiakis, steph, pekos}@uom.gr

II. Identification of E-Commerce Security Skepticism

Before we give a possible approximation that can be thought as definition of security is imperative to allocate the components of a security system. We have a set of actions (A) applying on a system, a set of processes (P) functioning as a domain and a set of outputs (O) resulting the reprocess of actions. When two domains want to establish a communication channel between them, in order to exchange information, the system must designate a set of rules (security policy). Given the options and the possibilities of the information flow we can verify that a system is secure

Internet is structured as an undirected connected graph where nodes in the graph are routers and links (subnets or sub-networks). Each node and link has a unique id specified by an IP (Internet Protocol) address. In addition, each link has a cost, which can vary in time, and the distance between the two nodes is the sum of the link costs in the path between them.

Reference [2] consider the amount of time (duration) needed for a message to proceed from a network link to another, as a random variable with expected duration where the probability density function $p(t)$ for this time is known. Thus, the expected duration for the transaction is, simply:

$$\langle t \rangle = \int_0^\infty tp(t)dt \tag{1}$$

And the risk of transaction is:

$$\sigma = \sqrt{Var[t]} = \sqrt{\langle (t - \langle t \rangle)^2 \rangle} \tag{2}$$

The first step in a security project must contemplate the identification of all security requirements that can be applicable to a specific environment (the web). Next, it is critical to identify the parties that will be involved in an e-payment transaction and partition the transactions into autonomous actions that can be linked into the parties participating in an e-commerce environment. This information constitute a group of security requirements that develop security architecture (by means of procedures, mechanisms and policies [3].

By Security Architecture we mean the consideration of how a company's systems (in the widest sense) should be designed to ensure that the company meets its security objectives. It relates the security policies, and affects both systems bought and built for general use and a specific solution. A security Infrastructure is the practical realization of a security Architecture in a tangible and usable form.

 COPYRIGHT © 2005 ENFORMATIKA

Computer security refers to the process of prevention, protection and detection of the system and the data stored therein against unauthorized access, modification, destruction or use [4]. Next a question can come up, on how do we secure a faceless, non-physical, remote transaction between individuals and organisations. We must notate that the transmission of information can be materialized in two types of channels, open and secure channels. Open channels are communication channels on which communication may be intercepted by an unauthorized party, in opposition secure channels are communication channels on which data cannot be read, written or altered. This security can be achieved either physically by securing the communication link or cryptographically by securing an open channel [5].

The critical factors for an economic organization or enterprise to both implement and operate an e-commerce mechanism are the flow of money, information flow and product flow. But security and implementation cost are the fundamental. Electronic Commerce (e-commerce) can be highly beneficial in reducing business costs and in creating opportunities for new, simple and improved customer services. Attempting to define e-commerce we can suppose that is the operation of maintaining business transactions (exchange of value) with the use of telecommunication networks

Reference [6] divides e-commerce into three classes:

1. Electronic Fund Transfer (EFT): the methods or the systems of paying electronically, transferring money or funds electronically and exchange digital information by means of electronic payments.
2. Electronic Commercial Information Transfer System: the system that exchange commercial information digitally.
3. Electronic Marketplace: the domains on the Internet where the expectant buyer can seek and purchase goods and services.

But e-commerce involves more than simple on-line transactions. We consider it as a mass of diametric unconventional activities that need to perform operation market research, identification of new opportunities, products, supplying services and exchange ways.

Reference [7] differentiates e-commerce in 1) Business-to-business transactions, 2) Consumer-to-business transactions and identifies that the transaction of e-commerce process can be visualized as a cycle of four phases:

1. **Request** (request of providence)
2. **Negotiation** (conditions of satisfaction)
3. **Performance** (fulfilment and notification of realization process)
4. **Settlement** (acceptation and payment)

Although the progress that has been made for the amplification of methods for achieving secure business transaction electronically, the use of e-commerce has not reach satisfactory limits and it is not considered being a concerted system for transactions, especially financial.

This can be identified as high transactional risk [8]. Transactional risk results when markets fail to provide standard level of security in payments and services.

Inadequacy of trust to electronic commercial and security is a result of the geographical separation of buyers and sellers, often coupled with a lack of real time physical presence [9].

The electronic systems that support the infrastructure of electronic commerce are vulnerable to three aspects of risk: abuse, misuse, and failure. Examining these risks from a business perspective we can identify the primary loss of asset (both in monetary and informational value) and lack of trust to conduct business electronically. What can outspread the universal acceptation, adoption and use of electronic commerce are secure, reliable, speeder, available, renovate able and user friendly communication infrastructure. From all these perspectives the motion of security is the one that distinguish and should be addressed with our whole attention. That does not mean that the rest residue in the extent of importance.

For Internet to be accepted as a medium of conducting monetary transactions, there will need to be a higher degree of confidence in the technology's reliability and security. As with any communications medium, it has both advantages (flow of information and digital assets) and disadvantages (the risk of loss transforming progressively to damage the asset). Reference [10] in a micro and macro analysis have concluded that for Internet to be accepted as a medium to conduct monetary transactions there will need to be a higher degree of confidence in the technology's reliability and security.

The risks of enabling commercial transactions on network operation can be vitiated by the enforcement of security management and policy.

There are therefore three goals in securing electronic communications:

1. prevention from the maximum of the threats
2. detection of violations as soon as possible after they occur
3. reaction to security violations within the minimum of time

Having in mind that businesses are looking for possible ways to provide cost-effective, secure communications services that will enable them to link their business processes more closely with the partners, in a supply chain network, the issue of trust is catalyst.

III. ENABLING THE TRUST FACTOR

Reference [11] identifies that the majority of trust theories and mechanisms put the emphasis on trust based on the history of transaction experiences the partners had. More specifically, the challenge of the first trade problem in electronic commerce is to develop on line services that will lead companies to build trust among them without any previous experience. To design for trust, it is necessary to determine if, and under what conditions trust mechanisms are brittle [12].

Trust is a function of context, identity, reputation, capability and stake. Trust is also conditioned by social and cultural factors; in certain cultures tradition may provide a strong influence [13] The need of trust in electronic commerce is usually explained by time asymmetry, lack of power, or inability to conclude perfect contracts. The time asymmetry argument draws on the fact that usually transactions are

 COPYRIGHT © 2005 ENFORMATIKA

performed over a period of time [14]. Reference [15] have reported that trust is a catalyst for human cooperation and that people will trust and embrace e-commerce if they perceive sufficient security. They mention that is often ignored the trade-off between functionality and security. In addition, an entity can be said to "trust" a second entity when it (the first entity) makes the assumption that the second entity will behave exactly as the first entity expects [16].

There is a strain to simulate off line an on line trust. The cases might be similar (commerce transaction) and the element to be the exchange might be common however, the nature of the environment, the type of process and many more make the issue of trust variable. Reference [17] sustains this aspect and suggests that in the on-line world, there are two approaches defining relationships between trustors and objects of trust; computer-mediated communication for individual-to-individual trust relationships mediated through technology and in contrast, technology as the object of trust.

Trust and trustworthiness are the foundations of security. The basis for these trust relationships and how they are formed can dramatically affect the underlying security of any system—be it home protection or online privacy [18]. A trust relationship is a relationship involving multiple entities to trust each other having or not certain properties (the so-called trust assumptions). If the trusted entities satisfy these properties, then they are trustworthy.

Given a network of (n) participating members we can consider individuals member trust as Direct or Indirect (Recommended). The direct trust relationship exists, as the word implies, from direct experiences two members develop. In a payment framework let us suppose customer *c* and merchant *m*. The preference of member *c* to pay a curtain amount (a) is represented by $\rho_c(a) \in \{0, 1\}$, where 0 indicates that member *c* does not have sufficient trust to proceed in a payment transaction and 1 indicates the acceptance to proceed. Next the member *m* in the network operates as *c ? m* and so the function that indicates how *c* trusts direct or not *m*:

$$\rho_{cm}(a) = \begin{cases} 1 \text{ if } \rho_c(a) = \rho_m(a) \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

Reference [19] defines a recommendation of trust as:

$$C \text{ trusts.rec}_x^{\text{seq}} M \text{when.path} S_p \text{when.t arg et} S_t \text{ Value} V$$

A recommendation trust relationship exists if C is willing to accept reports from M about experiences with third parties with respect to trust class x. Seq is the sequence of entities that mediated the experience excluding C and M. Let p be the number of positive experiences with Q which P knows about with regard to the trust class x. Then the value v_z of these experiences is computed as follows:

$V_z(p)=1-a^p$ (4)

This trust is restricted to experiences with entities in S_t (the target constraint set) mediated by entities in S_p (the path constraint set). If p and n represent positive and negative experiences respectively with the recommended entities, the recommendation trust value v_r is computed according to the following formula.

$$V_r(p,n) = \begin{cases} 1-a^{p-n} \text{ if } p>n \\ 0 \text{ else} \end{cases} \quad (5)$$

According to the Figure 1, V_2 represents direct trust and V_3, V_1 represent recommendation trust.

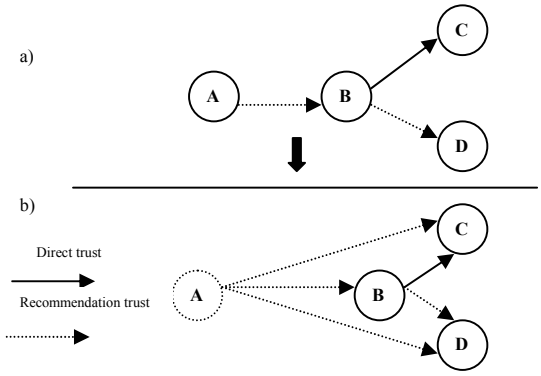

Fig. 1 Trust relationships

Due to an existing relationship, a new trust relationship can brought out between A and C as well as A and D can be derived. These processes are represented by the following equations:

Derived direct trust between A and C

$$V_1 \circ V_2 = 1-(1-V_2)^{V_1}$$
$$= 1-(1-(1-a^p))^{V_1} \quad p = \textit{number of positive experiences B had about C}$$
$$= 1-a^{v_1 p}$$

Derived recommendation trust between A and D

V_1V_3 = simply multiplication between V_1 and V_3

This multiplication shows that the value of derived recommendation trust decreases as the recommendation path grows.

IV. NEED FOR SECURITY

Protection of business transactions (referring to networks) and information within applications and web services from unauthorized use can be seen through the key security issues of:

Privacy: during the transmitting of the message the message in any form must be not altered or read.

Authentication: each party taking part in a communication must be sure of the identity of each other.

In addition we can see

Confidentiality: the process of keeping information in secret form.

Integrity: the role of proving that the information has not been tampered during its transit or its storage on the network.

No repudiation: the method of ensuring that the information cannot be disclaimed.

We need to ensure these properties or fundamentals security services in order to abjure and persevere the four types of possible security attacks:

Interruption: Attacks by unauthorized users can lead to system failure.

Interception: An unauthorized individual (C) intercepts the message content and changes it or uses it for malicious purposes.

Modification: The content of the message is modified by a third person C

 COPYRIGHT © 2005 ENFORMATIKA

Fabrication: Another user C, can produce messages and send them to B, by making them look just like they have been sent from A.

E-business requirements for security vary from company to company. The cost of security measures must be commercially justifiable. Risk analysis and the investigation of possible protective mechanisms must include an estimation of both the value of information and the likelihood of a security leak. They are essential tools for determining the appropriate security architecture. The final result of this process has to be a well defined security policy, which must be consistently implemented and frequently reviewed and up to last technology reports and methods.

Reference [20] distinguishes management of security networking environment in:

1. Defining a set of security policies which describes an organization's security,
2. Deploying, configuring, and monitoring a set of security, and
3. Monitoring the firing patterns of the security rules

Electronic payment systems are the most essential part of electronic commerce and electronic business. Electronic payment mechanisms provide the financial infrastructure needed to open the electronic marketplace.

There are three payment protocol models [21]:

1. Cash, tokens that can be authenticated independently by the issuer
2. Cheque, payment instruments whose validity requires reference (also called Credit/Debit instrument) to the issuer.
3. Cards, payment through existing credit card mechanism.

