Intelligent GPS-based predictive engine control for a motor vehicle

S.H. Lee∗, S.M. Begg, S.D. Walters and R.J. Howlett
Centre for SMART Systems, Engineering Research Centre, University of Brighton, Moulsecoomb, Brighton, BN2 4GJ, UK

Abstract. An intelligent Global Positioning System (GPS) based control system utilises information about the current vehicle position and upcoming terrain in order to reduce vehicle fuel consumption as well as improve road safety and comfort. The development of such in-vehicle control systems has provided static and dynamic road information. The vehicle running parameters have been mathematically defined whilst the engine control algorithms were derived from a custom-built engine test-rig. As the vehicle travelled over a particular route, road information such as gradient and position was stored with the past trajectory using a Neuro-Fuzzy technique. This road information was continuously updated and replaced by new data as the vehicle moved along, thereby adjusting the engine control parameters to reflect the actual current vehicle running data. The control system essentially used a fuzzy logic derived relief map of the test route and this was further validated and corrected based on the past trajectory from the in-vehicle GPS sensor. The simulation model demonstrated the feasibility and robustness of the control system for motor vehicle control applications.

1. Introduction

The UK and Europe have many major cities and surrounding villages which have an excessive level of congestion, often making travel difficult. The exhaust emissions and noise of the increasing number of vehicles going in and around the city add to the problem. In urban areas, the use of fuel-efficient public transport e.g. buses, could be a medium term solution for the improvement of traffic and more particularly for a healthier environment.

Buses spend most of their time on journeys around town which involves a lot of stop-and-go events. Most of these journeys are for example, between the railway station and the town centre. The current state of advancement and availability of vehicle positioning and sensor technology means that the Global Positioning System (GPS) can be used to accurately pinpoint the location of the vehicle. This GPS information provides a reference of the road ahead and thus could be used to predict the future direction and elevation of the vehicle; this information could then be used for dynamically tuning the engine and/or giving control commands to the Engine Management System (EMS) leading to reduced fuel consumption and reduced emissions.

This work has been co-funded by the Interreg IIIa European research programme, and has been carried out by the University of Brighton, Centre for Smart Systems ‘Intelligent Vehicle On-board System (VBIS)’. It was focused on the development of engine control systems using the fusion of externally-acquired positioning data and engine operating parameters; also from analysis and optimisation of these data using intelligent techniques and tools for application to motor vehicles. These are expected to facilitate the reduction of exhaust emissions and fuel consumption through precise control of the combustion process.

1.1. Background

Embedded control design has been widely applied in automotive systems such as vehicle and engine control. Such a system is dedicated to specific tasks; design engineers can optimise, reducing the size and cost of the product, and/or increasing the reliability and perfor-
mance. Some embedded systems are mass-produced, benefiting from economies of scale. Typically, several complex algorithms are running in this embedded system most of them are based on operating information of the vehicle. Thus a reliable estimation of running parameters is very important in determining that the control regime consequently improves vehicle performance. The vehicle mass, road gradient and air drag are essentially the major factors to influence a vehicle’s performance. These parameters are significant in the case of buses due to their weight and particularly large frontal area. Many modern vehicle control systems consist of engine, transmission, brake and auxiliary functions. There are large numbers of dedicated algorithms in these sub-systems ranging from pure control task to running resistance estimation. To develop such a system with precision requires time and knowledge, increasing complexity in some systems making design and development of such system even more difficult. Hardware-in-the-loop (HIL) simulation is a technique that is used in the development and testing of complex real-time embedded systems. HIL simulation provides an effective platform by adding the complexity of the system under control to the test platform. The complexity of the system under control is included in test and development by adding a mathematical representation of all related dynamic systems.