The problem is how to enable the traditional ways of paying for goods and services to work similarly and suitably over the Internet. Similar is the theme of what measures are needed to insure an open network as Internet, to transfer the digital image of information with compliance to security services.

There are many approaches to integrate and plan a strategy for business in the information technology place [22] The typical procedure is:

- Macro planning, for the realisation of business requirements and the outcome of this economic epicheirema.
- Micro planning, for step by step phase planning of each and every part of application.
- Information analysis, definition of technological equipment.
- System development process, the final process of design, development and implementation of e-business plan.

Skepticism can be developed about the support of e-business applications due to required level of trust that is needed to transact over the Internet. Security doubts are made about the ability to establish and perform the four services of security.

V. Involving Cryptography

As the information relies on security, cryptography plays the central role in an information security plan. It safeguards the integrity and the confidentiality of stored and transported data [23]. First we must distinct implementation between algorithms. Algorithm is a mathematical procedure with finite set of rules-actions for a problem solving and implementation is the process that defines/shows how this theoretic evaluation can be carried out in the real world. What we need to look for is the implementation. But we have to consider the existence of an algorithm that satisfies the following criteria: completion, decisiveness and affectivity. A communications system under no circumstances has to rely on the secret algorithm. The system relies on the secrecy of the deciphering key.

Cryptography can be the progenitor of every security solution. That means that a security policy, architecture and implementation cannot be without at least taking into consideration the use of cryptographic tools. Cryptography is fundamental in creating and maintaining secure information to sufficiently identify users in electronic business environments.

A typical cryptosystem consists of a plaintext P, ciphertext C, a cryptographic algorithm cipher, and a Key(s). The key or (s) is the secret information shared by the originator and the recipient that is used to secure the plaintext data by the application of the cipher. Encryption is a key-based mathematical transformation that changes plaintext to ciphertext, in such a way that the reverse operation - decryption - is very difficult without possession of the key.

Two basic kinds of cryptographic transformations exist. The single key or symmetric cryptography uses the same key for both encryption and decryption. The two key (pair) or public key or asymmetric cryptography (the one key encrypts and made public, the other decrypt and is kept secret). In the Public-Key Crypto-Systems the Keys are generated in matching pairs. The success of public key cryptography (PKC) is mainly based on the mathematical difficulty of factoring very large prime numbers.

The overall security of e-payments and online transactions is Cryptography, that can be seen as part of a hole security solution, in which her role is to obtain that the transmitted data of information in a communication systems are provable secure. Reference [24] in a confrontation between Provable Security and Practical Security led to conclusion that the perfect situation is reached when one manages to prove that, from an attack, one can describe an algorithm against the underlying problem, with almost the same success probability within almost the same amount of time.

Public Key Infrastructure is a mainstream method, to ensure key management and reliable authentication and encryption between two objects that are communicating over a single open network. The use of public-key cryptography requires a public-key infrastructure to publish and manage public-key values.

Reference [25] report that it is not clear what value a PKI brings in electronic commerce. What can be regarded as positive is that authentication protocols can verify the identity of an entity that one already knows about. On the Internet, users come into contact with businesses they have never met before.

 COPYRIGHT © 2005 ENFORMATIKA

VI. Electronic Payment (e-payment) Phase

Consumers and providers of products and services are not expected to use widely electronic commerce applications unless they are confident that electronic communications and transactions will be confidential, the origin of messages can be verified and the personal privacy can be protected [26].

Payments are considered to be the integral component of any commerce activity. The needfulness to accelerate the flow of e-commerce transaction leads to establish a scrutable, friendly and secure payment system. Acceptance of e-commerce depends on the confidence of discernible security. Only one security issue is solitary to electronic commerce, which is the electronic payment.

It is preferable to make a distinction between electronic transaction protocols and electronic payment protocols. Electronic payment deals with the actual money transfer, electronic transaction protocols deals with the transactions as a whole. Electronic transaction protocols group together operations and implement failure atomicity, permanence and serializability and electronic payment protocols transfer trust, either as cryptographically signed promises, or as digital cash [27].

Reference [28] defines "Electronic payment" or "e-payment" as the transfer of electronic means of payment from the payer to the payee through the use of an electronic payment instrument. An "electronic mean of payment" would be defined as a mean of payment that is represented and transferable in electronic form. In a similar vein, an "electronic payment instrument" can be understood to be a payment instrument where the forms are represented electronically and the processes that change the ownership of the means of payment are electronic.

Electronic payment mechanisms as mentioned before provide the infrastructure (financial) that is indispensable to open and then establish an aggregate electronic marketplace. Within similar types of electronic payment systems, the encoding and decoding mechanisms of individualized payment systems follow different procedures [29].

The first distinctive feature of e-payment systems is the money model.

- Token – when the medium of exchange represents a value
- Notational – when a value is stored and exchanged by authorisation

A payer and a payee are the conceptual parts that exchange money for goods or services, and a financial institution is the one which links "bits" to "money." Payments can be performed either on-line **(**real time authorisation) or off-line **(**without contacting any third party during payment) [30]. On-line payment means that the payment systems requires from the payee to contact a third party in order to verify the process of payment and Off-line that there is no need of contacting and verifying the transaction of payment). We can add semi-online category as the involvement of a trusted third party but not in every payment transaction. The element of order is the validation of payment

Next, the time when the monetary value is actually taken from the payer attributes e-payments into

- Pre-paid systems – customer's account debited before payment
- Pay-now systems – customer's account debited at the time of payment
- Post-pay systems – merchant's account credited before customer's account is debited

Last distinctive feature, but not final, can be considered the payment amount.

- Micro payments, when amount is less than 1€
- Small payments, amounts between 1€ and 15€
- Macro payments, when the amount is bigger than 15€

In the current evaluation process our concerns are the on-line, macro payment systems that offer the ability of interactivity and access to services and large amounts of value.

The stimulants to turn to electronic equivalent fermentations are the need to achieve inferior processing cost, payment anonymity and confidentiality and payer untraceability.

Payment Models classify the digital payment systems according to the necessary flow of information between the participants of an electronic transaction [31]. Considering payments that take action over the Internet the keys issues are to prevent double spending (digital cash is represented by bytes that can easily be copied and re-spent), counterfeiting (digital money can only represent real value) and privacy control (confidentiality, anonymity and untraceability).

VII. Using Public Key Infrastructure (PKI)

Public Key Infrastructure (acronym - PKI) is a set of services that enable the use of public key cryptography (Simplified Key distribution, Digital Signature, Long-Term encryption) in a networked environment [32]. A PKI is the set of components, people, policies and procedures which provides the foundation for the management of keys and certificates used by public key-based security services. A PKI assures the trustworthiness of public key-based security mechanisms.

Before utilization of PKI as a component of a whole security project, several issues must be addressed. Concisely we can distinguish:

- The range of interaction (global or national)
- Operational management (previous experience)
- The economic growth of entity that can lead to expand (assets)
- Acceptance of product(s) or service(s)
- The financial result (cost and outcome of implementation)

Starting to operate PKI we can find out two basics: Certification (the process of binding information such as the public-key value to an entity) and Validation (the process of verifying that a certification is still valid). The way these two operations are implemented is the basic defining characteristic of a PKI.

Public key infrastructure can only provide two basic functions [33]:

- Establish identity (by possession of the private key).

 COPYRIGHT © 2005 ENFORMATIKA

- Enable secure communication (through use of protocols that exploit properties of asymmetric algorithms) between two parties.

There are many approaches to fulfil a comprehensive list of PKI services that satisfy the security requirements [34], [35] and [36].

Security Policy: describes the business practices of the organization and defines the principles for the use of cryptography.

Certification Authorities (CA): issue digital certificates to valid applicants, set the expiry date for certificates and invalidate them when the validity period expires. There are in general two types of structure for a CA: the CA hierarchy and the cross certification. The first CA is built up in one root CA. The cross-certification is a flat and the top node in each hierarchy is connected through each other.

Registration Authorities (RA): is the interface between the user and CA. It authenticates the user and determines the level of trust.

A Certificate Distribution System divided into:

Certificate Holders: subjects or end-entities which get the certificates from CA

Certificate Repository: the storage area of PKI (storing and distributing of entities certificates)

Validation Server: an accessory sever (to provide certificate status, date of expiration etc)

The number of keys required for a setup of a communication system with n users is 2^n as against n (n-1)/2 required for a corresponding symmetric key system. It is obvious that as the number of users boosts, a symmetric key setup becomes rather incapable. Public keys can be published easily without peril the security of the pair keys especially the private or the system. The security of the cryptosystem is dependent upon the key lengths being used. The larger the key length is, the more difficult it is to attack. Regardless of strength the large length of a key lends to PKC, reduces the speed of computation analogue to symmetric key cryptosystems (the biggest disadvantage).

A PKI policy [37] contains a set of rules that must be enforced or applied by an element of the PKI. The rules include a specification about which are the sets of users controlled by them and one or more values related to the parameter. A typical categorization of rules is the Certification rules (control of validity period, key type, key length, certificate), Re-issuance rules (applied to certificates that are about to expire and control whether the certificate can be re-issued and the next validity period) and Revocation rules (specification about what should be done when a particular key is compromised).

PKI can provide higher levels of confidence for exchanging information over the Internet. It achieves this by:

- offering certainty of the quality, source and destination of information sent and received electronically;
- assuring the time that information was sent and received (if known); and
- ensuring the privacy of information sent.

Public key cryptography by its own means is not enough [38]. The reproduction of contractual commerce in the electronic environment shows that is required:

- Security policies to define the rules under which cryptographic systems should operate
- Products to generate store and manage certificates and their associated keys
- Procedures to dictate how keys and certificates are generated and distributed

A trusted and authenticated key distribution infrastructure is necessary to support the use of public keys in a public network such as the Internet.

Public Key Infrastructure is composed of three main entities [32]; the Certification Authority, the Registration Authority and the Repository or Directory Server. The PKI functions are Key Generation and Distribution, Certificate Validation, Generation, Revocation and Management of trust.

The public keys are stored in directory in order to be accessible over open networks such as Internet.

The identity of involved parties is provided by a unique key that can be used with encryption to stamp data or a transaction with a unique identification key. The transmitted data are guaranteed that they haven't been altered by digital signatures.

The implementation of a PKI requires an analysis of business objectives and the trust relationships that exist in their environment. The awareness of these trust relationships leads to the establishment of an overall trust model that the PKI enforces [38], [39].

A classification of trust models can be placed as the following:

Hierarchical: Can be considered as the simplest form of trust model, that allows end entities' certificates to be signed by a single CA.

Distributed (Web of Trust) or Pretty Good Privacy (PGP): Every entity has it's own root CA. It can be thought as a system without the incorporation of a CA. It is used by individuals to encrypt and digitally sign electronic mail messages and files.

Direct (Peer to Peer): A trusted third party does not exist in a direct trust model and each end entity in a peer-to-peer relationship establishes trust with every other entity on an individual basis.

VIII. Creation of Trust Factor and a Trusted Payment Framework

A trusted environment is characterized by a unique identification process and is considered to be the one that has a minimum number of a priori trusted entities. In a PKI a trust anchor is any CA, which is trusted without the trust being referenced through the PKI certificates [40], [41]. Simply the PKI enables the establishment of a trust hierarchy. The transaction entities are unfamiliar and they must establish a trust relationship with a CA. Next, the CA authenticates the entity (referring to established rules that noted in Certificate Practices Statement-CPS), and then issues for each entity a digital certificate. That digital certificate is now signed by the CA and can be considered as a personal identification. These certificates are capable of establishing trust between the

 COPYRIGHT © 2005 ENFORMATIKA

unknown entities as long as they trust the CA. The motive of this trust establishment is to offer a way to transmit data securely over insecure heterogeneous networks.

The public keys are placed in a storage area named *trusted party*. Both the name and public key of the entity with the digital signature of the trusted party is called a *Certificate*. The certificate is important for authentication because it is containing the name, key and signature of the entity.

To authenticate a transacting communication entity we need to authenticate it by the use of a third party. This process enables two diametrically singulars to trust indisputably each other, even though they have not previous personal relationship [42]. The party that necessitates trust to other entities participating in the information transaction, such as payment, is called Trusted Third Party (TTP). Because a TTP issues certificates, it is commonly referred as a Certification Authority or CA.

Electronic payment is confidential when all phases of the procedure are designed to satisfy the participants and their security expectations. To build up trust in the electronic payment system, three elements must be taken into consideration: data, identities and role behavior.

Knowing that commerce exchange is based on the trust between the parties and that internet is a distributed environment with no trust, we can put an end to this problem by the use of a trusted service mechanism (TSM) utilizing/exploiting the browser trust. The reason is that the browser trust list model is very common, simply, flexible and with a scalability of hierarchy. These elements satisfy the need for trust management for distributed environments.

The TSM distributes certificates of CA (root CA or subordinate CAs) in order to achieve the security services of authenticity and integrity. The relying party that has a need to consume the digital product or service of the TSM, simple corresponds as trust anchor to form a trust chain.

Figure 2 illustrates how possible user might use a trust service. The identified parts are a set of four entities (PKI system/ website/ TSM trust web and a client Customer). PKI publishes its certificate of root CA and optional subordinate CAs on the selected TSM in a security way, possible offline way (1). TSM evaluates the PKI security policies and assigned a security level to it. A future customer then, will visit the website (which has a certificate issued by a subordinate CA (2)) to buy digital goods. The n he points to the website (3) and downloads the certificate of the website (4), in order to verify the certificate and he sends a query message with the certificate to the selected TSM. TSM is capable to verify the certificate because it has known the certificate of root CA. If it validated, TSM will response a message to inform the customer about the status of the website's certificate and the level of security. Depending on the report of this message customer can decide whether or not trust the website and proceed to buy the needed digital goods from it.

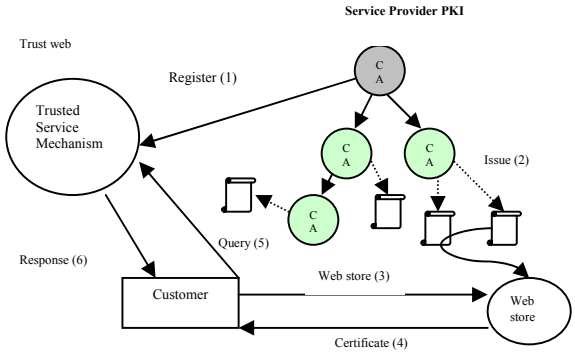

Fig. 2 A trusted framework

IX. CONCLUSIONS

If we consider establishing a PKI we must take into serious consideration, issues involving ease of use, ease of deployment, centralized management requirements and security service integration. Providence of end-users with integrated electronic identities (certificates), digital signatures and encryption facilities must be without doubt. We have to accomplish a security mechanism to lay the foundations of trust. PKI is the emergency tool for the establishment of a trust hierarchy. It is the underlying principal of every PKI, due to the fact that electronic commerce operates with trust mechanisms comfortable with risk management operation. The parties-entities (unknown to each other) transacting in open environment as the Internet, do not have sufficient trust established between them to perform business, contractual, legal, or other types of transactions. The implementation of a PKI using a CA provides this trust. This implementation of trust is capable of immunizing the essential part of electronic commerce, the electronic payments.

REFERENCES

[1] B. Corbitt, T. Thanasankit, H.Yi, *Trust and e-commerce: a study of consumer perceptions*, Electronic Commerce Research and Applications, 2, 2003, pp. 203–215

[2] R. Lukose, B. Huberman, *A methodology for managing risk in electronic transactions over the Internet*, Netnomics, 2000, pp. 25–36

[3] S. Gaines, Z. Norman, *Some Security Principles and Their Application to Computer Security*, the National Science Foundation under Grant No.MCS76-00720

[4] G. Whitson, *Computer security: theory, process and management*' Consortium for Computing Sciences in Colleges, JCSC 18, 2003

[5] D. Pipkin, *Information Security*. Prentice Hall PTR, 2000

[6] L. Fera, M. Hu, G. Cheung, M. Soper, *Digital cash payment systems*, Report, 1996

[7] S. Katsikas, *The Role of Public Key Infrastructure in Electronic Commerce*' The electronic journal for e-Commerce Tools & Applications, eJETA.org, Vol.1, No.1, 2002

[8] C. Westland, *Transaction Risk in Electronic Commerce*, Decision Support Systems 33, Elsevier, 2002, pp. 82-103

[9] P. Skevington, T. Hart, *Trusted third parties in electronic commerce*, BT Technology Journal, Vol. 15, No 2, 1997

[10] S. Lancaster, S. Yen, S. Huang, *Public key infrastructure: a micro and macro analysis*, Computer Standards & Interfaces 25, Elsevier Science, 2003, pp. 437–446

[11] Y. Tan, *A Trust Matrix Model for Electronic Commerce*, Trust Management, LNCS Springer-Verlag, 2692, 2003, pp. 33–45

[12] J. Camp, *Designing for Trust*, LNAI 2631, Springer-Verlag, 2003, pp. 15–29

 COPYRIGHT © 2005 ENFORMATIKA

[13] J. Daniel, *Patterns of Trust and Policy*, New Security Paradigms Workshop Langdale, 1998, Cumbria UK
[14] S. Brainov, T. Sandholm, *Contracting with Uncertain Level of Trust*, 1999, ACM 158113-176
[15] M. Patton, A. Josang, *Technologies for Trust in Electronic Commerce*, Electronic Commerce Research, Vol. 4, 2004, pp. 9–21
[16] ITU-T Recommendation X.509 (2000) Information Technology, Open systems interconnection - The Directory: Public-key and attribute certificate frameworks
[17] C. Corritorea, B. Krachera, S. Wiedenbeck, *On-line trust: concepts, evolving themes, a model*, Int. J. Human-Computer Studies 58, 2003, pp. 737–758
[18] J. Viega, T. Kohno, B. Potter, *Trust (and mistrust) in secure applications*, Communications of the ACM, Vol. 44, No. 2, 2001
[19] T. Beth, M. Borcherding, B. Klien, *Valuation of Trust in Open Networks*, Proceedings of the European Symposium on Research in Computer Security, Brighton, 1994
[20] L Ho, *Distributed Security Management in the Internet*, Journal of Network and Systems Management, Vol. 7, No. 2, 1999
[21] H.-W.-P. Beadle, R. Gonzalez, R. Safavi-Naini, S. Bakhtiari *Review of Internet Payment Schemes*, Proceedings of ATNAC'96, 1996
[22] M. Chesher, R. Kaura, *Electronic commerce and business communications*, Springer-Verlag, 1998
[23] E. Verheul, B. Koops, H. Tilborg, *Public key infrastructure - Binding cryptography -- A fraud-detectible alternative to key-escrow proposals*, Computer Law and Security Report, Vol. 13, no.1, 1997, pp. 3-14
[24] D. Pointcheval, *Practical Security in Public-Key Cryptography*, ICICS 2001, Lecture Notes in Computer Science Vol. 2288, 2002, pp. 1–17
[25] T. Aura, D. Gollmann, *Communications security on the Internet*, Focus Software, No. 105, Volume 2, Issue 3, 2001, pp. 104-111
[26] I. Mavridis, G. Pangalos, T. Koukouvinos, S. Muftic, *A Secure Payment System for Electronic Commerce*, 10th International Workshop on Database & Expert Systems Applications, Florence, Italy, 1999
[27] P. Havinga, G. Smit, A. Helme, *Survey of electronic payment methods and systems*, University of Twente, department of Computer Science
[28] Electronic Payment Systems Observatory (ePSO), Building Security and Consumer Trust in Internet Payments, Background Paper No. 7, 2002
[29] Yu Hsiao-Cheng, His Kuo-Hua, Kuo Pei-Jen, *Electronic payment systems: an analysis and comparison of types*, Technology in Society 24, 2002, pp. 331–347
[30] D. Abrazhevich, *Classification and Characteristics of Electronic Payment Systems*, Lecture Notes in Computer Science, Vol. 2115, 2001, pp. 81-90
[31] J. L. Abad-Peiro, N. Asokan, M. Steiner, M. Waidner, *Designing a generic payment service*, Technical Report 212ZR055, IBM Zurich Research Laboratory, 1996, Availiable: http://www.semper.org/info/212ZR055.ps.gz,
[32] D. Bruschi, A. Curtil, E. Rosti, *A quantitative study of Public Key InC. Sundt, PKI — Panacea1 or Silver Bullet*, Information Security Technical Report, Vol 5, No. 4, 2000, pp.53-65frastructures, Computers & Security, Vol 22, No 1, 2003, pp. 56-67
[33] C. Sundt, *PKI — Panacea1 or Silver Bullet, Information Security Technical Report*, Vol 5, No. 4, 2000, pp.53-65
[34] S. Gritzalis, S. Katsikas, D. Lekkas, K. Moulinos, E. Polydorou, *Securing The Electronic Market: The KEYSTONE Public Key Infrastructure Architecture*, Computers & Security, Vol. 19, No. 8, 2000, pp. 731-746
[35] K. Liaquat, *Deploying Public Key Infrastructures*, Information Security Technical Report, Vol. 3, No. 2, 1998, pp. 18-33
[36] R. Hunt, *PKI and Digital Certification Infrastructure*, Proceedings of the 9th IEEE International Conference on Networks (ICON.01), 2001, pp. 234-239
[37] A. Gómez, G. Martýnez, Ó. Cánovas *New security services based on PKI*, Future Generation Computer Systems 19, 2003, pp. 251–262
[38] J. Weise, *Public Key Infrastructure Overview*, Sun BluePrints™, 2001
[39] RSA Inc. Understanding Public Key Infrastructure (PKI), An RSA Data Security White Paper, RSA Data Security, Inc., 1999
[40] M. Henderson, R. Coulter, *Modelling Trust Structures for Public Key Infrastructures*, ACISP 2002, Lecture Notes in Computer Science, Vol. 2384, 2002, pp. 56–70
[41] S. Gritzalis, D. Gritzalis, *A Digital Seal solution for deploying Trust on Commercial Transactions*, Information Management and Computer Security, Vol.9, No.2, 2001, pp.71-79
[42] M. Benantar, *The Internet public key infrastructure*, IBM, 2001

Theodosios Tsiakis is a Research Assistant teaching Introduction to Computer Science and Cryptography in the University of Macedonia, Dept. of Applied Informatics. His main research interests are financial cryptography and trust management.

George Stephanides is an Assistant Professor similarly in the University of Macedonia, Dept. of Applied Informatics teaching Object Oriented Programming, Computational Mathematics, Cryptography and Algorithms. His scientific research focus on computational number theory, cryptography and computer programming.

George Pekos is a Professor in the University of Macedonia, Dept. of Applied Informatics teaching Computational Mathematics, Cryptography and Statistics. His scientific research focus on computational mathematics, cryptography and applied economics.

 COPYRIGHT © 2005 ENFORMATIKA

Detection and Correction of Ectopic Beats for HRV Analysis Applying Discrete Wavelet Transforms

Desmond B. Keenan

***Abstract*—**The clinical usefulness of heart rate variability is limited to the range of Holter monitoring software available. These software algorithms require a normal sinus rhythm to accurately acquire heart rate variability (HRV) measures in the frequency domain. Premature ventricular contractions (PVC) or more commonly referred to as ectopic beats, frequent in heart failure, hinder this analysis and introduce ambiguity. This investigation demonstrates an algorithm to automatically detect ectopic beats by analyzing discrete wavelet transform coefficients. Two techniques for filtering and replacing the ectopic beats from the RR signal are compared. One technique applies wavelet hard thresholding techniques and another applies linear interpolation to replace ectopic cycles. The results demonstrate through simulation, and signals acquired from a 24hr ambulatory recorder, that these techniques can accurately detect PVC's and remove the noise and leakage effects produced by ectopic cycles retaining smooth spectra with the minimum of error.

***Keywords*—**Heart rate variability, vagal tone, sympathetic, parasympathetic, wavelets, ectopic beats, spectral analysis.