The use of externally acquired information such as GPS data is believed to be useful in engine and vehicle control. It has been increasingly used in real-time tracking of vehicles, especially when GPS is integrated with ever increasingly powerful Geographic Information System (GIS) technologies. The accuracy and reliability of low-cost, stand-alone GPS receivers can be significantly improved to meet the technical requirements of various transportation applications of GPS, such as vehicle navigation, fleet management, route tracking, vehicle arrival/schedule information systems (bus/train) and on-demand travel information. Systems that were previously only intended for fixed installation in vehicles are gradually being replaced on the market by portable systems that require no connection to the vehicle other than the power supply. To an increasing extent, GPS navigation is becoming a software product that can also be installed on handheld computers, laptops and mobile phones.

Global positioning determination is based primarily on the use of GPS. Stand-alone systems, such as handheld computers, use this exclusively, whereas fixed installation systems also run ‘dead reckoning’ if they have additional in-vehicle sensors. Dead reckoning ensures exact position determination even if no GPS signals can be received, e.g. in tunnels. To measure the distance travelled, all that is needed is a speedometer output signal. The change of direction is ascertained by a rotation rate sensor or gyroscope. Hence, the absolute direction of travel can be determined by the Doppler effect of the GPS signals [1]. The levels of accuracy that can be achieved is in the range of 3 to 5m, and 10 to 20m in the case of measuring altitude relative to sea level. With the autonomous European Satellite Navigation System Galileo, an opportunity of a joint system ‘GPS + Galileo’ with more than 50 satellites will provide many advantages for civil users and vehicle systems, in terms of availability, reliability and accuracy [2].

Future GPS may not only be used to guide the vehicle but information from the system may also be used to control or influence the engine, through given control parameters in a safe and cost-effective manner. A GPS receiver provides reliable reference position data which can be manipulated to provide more significant road information such as gradients or even road traffic congestion updates when it is combined with the vehicle telematics. It is a technology integrated with computers and mobile communications technology in vehicle navigation systems. This information can be used to not only inform the driver but also to enhance the control of several systems of the vehicle. Ultimately, for example, the vehicle speed, gear selection and even the application of brakes could be appropriately chosen and strategically designed. The idea is to provide the control system with this essential information that the driver normally uses when driving. Good driving requires consideration of several inputs. It can be a complex, exhausting and demanding task, even for commercial vehicle drivers and thus supporting control functionality is of great interest. It is believed to be even valuable to obtain road information beyond the line of sight of the driver. Whilst all of these driving decisions have to be made manually by the driver in the interest of comfort and fuel efficiency, the newly intelligent vehicle controller aims to address these tasks.

This work towards sustainable transportation requires dramatically reduced fuel consumption and emissions. This project has addressed the issues and challenges imposed by legislation and guidelines with the aim of facilitating the reduction of exhaust emissions and fuel consumption through precise control of the vehicle. Techniques include the fusion of data from sources that are external as well as internal to the ve-
hicle; also from analysis of these data using special intelligent systems techniques and tools. The resulting system essentially used a fuzzy logic derived relief map of the test route, and this was further validated and corrected based on the past trajectory from the GPS sensor. The information was then processed and translated in order to estimate the future elevation of the vehicle. Similar techniques based on predictive parameters have been proven useful and achieved better results. Model Predictive Control (MPC) is an optimisation algorithm which has shown that a 2.5% reduction of fuel consumption can be achieved by controlling the speed of a vehicle. The control signals were; percentage of throttle opening, activation of brakes and gear selection. The control algorithm was tuned and optimised according to some criteria, e.g. the main issues were to minimise costs, time and fuel consumption [6]. Similar work described in another publication has been designed and simulated on cruise control [7]. The simulation showed that a reduction of fuel consumption in the range of 1.5 to 3.4% was achieved. It used a dedicated logic in a finite number of simulated driving situations, given that the topography of the road such as gradient was a known input to the system. Control of the vehicle powertrain has been undertaken by DaimlerChrysler; the research suggested usage of a three-dimensional digital road map in order to let the cruise control replicate a skilled driver [8]. A reduction in fuel consumption of 4.1 to 5.2% was attained. Furthermore, cruise control has now been incorporated with radar technology to record the distance and speed relative to the vehicle in front as well as additional data such as position of other vehicles in the vicinity. The system used such information to regulate the time gap between vehicles. The interface was developed in the European project MAPS&ADAS to obtain the map data from the on-board data provider [8]. This is an advance system which adapts the speed to the surrounding vehicles and keeps a safe distance.