I. Introduction

HEART rate variability is a measure of alterations in heart rate by measuring the variation of RR intervals and has been shown to provide an assessment of cardiovascular disease [1]. Heart rate is influenced by both sympathetic and parasympathetic (vagal) activity. The influence and balance of both branches of the anatomic nervous system (ANS) is known as sympathovagal tone reflected in the RR interval changes. HRV measurements can be made by spectral analysis where a measure of the power in each of four frequency bands is representative of the four components of HRV. A low frequency (LF) component provides a measure of sympathetic effects on the heart and generally occurs in a band between 0.04 Hz and 0.15 Hz. Non regular effects such as chemoreceptors, thermoreceptors, and the reninangiothensin system can be monitored at ultra-low frequencies. A measurement of the influence of the vagus nerve in modulating the sinoatrial node can be made in the high frequency band (HF) loosely defined between 0.15 and 0.4 Hz known as respiratory sinus arrhythmia (RSA), and is a measure of cardiac parasympathetic activity. This rhythm is generated by the sinoatrial node; sometimes referred to as the pacemaker of the heart, and is nodal tissue located on the upper wall of the right atrium and sets the rate of contraction by generating nerve impulses causing the atria to contract. This frequency band or rate of contractions can be variable though closely related to the frequency of respiration. The ratio of power in the LF and HF bands (LF/HF) provides a measure of cardiac sympathovagal balance. Several studies have indicated that the powers of the LF and HF components, occurring in synchrony with vasomotor waves and respiratory acts, respectively, appear to reflect in their reciprocal relationship the state of sympathovagal balance in numerous physiological and pathophysiological conditions [2-4]. A correlation between body fat content and the LF/HF ratio after glucose ingestion have been correlated in nonobese subjects with various levels of body fat content [5]. Decreased values of HRV measures are indirectly proportional to pressure, body mass index and insulinemia in young and obese patients without clinical symptoms of cardiovascular disease, diabetes or damage of target organs. In a study by Ravagli et al [6], pressure values correlated closely with the insulin level. These HRV measurements in the simplest form can be sensitive to stress i.e. the mental load on the brain. This measure can be seen to decrease with age, which has been attributed to a decrease in efferent vagal tone. Exercise on the other hand increases vagal tone. The scope of what HRV is capable of diagnosing or determining is far from understood and it is a busy area of research. It has been used as a measure of mortality primarily with patients who have undergone cardiac surgery. Clinical depression strongly associated with mortality with such patients may be seen through a decrease in HRV [7]. Spectral HRV was previously described as a measure of power in various frequency bands. To determine the RSA amplitude over a period of time, frequency domain, time domain and phase domain approaches have been analyzed. A time-domain approach by Grossman [8] known as the peak-valley algorithm measures the maxima and minima values of RR time intervals within each breath. In the frequency domain a value for RSA can be derived by applying an appropriate window function to a given time series to reduce spectral leakage from random events and applying the Fourier transform to the filter residuals. When analyzing the HF band

D. B. Keenan is with Research and Development, Medtronic MiniMed, Northridge, CA 91325 USA (phone: 818-576-4472; fax: 818-576-6206; e-mail: barry.keenan@ medtronic.com).

 COPYRIGHT © 2005 ENFORMATIKA

the respiratory frequency should be predetermined through some kind of respiratory monitor, determined through paced breathing or at worst estimated. This frequency domain approach takes the average response over time, nominally a 5-minute period. This approach provides little information about the RSA waveform as it is time averaged and may not reflect the cardiac vagal dynamics. Another approach to evaluating RSA has been to analyze dynamics of heart rate with respect to phase [9] and is referred to as the phase domain approach. Frequency domain approaches to HRV, which are essential for the low frequency sympathetic measures are greatly hindered by the presence of ectopic beats [10-11]. Ectopic beats, when activated in the ventricles, are fairly common in subjects suffering from heart failure. There exist algorithms that detect and classify ectopic beats [12], but for HRV analysis these beats must be removed either by editing [13], or some means of interpolation or filtering. This paper describes a wavelet-based approach using Daubechies wavelets for detecting and filtering PVCs, and compares the filtering technique to linear interpolation based on simulation data and data from a heart failure patient.

II. METHODS

A. Data Collection

The data in this investigation was collected from an ambulatory monitoring device, the LifeShirt. The LifeShirt employs the Konno and Mead [14] two compartment-breathing model of the respiratory system. This approach shows changes in tidal volume measured at the mouth to be comparable to the sum of changes in ribcage and abdominal contributions. These volume changes are normally obtained by measuring variations in the thoracic and abdominal regions by inductive plethysmography (IP), magnetometers, or strain gauges [15-17]. The LifeShirt contains two IP sensors encircling the ribcage and abdomen used to measure tidal volume. A low oscillating current is passed through these inductive bands creating a magnetic field that is required to measure the self-inductance of the band's coil. The self-inductance of the coil is proportional to the cross sectional area of the band. As this cross sectional area changes the coils self-inductance changes. Changes in self-inductance are measured by integrating an oscillator circuit whose resonant frequency varies with changes in self-inductance. A counter measures this resonant frequency by counting the pulses produced by the oscillator over time, creating a waveform proportional to changes in the cross sectional area. These bands are then calibrated to adjust the contribution of each band to overall tidal volume base on Equation (1)

$$Vt = k.RC + l.AB \tag{1}$$

where Vt is the tidal volume measurement, RC and AB are the ribcage and abdominal bands, and k and l are calibration coefficients that apply the appropriate gain to each signal based on a calibration algorithm [18][19]. The LifeShirt acquires single lead ECG (lead II) sampled at a rate of 200 Hz, which is linearly interpolated in software by a factor of 5, producing an ECG with 1 kHz resolution. Heart rate is determined based on the Pan-Tompkins algorithm [20] for QRS detection. The RR interval signal is derived from the temporal R wave peak, which is determined by locating the peak point within the QRS complex. The RR signal is then processed by filters and uniformly sampled prior to spectral analysis.

B. Spectral Analysis

The instantaneous RR interval is preprocessed creating a uniformaly sampled signal. This process is illustrated in Figure 1 where the instantaneous RR interval is uniformly sampled at 50Hz and resampled at 5Hz by low pass filter decimation.

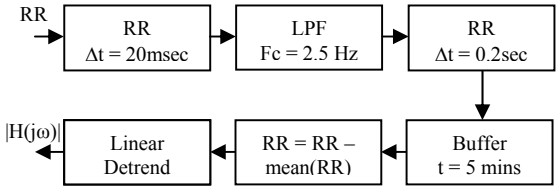

Fig. 1 Block diagram of HRV preprocessing

Five-minute time windows are detrended using a best straight line fit approach with the segment mean removed. Welch's [21] power spectral density estimation approach is then applied. Welch's averaged, modified periodogram method that applies sections of the RR signal with 50% overlap, with each section windowed with a Hamming window and nine modified periodograms are computed and averaged. A five-minute window is decomposed into 1-minute windows with 300 samples; sampled at 5 Hz. Each 1 minute window is zero padded to a 2048 sample length and the power of the FFT averaged. Various window functions may be applied, although a rectangular window will normally introduce some spectral leakage, and other window functions such as Blackman-Harris and Nutal can smear the spectra, but are also preferred in certain circumstances. The power in the LF and HF bands is determined [22] and the ratio of LF and HF calculated to evaluate sympatovagal balance.

C. Discrete Wavelet Transform

Prior to spectral analysis ectopic beats are identified and processed. The discrete wavelet transform (DWT) is used for both identification and filtering of ectopic beats. Wavelet coefficients at the highest level of detail are analyzed to locate the ectopic cycles. The DWT coefficients are obtained from the unprocessed RR interval signal by decomposing it into a set of frequency bands by applying low pass and high pass filter banks. Theses subbands are distributed logarithmically in frequency, each sampled at a rate that has a natural proportion to the frequencies in that band. Figure 2 illustrates this subband decomposition where the original signal x[n] or

 COPYRIGHT © 2005 ENFORMATIKA

in this case the instantaneous RR interval is passed through a halfband high pass filter h[n] and a low pass filter g[n] creating two subbands both sampled at half the original frequency.

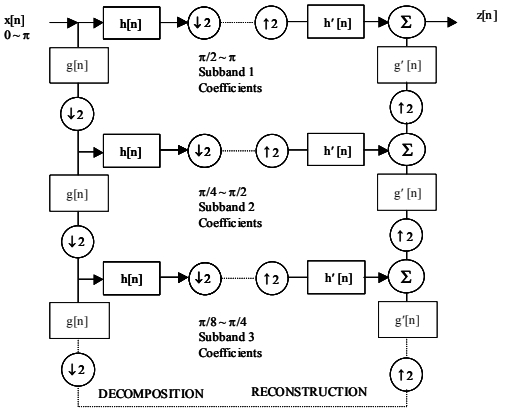

Fig. 2 Wavelet decomposition and reconstruction

These filters approximate halfband FIR filters that are determined by the choice of wavelet. The bandwidth of each filter output and subband is illustrated by a fraction of the sampling frequency where π is equal to the Nyquist rate or half the sampling frequency, therefore half of the samples can be eliminated. This demonstrates one level of wavelet decomposition, which can be expressed as:

$$y_h[k] = \sum_n x[n]h[2k-n] \tag{2}$$

$$y_g[k] = \sum_n x[n]g[2k-n] \tag{3}$$

where the g terms denote the low pass output and h terms denote the high pass output. This decomposition is reapplied to the low subband output repeatedly creating logarithmic resolution in the frequency domain. This has the effect of doubling the frequency resolution and reducing the time resolution by a factor of 2 since half the number of samples now make up the signal. This subband coding procedure is repeated for further decomposition until there is no data left to decompose, where every decomposition results in half the number of samples, thereby decreasing the time resolution by a factor of 2 and taking half the frequency band thereby doubling the frequency resolution. Only ideal halfband filters such as various wavelet packets allow for perfect reconstruction to the original signal, for example Daubechies set of wavelets [23]. Also illustrated in Figure 2 is the reconstruction process of this decomposed signal. Reconstruction normally takes places after some kind of soft or hard thresholding, or compression. The reconstruction in the case of wavelets uses the fact that these halfband filters form orthonormal bases. Therefore the decomposition procedure is followed in reverse order for the reconstruction. The wavelet subband coefficients at every level are upsampled by two, passed through synthesis filters g'[n], and h'[n] (high pass and low pass) which make up the inverse discrete wavelet transform (iDWT), and are added. The analysis and synthesis filters are identical to each other, except for a time reversal. Signal reconstruction is therefore the inverse DWT. The reconstruction process is shown on the right side Figure 2. The signal is interpolated by filling in zeros for every other sample prior to filtering thereby restoring the subband of the signal to its original length. This process is expressed by Equation (4).

$$z[k] = \sum_b h'[n]y'[2k-n] + \sum_n g'[n]d[2k-n] \tag{4}$$

D. Thresholding

Following decomposition and prior to the wavelet reconstruction process, ectopic beats are removed by a thresholding process. Soft thresholding or shrinkage requires selecting a threshold where all wavelet coefficients in each subband falling below this threshold are reduced to zero and coefficients above this thereshold have the threshold subtracted from it, so the coefficients tend toward zero. Hard thresholding requires reducing the coefficients below this threshold to zero leaving coefficients above this threshold constant. This nonlinear approach is very different from conventional filtering and has been shown to be very effective denoising images. Donoho and Johnstone [24-27] have demonstrated the usefulness of this approach against traditional linear methods of smoothing which suppress noise, but can also broaden the signals features, which can lead to error in HRV measures. This hard thresholding approach is demonstrated by Equation (5):

$$w[n] = \begin{cases} 0, & w[n] > T \\ w[n], & w[n] < T \end{cases} \tag{5}$$

where the threshold T is chosen for the higher frequency RR intervals, which will allow the removal of ectopic beats whilst retaining the signal quality. Therefore coefficients above this threshold are set zero.

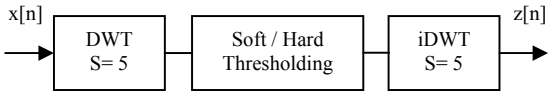

Fig. 3 Wavelet thresholding and filtering

The block diagram of Figure 3 outlines the process of filtering ectopic beats from RR interval signals using wavelet filtering. The instantaneous RR intervals are first passed through the DWT with up to 5 levels of decomposition (scales S=5) illustrated on the left side of Figure 2. Hard thresholding is applied to each wavelet coefficient for each scale or subband using Equation (5), where T is precomputed based on the average level of DWT coefficients produced from ectopic beats. Following hard thresholding the iDWT is applied to the coefficients where higher amplitude coefficients generated by ectopic beats are replaced with zero. The iDWT process is illustrated on the right side of Figure 2 showing the

 COPYRIGHT © 2005 ENFORMATIKA

reconstruction of the filtered signal. This filtering process has the effect of interpolating bad beats.