All in all, a number of approaches have been researched. A substantial amount of work has been carried out on how the interface between vehicle control system and the GPS system should be designed. The investigation was focused on information retrieval and processing. Location data could be available to the vehicle control unit in a variety of formats, resolutions and temporal accuracies. Data processing and fusion forms the main part of this project. This information was made available and able to combine with other sensory data of the vehicle.

The following sections describe the project and report on experimental results. The simulation model generated using Matlab/Simulink showed the effectiveness of the system. Simulink is a software package for modelling, simulating, and analyzing dynamic systems. It supports linear and nonlinear systems, modelled in continuous time, sampled time, or a hybrid of the two. It offers a graphical user interface for creating block diagram models. A system is configured in terms of block diagram representation from a library of standard components. In the middle of a simulation, algorithms and parameters can still be changed to get intuitive results, thus providing the user with a readily accessible learning tool for simulating many of the operational problems found in the real world. It also provides immediate access to the mathematical, graphical, and programming capabilities of Matlab. The effective engine/vehicle control system devised using Simulink could potentially be used in vehicle control for reduced fuel consumption and emissions.

2. Neuro-fuzzy techniques

Intelligent systems, i.e. software systems incorporating artificial intelligence, have shown many advantages in engineering system control and modelling. They have the ability to rapidly model and learn characteristics of multi-variate complex systems, exhibiting advantages in performance over more conventional mathematical techniques. This has led to diverse applications in power systems, manufacturing, optimisation, medicine, signal processing, control, robotics, and social/psychological sciences [3,4]. In an industrial automation and process control, fuzzy logic technologies enable the efficient and transparent implementation of human control expertise. For example, an individual control loop of a single industrial process has variables mostly controlled by conventional models such as proportional-integral-derivative (PID). A fuzzy logic system can then give the set values for these controllers based on the process control expertise put in the form of fuzzy logic rules. In Japan, Germany and France, cars with intelligently controlled components are quite common; the reasons being the control systems in cars are complex and involve multiple parameters. The optimisation of these systems is based on engineering expertise rather than mathematical models. Criteria such as ride-comfort and handling are optimisation goals that cannot be defined mathematically.

Adaptive Neuro-Fuzzy Inference Systems (ANFIS), developed in the early 1990s by Jang [5], combine the concepts of fuzzy logic and neural networks to form
a hybrid intelligent system that enhances the ability to automatically learn and adapt. Hybrid intelligent systems have been used by researchers for modelling and prediction in various engineering systems. The basic idea behind the neuro-adaptive learning techniques is to provide a method for the fuzzy modelling procedure to learn from a data set, in order to automatically compute the membership function parameters that best allow the associated Fuzzy Inference System (FIS) to track the given input/output data. These parameters associated with the membership functions will change through the learning process similar to the update of weights in a classic neural network. Their adjustment is facilitated by a gradient vector, which provides a measure of how well the FIS is modelling the input/output data for a given set of parameters. Once the gradient vector is obtained, any of several optimisation routines can be applied in order to adjust the parameters so as to reduce error between the actual and desired outputs. This allows the fuzzy system to learn from the data it is modelling.