E. Interpolation

Another approach for removing ectopic beats is by interpolation. Ectopic beats are first identified by DWT and then 2 beat cycles are low pass interpolated to create a smooth signal. Ectopic beats are typically beats with a shorter cardiac cycle (coupling interval), followed by a longer cardiac cycle (post-extrasystolic pause). The ectopic beats are thus replaced by linear interpolated samples expressed in Equation (6)

$$X = \sum_{n=0}^{N-1} a_n x_N \tag{6}$$

where N = 17 and the a terms are FIR filter coefficients modeling a symmetrical filter which allows the original data to pass through unchanged and interpolates between samples so as the mean square error between them and their ideal values is minimized. The sample outputs are the last samples in the 17 long sequence going back 9 cycles and placing zeros between each cycle to be interpolated.

F. Simulation

Before applying the detection and correction procedures to real signals a simulation was designed to be roughly representative of the characteristics of real life respiratory and heart rate signals. This enables simple testing of detection accuracy. Equations (7) and (8) simulate the RR interval and tidal volume signals respectively,

$$s(t) = A\sin(2\pi f_h t + \alpha_h) + B\sin(2\pi f_l t + \alpha_l) + C \tag{7}$$

$$Vt(t) = D\cos(2\pi f_h t) + E \tag{8}$$

where A is the peak-to-peak RSA amplitude per breath expressed in ms, B is the sympathovagal balance expressed as a ratio of LF to HF power and C is the mean RR interval. Tidal volume is expressed by D, and E is a DC offset whose value depends on ribcage and abdominal cross sectional area, which varies with changes in posture and mass, as the signals are DC coupled. The parasympathetic (HF) and sympathetic (LF) components have frequencies fh , fl and phases αh, αl respectively. The signals are sampled with 1ms resolution where t = 0, 0.001..299, for a 300 second or 5 minute duration. The instantaneous RR interval signal is thus derived from Equations (9) and (10).

$$RR(0) = s(0) = C \tag{9}$$

$$RR(i) = s\left(1000\sum_{n=0}^{i-1} RR(n)\right) \tag{10}$$

The sum of all previous RR intervals determines the time of the beat that is applied to Equation (11), where i is the beat number. Each sample is then repeated for the duration of the cycle length in 20ms sample intervals providing the 50 Hz sample rate expressed as follows:

$$RR_{cont}(k) = RR(i), \quad k = i, 0.02....i+1 \tag{11}$$

III. RESULTS

Each example presented uses Daubechies p = 4 wavelet filters. Higher orders were examined with each set having similar results. The first set of results presented is based on the heart rate simulation previously described. The output is illustrated in Figure 4 for a parameter set with high and low frequencies of 0.2 and 0.1 Hz respectively with no phase variation applied. A high frequency amplitude of A = 200 ms, B = 50% for a LF/HF ratio of 1/2 and C = 1000 ms is applied. Ectopic beats were inserted randomly where the RR interval would drop significantly indicating a high heart rate for one cycle, and increase significantly above the mean for one cycle. Figure 4 shows the RR interval sampled uniformly at 50Hz and decimated to 5Hz respectively for HRV spectral analysis. The following 2 signals illustrate the instantaneous RR interval followed by a set of wavelet coefficients from the highest frequency band. Figure 5 illustrates DWT coefficients for each band, where it is obvious band 1 effectively identifies ectopic beats.

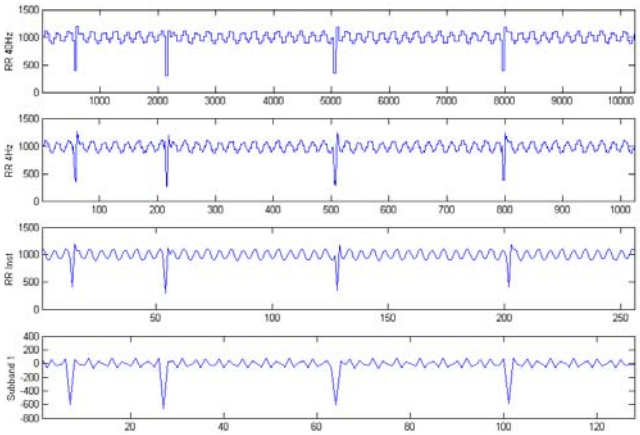

Fig. 4 Simulation with ectopic beats

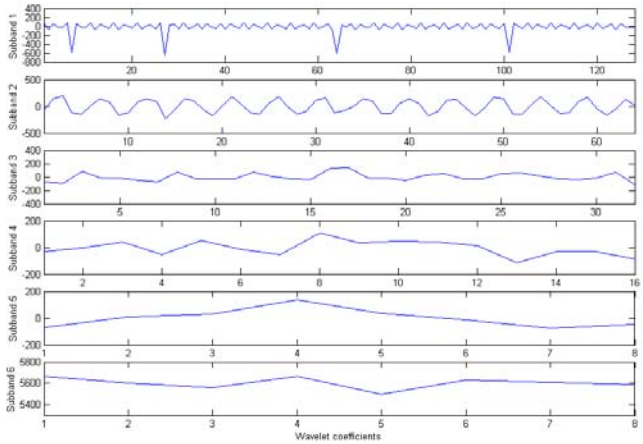

Fig. 5 Wavelet coefficients for 6 subbands

 COPYRIGHT © 2005 ENFORMATIKA

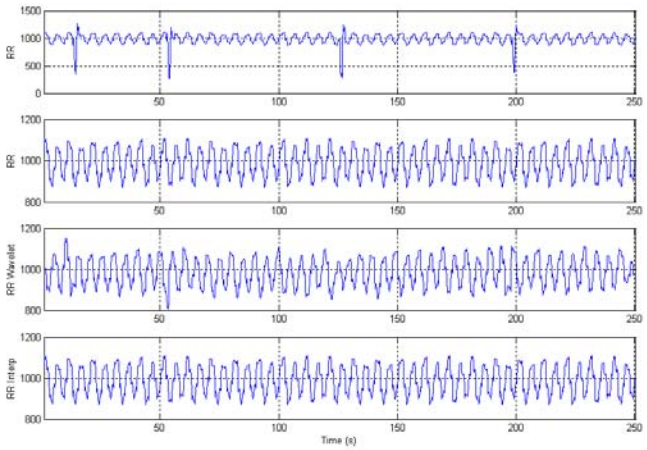

Fig. 6 Filtered and ectopic RR signals

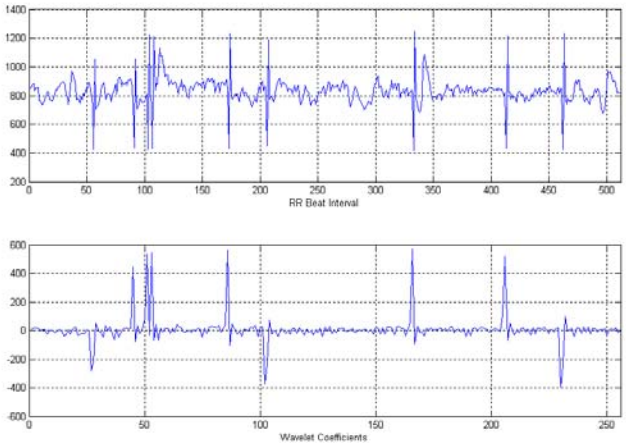

Fig. 8 RR signal with DWT coefficients

The traces illustrated in Figure 6 include the RR interval signal with ectopic beats inserted, before insertion, DWT filtered and interpolated signals respectively. All signals resemble the simulation with no ectopic beats inserted. These signals were fed through the spectral analysis routine of section 3 with the results presented in Figure 7. The first spectral plot is the RR signal with no ectopic beats added. The second spectrum is that of the DWT filter and third the simulated RR signal with ectopic beats and no filtering. The interpolated signal is not presented as it perfectly replaces simulated data. The degrading effects of ectopic beats on power spectral analysis can be seen from the third trace where there exists spectral leakage across the complete spectrum. This has the effect of greatly increasing the power in each frequency band. For instance, the LF band has increased from 39.78 ms^2 to 100.2 ms^2 and the HF band has increased from 186.84 ms^2 to 249.5 ms^2. The second spectrum presented illustrates the ability of the DWT hard thresholding approach in removing these beats from the analysis. The power in LF band has increased from 39.78 to 46.17 ms^2. The HF band power has only increased by a negligible amount. A segment of heart rate data captured from the LifeShirt from a patient suffering from heart failure is illustrated 8. There exist 9

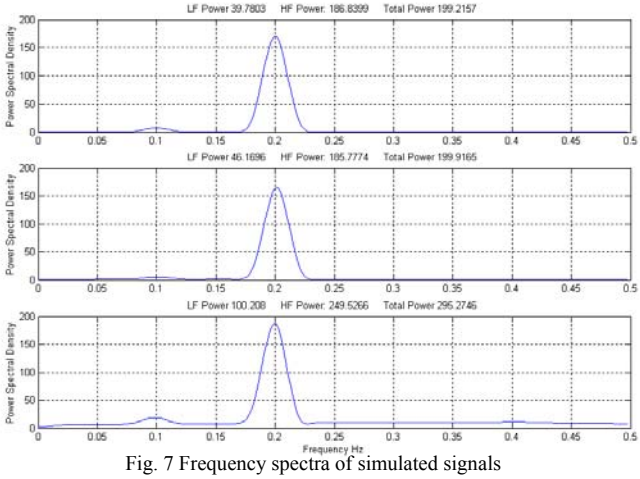

Fig. 7 Frequency spectra of simulated signals

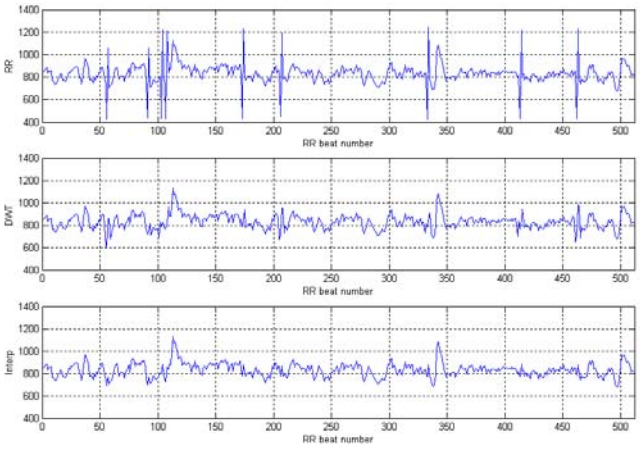

Fig. 9 RR signal with filtering and interpolation

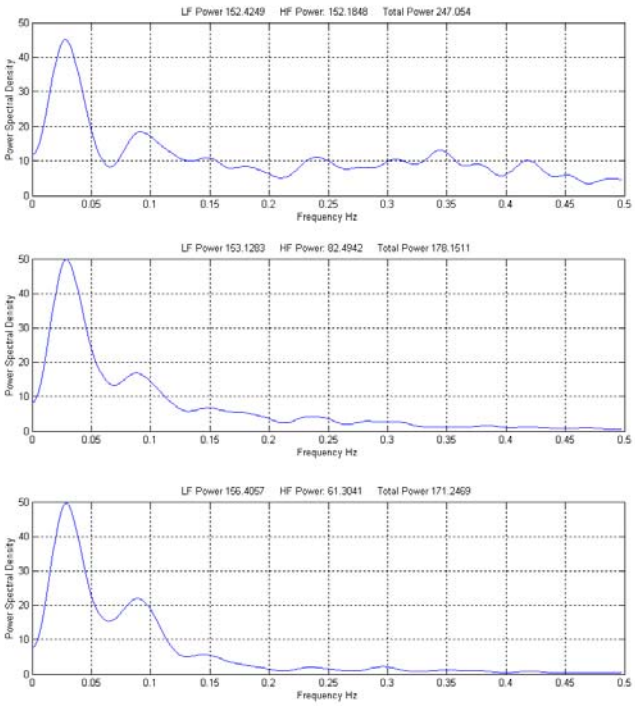

Fig.10 Frequency spectra of raw and filtered signals

 COPYRIGHT © 2005 ENFORMATIKA

ectopic beats in this segment. Large negative or positive wavelet coefficients, several times in excess of the baseline identify each ectopic beat.