ANFIS largely removes the requirement for manual optimisation of fuzzy system parameters. A neural network is used to automatically tune the system parameters, for example the membership function shapes, leading to improved performance without operator intervention. In addition to a purely fuzzy approach, an ANFIS was also developed for the estimation of spray penetration because the combination of neural network and fuzzy logic enables the system to learn and improve its performance based on past data. A Neuro-Fuzzy system with the learning capability of a neural network and with the advantages of rule-based fuzzy systems can improve the performance significantly and can provide a mechanism to incorporate past observations into the classification process. In a neural network, the training essentially builds the system. However using a Neuro-Fuzzy scheme, the system can be initialized with an initial fuzzy model and then refined using neural network training algorithms.

3. Vehicle model

The model developed using Simulink was based on a Volkswagen Golf with a standard gasoline-powered internal combustion engine. This model has formed a core part of the simulation intending to predict the amount of torque required to balance the loads exerted on the vehicle. This model consisted of two major input components, the engine speed and the predictive gradient of the road. The outputs from the model are the ignition timing and the engine torque. The objective here is to control the ignition timing so that the engine is operated in its optimised condition. Ignition timing in an internal combustion engine is the process of setting the time that a spark will occur in the combustion chamber relative to piston position and crankshaft angular velocity. Setting the correct ignition timing is crucial in the performance of an engine. The ignition timing affects many variables including engine longevity, fuel economy and engine output.

3.1. Vehicle dynamics

Consider a vehicle with a mass \( m \) travelling on a road with an incline \( \theta \) showed schematically in Fig. 1. The total force acting on the vehicle is simplified to the sum of the driving force generated by the engine, the air drag and the gravitational force. The rolling resistance between the road and the tyres was assumed to be zero. The car is travelling in a straight line and must maintain constant speed with change in \( \theta \).

Applying Newton’s second law, the resultant motive force, \( F_m \), on the vehicle is given by the resolved component parallel to the slope:

\[
F_m = ma = F_{\text{engine}} - F_{\text{drag}} - mg \sin \theta
\]  

(1)

where \( m \) is the vehicle mass, \( a \) is the acceleration of the vehicle, \( F_{\text{engine}} \) is the driving force produced by the engine and \( F_{\text{drag}} \) is the resistance due to aerodynamic drag.
Given that the engine controller was to maintain the vehicle at constant speed regardless of the change in road gradient as shown in Fig. 2. In the case of uphill scenario, the controller adjusts the ignition timing according to the loads i.e. air drag and gravity, and associated with the advance road gradient derived from the predictive algorithm.

For a vehicle travelling at constant speed, i.e. $a = 0$, Eq. (1) is reduced to:

$$F_{\text{engine}} = F_{\text{drag}} + mg \sin \theta$$

The vehicle drag force is given by:

$$F_{\text{drag}} = \frac{1}{2} \rho v^2 AC_d$$

The gross indicated power, $I_p$, is given by:

$$I_p = \left( \frac{\text{GIMEP} \times V_h}{2\pi} \right) \dot{\theta}$$

By rearranging and substituting relevant parameters in the Eqs (2), (3) and (4), the power balance of the system is given by:

$$\left( \frac{\text{GIMEP} \times V_h}{2\pi} \right) \dot{\theta} = \left( \frac{1}{2} \rho v^2 AC_d + mg \sin \theta \right) v$$

where GIMEP is the gross indicated mean effective pressure of the engine, $V_h$ is the engine capacity, $\dot{\theta}$ is the engine speed, $v$ is the speed of the vehicle, $\rho$ is the air density, $A$ and $C_d$ are the frontal area and the drag coefficient of the vehicle, respectively.

The term $I_p$, gross indicated power of an engine is the theoretical power of an internal combustion engine, consider that it is completely efficient in converting the energy contained in the expanding gases in the cylinders. The term GIMEP is effectively torque without losses. Losses and efficiency can be built into a more complex model. An essential part of this vehicle mod-

el however is to obtain optimised engine operating parameters in order to achieve reduced fuel consumption and emissions. As the GIMEP is directly related to the engine control parameters ‘ignition timing’ and ‘engine operating speed’, a series of engine tests were carried out so that the optimised control map could be obtained and included in the simulation model.