The first subband, which is the high frequency detailed level, is used to identify ectopic beats. As it is the first subband and the first level of decomposition, its length is half the RR segment length. The ectopic beats are thus identified as twice the largest coefficient index and the proceeding index.

The resultant signals from applying the linear interpolation algorithm, and the DWT thresholding approach are illustrated in Figure 9. It is clear both approaches have effectively filtered the appropriate beats. The resultant spectra from the filter outputs are shown in Figure 10. The first frequency spectrum is as expected with excessive spectral leakage created by the ectopic beats, and therefore the total spectral power is overestimated. The second spectrum demonstrates the output from the DWT threshold filter. As expected, the majority of spectral leakage is minimized, as is the total power. The third spectrum closely resembles that of the DWT with the exception that there is less power in the HF band. A more difficult signal to process is illustrated in Figure 11. 512 beats are shown with an average RR interval of approximately 600 ms. Furthermore the subject is ambulatory and there exists a non-stationary baseline. There exist 11 ectopic beats in this example which can all be clearly identified in the following trace where DWT coefficients of ± 50 of the previous beat is classified ectopic. These detected beats are filtered and interpolated with both techniques and illustrated in Figure 12. The two outputs look very similar where the DWT output has smoothed more beats. Whether or not the beats are ectopic is questionable, although would still impact spectral analysis somewhat. The spectra for each signal in Figure 12 are illustrated in Figure 13. The spectra in this case doesn't appear to have had as much spectral leakage introduced by ectopic beats as the previous examples. With closer examination of the power in each frequency band, the LF band power in each example is comparable. The spectral leakage power in the HF band of the interpolated approach appears to have attenuated the most noise.

IV. Conclusion

The examples presented show the accuracy of the DWT's ability to detect ectopic beats in the higher frequency subbands. This higher frequency component occurs as an ectopic beat generates a shorter cardiac cycle (coupling interval), followed by a longer cardiac cycle (post-extrasystolic pause). Therefore, this detection technique is robust throughout the population. Although the DWT hard thresholding approach appears to adequately filter the ectopic beats from the analysis, it does, in some circumstances alter the characteristic of the natural sinus rhythmic beats adjacent to it. Although this approach is much better than nothing at all, it does have an impact on the resultant spectrum. This approach is acceptable for displaying and time-domain measures, but problematic for spectral measures. Linear interpolation has been shown to adequately replace the ectopic

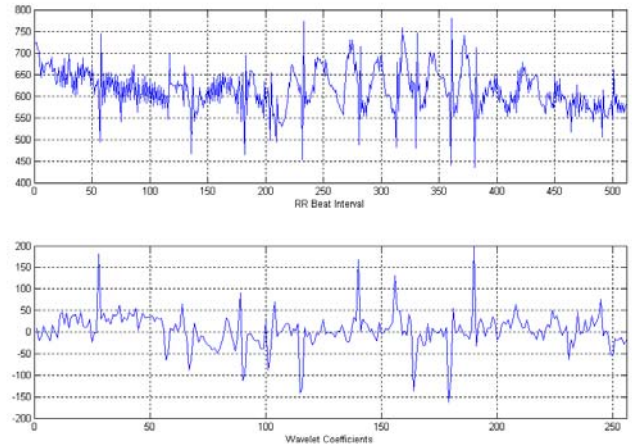

Fig. 11 RR interval with DWT coefficients

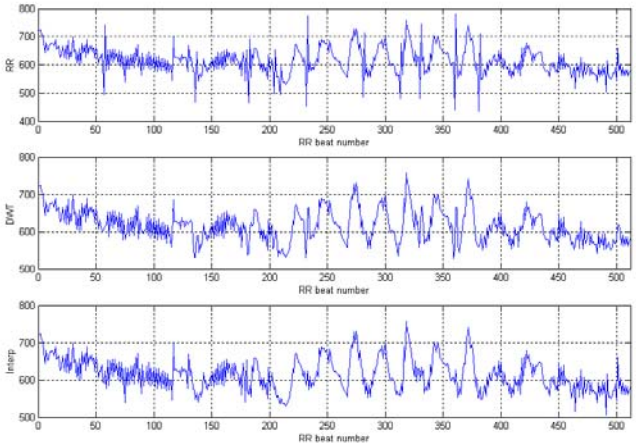

Fig. 12 RR signal with filtering and interpolation

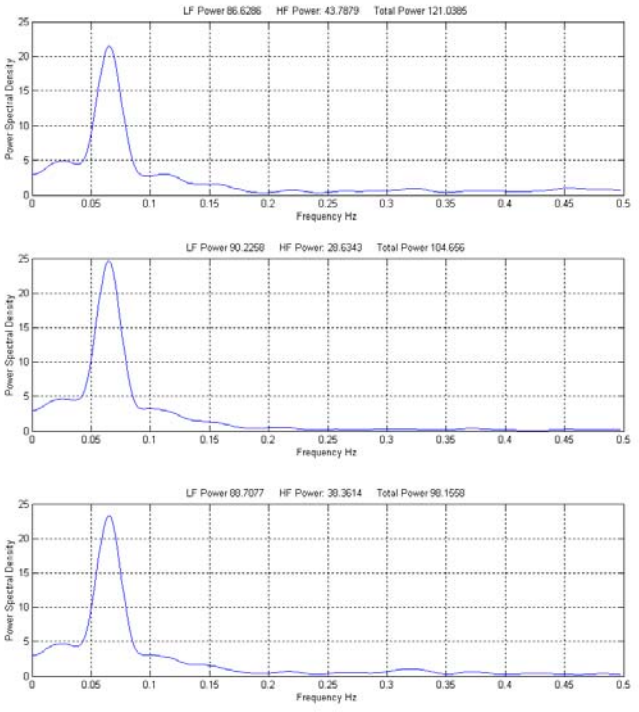

Fig. 13 Spectra of raw and filtered signals

 COPYRIGHT © 2005 ENFORMATIKA

beats with samples that can preserve the power spectrum. It is therefore concluded that the DWT detection algorithm with linear interpolation is a suitable automated combination for large datasets.

REFERENCES

[1] M. H. Crawford, S. Bernstein, P. Deedwania, "ACC/AHA guidelines for ambulatory electrocardiography," *Circulation*, vol. 100, pp. 886-893, 1999.

[2] N. Montano, T. Gnecchi Ruscone, A. Porta, F. Lombardi, M. Pagani, A. Malliani, "Power spectrum analysis of heart rate variability to assess the changes in sympathovagal balance during graded orthostatic tilt," *Circulation*, vol. 90, pp. 1826–1831, 1994.

[3] A. Malliani, M. Pagani, F. Lombardi, S. Cerruti, "Cardiovascular neural regulation explored in the frequency domain," *Circulation*, vol. 84, pp. 482–492, 1991.

[4] M. Pagani, N. Montano, A. Porta, "Relationship between spectral components of cardiovascular variabilities and direct measures of muscle sympathetic nerve activity in humans," *Circulation*, vol. 95, pp. 1441–1448, 1997.

[5] G. Paolisso, D. Manzella, N. Ferrara, et al. "Glucose Ingestion Affects Cardiac ANS in Healthy Subjects with Different Amounts of Body Fat," *American Journal of Physiology*, vol. 273, pp. E471–E478, 1997.

[6] A. Ravogli, G. Parati, E. Tortorici et al, "Early cardiovascular alterations in young obese subjects: evidence from 24-hour blood pressure and heart rate monitoring," *Europ. Heart J*., vol. 98, pp. 3405A, 1998.

[7] R. M. Carney, J. A. Blumenthal, P. K. Stein, L. Watkins, D. Catellier, L. F. Berkman, S. M. Czajkowski, C. O'Connor, P. H. Stone, K. E. Freedland, "Depression, Heart Rate Variability, and Acute Myocardial Infarction," *Circulation*, vol. 104, no. 17, pp. 2024 – 2028, 2001.

[8] P. Grossman, J. A. Van Beek, "A Comparison of Three Quantification Methods for Estimation of Respiratory Sinus Arrhythmia," *Psychophysiology*, vol. 27, no. 6, pp. 702-714, 1990.

[9] K. Kotani, I. Hidaka, Y. Yamamoto, S. Ozono, "Analysis of Respiratory Sinus Arrhythmia with Respect to Respiratory Phase," *Methods of Information in Medicine*, vol. 39, pp. 153-156, 2000.

[10] J. Mateo, P. Gascón, L. Lasaosa, "Analysis of Heart Rate Variability in the Presence of Ectopic Beats Using the Heart Timing Signal," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3, pp. 334-343, 2003.

[11] M. A. Murda'h,, W. J. McKenna, A. J. Camm, "Repolarization Alternans: Techniques, Mechanisms, and Cardiac Vulnerability," *Pacing Clinical Electrophysioly*, vol. 20, pp. 2641-2657, 1997.

[12] P. Laguna, R. Jané, P. Caminal, "Adaptive estimation of QRS complex by the Hermite model for classification and ectopic beat detection," *Medical and Biological Engineering and Computing*, vol. 34, pp. 58-68, 1996.

[13] M. A. Salo, T. Seppänen, H. V. Huikuri, "Ectopic Beats in Heart Rate Variability Analysis: Effects of Editing on Time and Frequency Domain Measures," *Annals of Noninvasive Electrocardiology*, vol. 6, no. 1, pp. 5-17, 2001.

[14] K. Konno and J. Mead, "Measurement of the separate volume changes of rib cage and abdomen during breathing," *J. Appl. Physiol*., vol. 22, pp. 407-422, 1967.

[15] J. Sackner, A. Nixon, B. Davis, N. Atkins, and M. Sackner, "Non-invasive measurement of ventilation during exercise using a respiratory inductive plethysmograph. I," *Am. Rev. Respir. Dis*., vol. 122, pp. 867-871, 1980.

[16] J. A. Adams, I. A. Zabaleta, D. Stroh, M. A. Sackner, "Measurement of breath amplitudes: comparison of three noninvasive respiratory monitors to integrated pneumotachograph," *Pediatr. Pulmonol*., vol. 16, pp. 254-258, 1993.

[17] M. A. Sackner, and B. P. Krieger, "Non-invasive respiratory monitoring," In: Heart-Lung Interactions in Health and Disease, edited by S.M. Scharf, and S. S. Cassidy, New York: Marcel Dekker, pp. 663-805, 1989.

[18] M. A. Sackner et al, "Calibration of respiratory inductive plethyomograph during natural breathing," *J. Appl. Physiol*. vol. 66, pp. 410-420, 1989.

[19] M. A. Sackner et al, "Qualitative diagnostic calibration technique," *J. Appl. Physiol.*, vol. 87, 869-870, 1999.

[20] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng*., vol. 32, no. 3, pp. 230-236, 1985..

[21] P. D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms," *IEEE Trans. Audio Electroacoustics*, vol. AU-15, pp. 70-73, 1967.

[22] A. J. Camm, M. Malik, "Heart Rate Variability: Standards of Measurement, Physiological Interpretation and Clinical Use," Task Force of the Working Groups on Arrhythmias and Computers in Cardiology of the ESC and the North American Society of Pacing and Electrophysiology (NASPE). *European Heart Journal*, vol. 93, pp. 1043-1065, 1996.

[23] I. Daubechies, "Orthogonal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, pp. 909-996, 1988.

[24] D. L. Donoho, I. M. Johnstone, "Wavelet Shrinkage: Asymptopia?," *Journal of the Royal Statistical Societ*, Series B, vol. 57, pp. 301-369, 1995.

[25] D. L. Donoho, I. M. Johnstone, "Ideal time-frequency denoising, Technical Report," Department of Statistics, Standford University, 1994.

[26] D. L. Donoho, I. M. Johnstone, "Ideal Spatial Adaptation via Wavelet Shrinkage," *Biometrika*, vol. 81, pp. 425-455, 1994.

[27] D. L. Donoho, "De-noising by Soft Thresholding," *IEEE Transactions on Information Technology*, vol. 41, pp. 613-627, 1995.

 COPYRIGHT © 2005 ENFORMATIKA

Visual Object Tracking in 3D with Color Based Particle Filter

Pablo Barrera, José M. Cañas, Vicente Matellán

***Abstract*—** This paper addresses the problem of determining the current 3D location of a moving object and robustly tracking it from a sequence of camera images. The approach presented here uses a particle lter and does not perform any explicit triangulation. Only the color of the object to be tracked is required, but not any precise motion model. The observation model we have developed avoids the color ltering of the entire image. That and the MonteCarlo techniques inside the particle lter provide real time performance. Experiments with two real cameras are presented and lessons learned are commented. The approach scales easily to more than two cameras and new sensor cues.

***Keywords*—** Monte Carlo sampling, multiple view, particle lters, visual tracking.

I. Introduction

OBJECT tracking is a useful capability for autonomous systems like ambient intelligence or mobile robotics, and even for computer-human interaction. Cameras are cheap and ubiquitous sensors. Images may provide much information about the environment, but usually it takes lot of computing power to extract relevant data from them. A basic information they may provide is the 3D location of an object or person who is moving around the camera environment.

Many commercial applications may take bene t from a robust object tracking. For instance, Gorodnichy et al [1] employ a tracking system which allow a person to use her nose as a mouse in front of a personal computer endowed with two off-the-self cameras. In security applications, unsupervised cameras may autonomously track moving persons and trigger an alarm if the person approached to any protected location.

The object tracking techniques may be also used in robotics to cope with the self-localization problem. If the mobile robot tracked the relative 3D positions of some surrounding objects, and their absolute locations are known, then it could infer its own position in such absolute frame of reference. Such objects may not be dynamic, but the robot's motion causes a relative movement which demands a tracking. Davison work [2] is a good example for this. Actually, object tracking and localization share much mathematical background with dynamic state estimation like Kalman lters, probabilistic grid based methods [3] and MonteCarlo sampling methods [4], [5]. There are also approaches to the visual 3D tracking based on genetic algorithms [6][7], similar in spirit to the sequential MonteCarlo techniques but without the probabilistic foundations.

The approach described here uses a particle lter based on the CONDENSATION algorithm [8], but in a different scenario from the contour tracking inside an image. A similar approach [9] has been recently followed to track objects inside images based on movement, color and speech cues. While our algorithm shares the 2D observations (the images), it focuses on a 3D tracking and uses only color. It requires calibrated cameras and uses projective geometry to forward project particles into all the images. The color of such projection and its neighbors provides feedback about the closeness of the particles to the real 3D location of the coloured object. A gaussian random noise is used as the motion model of the particles. Such model allows the particle population to follow any object movement.

The rest of the paper is organized as follows. Second section explains our particle lter , detailing the observation and motion models used. Third section introduces the experimental setup and some tests of the system. Finally some conclusions and future lines are sketched out.

II. Color Based Particle Filter for 3D Tracking

Our approach uses the CONDENSATION algorithm [8] to estimate location of a coloured object. This is an iterative algorithm which includes three steps on each iteration: prediction, update and resampling. CONDENSATION is a Bayesian recursive estimator which uses Sequential MonteCarlo Importance Sampling.

In short, it estimates the current multidimensional state $X(t)$, using a collection of sequential observations $[obs(t), obs(t-1), obs(t-2)..., obs(t_0)]$. The observations are related to the state through a probabilistic observation model $p(X(t)|obs(t))$. The state itself may be dynamic, and such dynamism is captured in a motion model $p(X(t)|X(t-1))$. The sequential nature of the algorithm provides iterative equations, and its sampling nature makes it to manage a set of N particles to represent the $p(X(t)|obs(t), obs(t-1)...)$. A more rigorous and broad description of probabilistic estimators can be found in [5] and [10].

Each particle $s_i(t)$ represents an state estimate and has a weight $w_i(t)$ associated, regarding the importance sampling. Global estimates can be made from the whole particle set, like choosing that of the higher weight (Maximum a Posteriori) or a weighted mean (Minimum Mean Square Error).

The *prediction step* in each iteration of CONDENSATION samples the motion model for every particle, obtaining a new $s_i(t)$, and so building a new particle set. In the *update step*, the

Manuscript received April 26, 2005. Support for this work has been provided by project DPI2004-07993-C03-01 from Spanish Education and Science Ministry

P. B. Author is supported by a research grant from Universidad Rey Juan Carlos

J.C. and V. M. Authors are with the Universidad Rey Juan Carlos, Móstoles, España (e-mail: jmplaza@gsyc.escet.urjc.es).

 COPYRIGHT © 2005 ENFORMATIKA

weights of all particles are computed following the observation model: $w_i(t) = p(s_i(t)|obs(t))$. Those particles which are likely given the current observation increase their weights. In the *resampling step*, a new set of particles is built sampling from the weighted distribution of current particles. The higher the weight, the more likely that particle appear in next set. Full details are provided at the original paper [8].

In our approach the state to be estimated is the 3D location of the object $X = [x, y, z]$, so the particles have the shape of $s_i(t) = [x_i(t), y_i(t), z_i(t)]$, they are 3D positions. The observations are just the color images on M cameras and the motion model is just a simple one that randomly move the particles through the three dimensions following a gaussian distribution for each step.

A. Movement model

Similar to [11], our approach uses a weak motion modelling, in order to accomodate to any real movement of the object. This provides robustness to the tracking algorithm as it avoids the need of a precise movement modelling to perform properly.

The motion model is a gaussian distributed one, with the same typical deviation σ_m for x, y and z axis. It follows the equations (1), (2) and (3). There is no privileged motion direction, as the object may equally move in any of them. The size of σ_m has in uence on the particle speed while walking inside the state space.

$$x_i(t) = x_i(t-1) + N(0, \sigma_m) \quad (1)$$
$$y_i(t) = y_i(t-1) + N(0, \sigma_m) \quad (2)$$
$$z_i(t) = z_i(t-1) + N(0, \sigma_m) \quad (3)$$

B. Observation model

The update step gives the new weights of the particles according to the last sensor observation. Our observation model is color based and works with any number M of cameras. It takes each camera separatedly, treates them as if they were independent observations and so multiplies all the partial conditioned probabilities. For two cameras it takes the form of (5).

$$w_i(t) = p(s_i(t)|img_1(t), img_2(t), \ldots) \quad (4)$$
$$w_i(t) = p_1(s_i(t)|img_1(t)) * p_2(s_i(t)|img_2(t)) \quad (5)$$

Each individual conditioned probability like $p_1(s_i(t)|img_1(t))$ is computed as follows. First, we project the particles into the corresponding image plane using a pinhole camera model. We assume cameras have no distortion.

- If such projection falls outside the image limits, then $p_1(s_i(t)|img_1(t)) = 1/25$
- If the projection falls inside the image limits, its vicinity is explored to count the number m of pixels with a color similar to the target color. The vicinity is a 5x5 window around projected pixel. This can be seen in Fig. 1. The equation (7) assigns a probability proportional to m. To avoid probability locks with zeroes and to tolerate occlusions, m is set to 1 when no pixel matches the target color description.

$$outside : p_1(s_i(t)|img_1(t)) = 1/25 \quad (6)$$
$$inside : p_1(s_i(t)|img_1(t)) = m/25 \quad (7)$$

The color is described in HSI space which is more robust to changes in illumination than RGB. A target color is de ned with two pairs H_{min}, H_{max} and S_{min}, S_{max}. Pixels with very low or very high intensity are silenty discarded and do not match any color description.

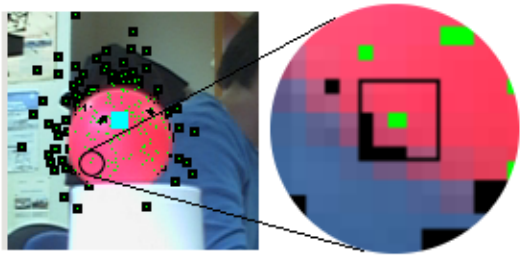

Fig. 1 5x5 vicinity window for observation model computation

The observation model in (5) clearly rewards those 3D locations which are compatible with several cameras simultaneously. In the case of two, the 3D locations compatible with one camera but which project badly in the other score poorly, because $p_1(s_i(t)|img_1(t))$ or $p_2(s_i(t)|img_2(t))$ is set to a minimum, and that keeps the $w_i(t)$ at small values. This combined reward will lead the particles to the right 3D positions.

Another advantage of this observation model is that it avoids the need to color ltering of the entire image. Depending on the number of particles this can be convenient and reduce the number computations. In our experimental setup, for instance, ltering the whole image requires 320x240 pixel evaluations and the model requires Nx25 pixel evaluations. So for $N < 3072$ it is worthwhile.

In addition the observation model doesn't require any segmentation in the images neither the search for salient points.

C. Considerations

Our approach requires calibrated cameras, but no back-projection or triangulation is performed. Only the forward projections, from 3D particles into image planes, are used. Actually, there is no matching between the stereo images, no correlation involved, and no explicit triangulation are carried out. The observation model rewards those 3D locations with are color compatible in all images. This may include more space areas than the true one, and may lead to the particle cloud to be splitted into such areas. This re ects the fact that particle lte r can represent multiple simultaneous hypothesis about the state. New observations will eventually break the ambiguity and the population will converge to the real object position.

 COPYRIGHT © 2005 ENFORMATIKA

The developed algorithm is a true multi-image algorithm [12]: there is no privileged camera, all images are treated equal and it may be used with an arbitrary number of cameras.

III. Experiments

The algorithm has been tested in our lab with two real cameras to track a pink ball. The setup is shown in Fig. 2, where camera 1 is located at $(0.5, 0.0, 0.195)$ (m) and camera 2 at $(0.07, 0.485, 0.085)$ (m) of that coordinate system. The cameras are two webcams, which have been calibrated using OpenCV library. Their external parameters like absolute position and orientation have been manually adjusted using a tape measure and projecting an absolute 3D grid into the images. They provide 320x240 color images through the video4linux API. Right camera was rotated 90^o so it delivers 240x320 images.

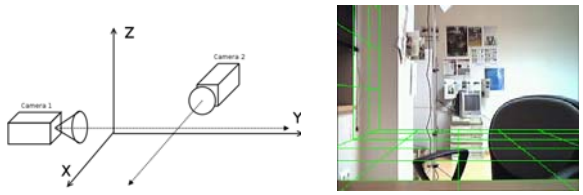

Fig. 2 Experimental setup (left) and projected grid (right)

The particle lter has been tuned to 200 particles, and a typical deviation $\sigma_m = 0.03$ (m). The vicinity window for observation model was set up to 5x5 pixels, as described previously. A typical lter iteration including all the prediction, update and resampling steps takes around 5 ms (on a Pentium IV, 2'7 GHz with HyperThreading) which is enough to real time performance.

A. Typical execution

The Fig. 3 shows a regular run of the particle lter displaying their projection in both images at three different times (iteration 2, 50 and 60). The Fig. 4 shows the projection of the same particle cloud in the XY plane.

The particles are initially located at position $(0.4, 0.1, 0.2)$ (m), just in front of the camera 1 (this initalization will be justi ed later on). In two iterations they spread following the gaussian motion model. As can be seen in the upper pair of Fig. 3, particles project around the pink ball for the left camera (camera 1), but are out of scope of the right camera (camera 2).

After 50 iterations, the particle cloud has moved itself away from the camera 1, along its optical axis and always keeping their projections around the pink ball in such camera. In Fig. 4, the typical deviation in Y (optical axis of camera 1) is greater than in X. Also, the middle pair of Fig. 3 shows some of the particles entering inside the scope of right camera while keeping around the ball in left image.

Finally, at 60th iteration the particles converge around real 3D location of the ball, and their projections into both cameras are coherent with the ball. If the ball doesn't move, the population remains stable around its real position. Once the population has converged, smooth movements of the ball are successfully tracked in any direction. It can be noticed that convergence of the whole population speeds up as soon as some particles enter into the ball projection.

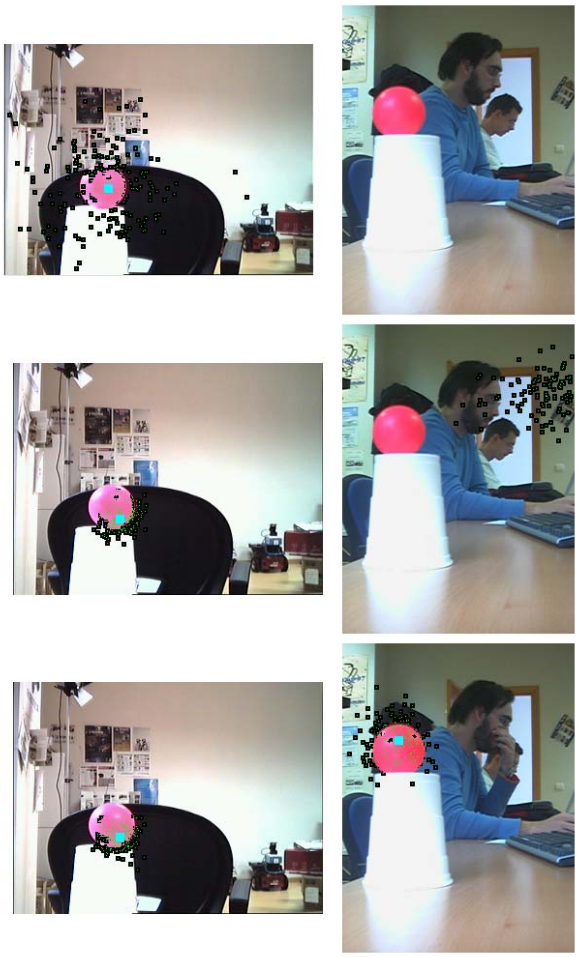

Fig. 3 Particle projections in left (camera 1) and right cameras (camera 2) at iterations 2, 50 and 60

The position error is de ned as the distance between the position estimate from the particle lter and the ground truth position of the tracked object. In all the experiments such error lied under 3 cm after the particle set has converged. Low position error means that the particle cloud has settled around the right location.

B. Systematic drift

We have observed a systematic pernicious trend of the particles to move far away from the cameras along their optical axis. This can be noticed in the Fig. 4. Experiments were also carried out with random initalization, and starting the particles at a point further than the ball to a given camera. All such runs resulted in no convergence at all: the systematic drift evolved the particles cloud consistently with one image, but always moving away from the other.

 COPYRIGHT © 2005 ENFORMATIKA

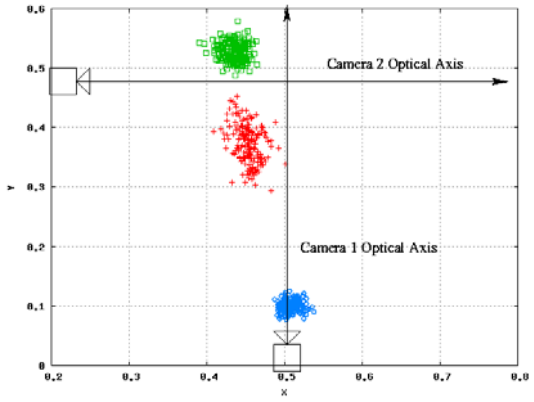

Fig. 4 Three eyebird snapshots of the particles at iterations 2, 50 and 60

Given the conic shape of the projective observation model, after the prediction step there is equal chance to fall closer or further to the camera than the current location, but falling closer makes less likely to project inside the image, makes harder to achive a good observation likelihood, and so, smaller the chance of surviving after the resampling. This can be seen in Fig. 5.

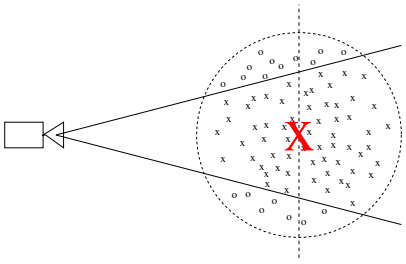

Fig. 5 Particle and its likely positions at $t + 1$

C. Particle filter dynamics

In a different set of experiments we have studied the particle lter dynamics, in particular we were interested in its convergence speed. It was studied measuring the evolution of the position error of the lter when the tracked objet suddenly moved to a different location. Instead of the 2D observations described we used the pink ball 3D position estimate as observations for the particle lter . A 3D gaussian observation model was developed for such observations. In real experiments such estimates were built triangulating the centers of masses of the pink pixels on each camera image.

In Fig. 6 the evolution of the position error is displayed. The vertical axis means the position error in (m), and the horizontal axis represents time, in iterations of the lter . The object is stable at the initial position until 30th iteration, and then moves to a new location 2 meters away. To avoid the effect of noise in the observations while studying the lter dynamics, in this experiment they were simulated and set to the ground truth 3D position of the pink ball.

As can be seen in Fig. 6, it takes 5 iterations to the lter to converge at the initial position of the object. At the 30th

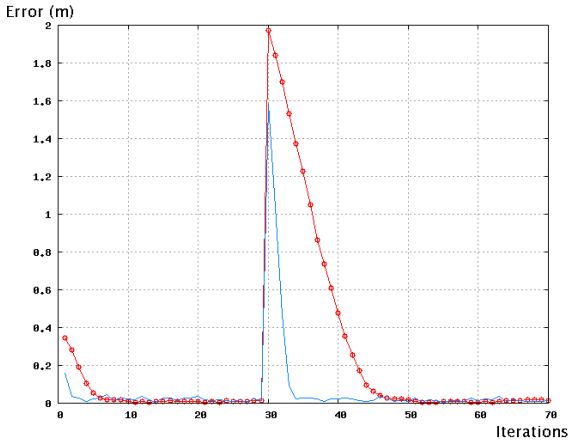

Fig. 6 Time evolution of the position error

iteration the error reaches 2 meters, just the distance that the object has moved. Then the lter needs around 15 iterations more to converge at the new object location.

The convergence speed of the particle lter has been studied for different σ_m values, in order to determine the right value for such parameter. In Fig. 7 the evolution of the position error is displayed. The horizontal axis represents the time, from the initial iteration at the left and increasing number of iterations to the right. The vertical axis represents different values of σ_m, from 0.001 (m) to 0.3 (m) at regular increments upwards. The pixel color represents the position error, the darker the higher. Colors close to black mean the lter estimation of position is far away from real one, colors similar to white mean they are close. The tracked object moved suddenly at 30th iteration. This way, each row shows an evolution like that in Fig. 6, but encoded in color.

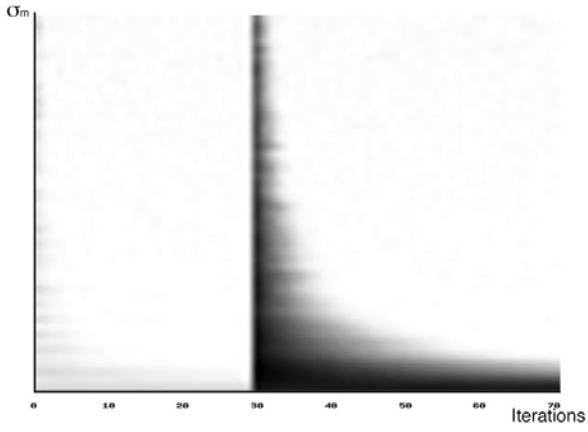

Fig. 7 Evolution of position error at different σ_m values

The experiments in Fig. 7 show that high values of σ_m speed up the convergence. The time needed to nd the new object location is the dark gap after the 30th iteration, inside each row. As σ_m increases, such time asymptotically decreases. For very low values of σ_m, the particle lter is not able to nd

 COPYRIGHT © 2005 ENFORMATIKA

the new object position before 70th iteration, as can be seen in the left side of the Fig. 7.

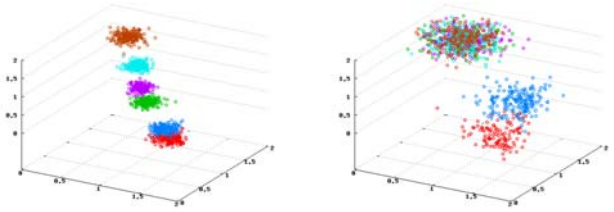

Fig. 8 Spatial evolution of the particle set

But such improvement in convergence speed does not come for free. We have observed that for high values of σ_m the typical deviation of the particle cloud increases. In Fig. 8 the evolution for two particle set is displayed, with snapshots of the particle clouds at six different instants. The same color means the same iteration in both lters. With small σ_m values (left), the population slowly approaches to the location of the object, keeping itself compact. With high σ_m values (right) the cloud reaches sooner the new position of the tracked object, but it is spread over a wide area.

IV. Conclusions and Future Lines

The work presented here summarizes the preliminary results on particle lter for object 3D tracking based on color information. The algorithm doesn't need any explicit triangulation or stereo matching at all, and it scales to an arbitrary number of cameras. The observation model used avoids the color ltering of the whole images and looks at the vicinity of the particle projections to estimate the particle's likelihood.

The results are promising as convergence has been validated in real experiments and the algorithm implementation exhibits real time performance. The real location of the object is an stable point for the particle cloud, and the particles successfully track smooth movements of the object. An interesting systematic drift in the particle behavior has been discovered and explained.

The experiments carried out are just a proof of concept. More experiments are necessary in order to validate the algorithm. Further improvements of the algorithm are coming. First, the use of more than 2 cameras simultaneously, in order to expand the volume inside which objects are succesfully tracked. Second, we are also exploring some proposal distributions inside the lter which hopefully would increase convergence speed of the cloud and its recovery capacity in case of losing the object.

References

[1] D. Gorodnichy, S. Malik and G. Roth, *Affordable 3D face tracking using projective vision*, Proc. of Int. Conf. on Vision Interface, pp. 383-390, Calgary (Canada) May 2002.

[2] A.J. Davison, Y. González-Cid and N. Kita, *Real-time 3D SLAM with wide-angle vision*, 5th IFAC/EURON Symposium on Intelligent Autonomous Vehicles, July 2004.

[3] D. Margaritis and S. Thrun, *Learning to locate an object in 3D space from a sequence of images*. Proc. of Int. Conf. on Machine Learning, pp. 332-340, 1998.

[4] D. Fox, W. Burgard, F. Dellaert and S. Thrun, *Monte Carlo localization: efficient position estimation for mobile robots*, In Proc. of 16th. AAAI Nat. Conf. on Arti cial Intelligenge, pp. 343-349, Orlando (USA), July 1999

[5] S. Arulampalam, S. Maskell, N. Gordon and T. Clapp, *A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking*, IEEE Transactions on Signal Processing, vol. 50, no. 2, pp. 174-188, 2002.

[6] A.M. Boumaza and J. Louchet, *Mobile robot sensor fusion using flies* In Günter Raidl et.al. editors, Applications of Evolutionary Computing, EvoWorkshops 2003, Lecture Notes in Computer Science, vol. 2611, pp. 357-367, Springer, 2003.

[7] Jean Louchet, *Using an individual evolution strategy for stereovision*, Genetic Programming and Evolvable Machines, vol. 2, pp. 101-109, 2001

[8] M. Isard and A. Blake, *CONDENSATION- conditional density propagation for visual tracking*, Int. Journal of Computer Vision, vol. 20, no. 1, pp. 5-28, 1998.

[9] P. Pérez, J. Vermaak and A. Blake, *Data fusion for visual tracking with particles*, Proceedings of IEEE, vol. 92, no. 3, pp. 495-513, March 2004.

[10] D. Mackay, *Introduction to Monte Carlo methods*, In M. Jordan editor, Learning in graphical models, pp. 175-204, MIT Press, 1999

[11] A. J. Davison, *Real-time simultaneous localisation and mapping with a single camera*, IEEE Int. Conf. on Computer Vision, ICCV-2003, pp. 1403-1410, Nice (France), October 2003.

[12] R.T. Collins, *Multi-image focus of attention for rapid site model construction*, IEEE Int. Conf. on Computer Vision and Pattern Recognition, 1997, pp. 575-581, San Juan, Puerto Rico, June 1997.

 COPYRIGHT © 2005 ENFORMATIKA