### 3.2. Test engine

A single-cylinder Ricardo Mk I Hydra engine was built using a production B230 Volvo cylinder head, cut from a multi-cylinder head. Modifications to the camshafts, the oil and water systems were carried out. An intake manifold was fabricated from the multi-cylinder manifold (including the injector boss) to fit the cylinder head, throttle body and intake plenum. The fuel injector was directed down the intake port. The angle of the injector to the machined gasket face of the cylinder head was the same as the production engine configuration. A low-pressure fuel rail was manufactured to fit the injector. An exhaust manifold was fabricated to fit the cylinder head with bosses for a lambda sensor and thermocouple. The basic engine specifications are given in Table 1.

#### 3.2.1. Test bed and installation

The engine was installed on a test bed in the Sir Harry Ricardo Laboratories at the University of Brighton. The facility was equipped with a Plint dynamometer and electrically-driven pumps for oil, coolant and fuel supplies. The oil and coolant temperatures were maintained at 80°C and 90°C ± 2°C respectively. Oil, coolant, fuel, intake air and exhaust gas temperatures were recorded using type-K thermocouples. Intake manifold pressure was recorded using two, 2 bar absolute pressure transducers (Kistler 4045A2, Druck DPI 201). In-cylinder pressure was recorded using a gauge pressure transducer (Kistler 6125).

The rotational speed of the engine was measured using an optical encoder (Leine and Linde) with a resolution of 720ppr directly coupled to the crankshaft. The engine speed was maintained to an accuracy of ± 5 rpm. Mass flow of air through the engine was measured using a thermal mass flow meter (Endress and Hauser, AT70F). The minimum flow measurement (and the greatest uncertainty) was approximately 4 kg/hr. It was not possible to record the mass flow of air for engine speeds less than or equal to 1000 rpm. The throttle valve was driven by a geared stepper motor. The throttle position was controlled with a multi-turn potentiometer.
### Table 1

<table>
<thead>
<tr>
<th>Engine specification</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bore</td>
<td>92 mm</td>
</tr>
<tr>
<td>Stroke</td>
<td>80 mm</td>
</tr>
<tr>
<td>Number of cylinders</td>
<td>1</td>
</tr>
<tr>
<td>Head</td>
<td>one cylinder of B230 4 valve multi-cylinder</td>
</tr>
<tr>
<td>Compression ratio</td>
<td>10:1 (nominal)</td>
</tr>
<tr>
<td>Maximum valve lift</td>
<td>9.8 mm</td>
</tr>
<tr>
<td>Ignition system</td>
<td>Brighton with Mitsubishi coil-on-plug</td>
</tr>
<tr>
<td>Coil charge duration</td>
<td>3.4 ms</td>
</tr>
<tr>
<td>Spark plug</td>
<td>NGK BP8EVX</td>
</tr>
<tr>
<td>Spark plug gap</td>
<td>0.85 mm</td>
</tr>
<tr>
<td>Fuel injection system</td>
<td>Brighton with Bosch injector</td>
</tr>
<tr>
<td>Fuel pressure</td>
<td>3.5 bar</td>
</tr>
<tr>
<td>Injection timing</td>
<td>90 CA BTDC firing (F)</td>
</tr>
</tbody>
</table>

Air to fuel ratio (AFR) was measured close to the exhaust port using a calibrated wide-range lambda sensor (ETAS LA3). The AFR is often defined in terms of the excess air factor, or lambda. Lambda is defined such that a lambda factor of unity corresponds to an AFR of 14.7:1 at normal temperature and pressure. This is termed the stoichiometric ratio, corresponding to the proportions of air and fuel which are required for complete combustion. A greater proportion of fuel gives a lambda of less than unity, termed a rich mixture, while a greater proportion of air gives a lambda of greater than unity, termed a weak or lean mixture. The calibration of the sensor was checked periodically against a Horiba MEXA 7170DEGR exhaust gas analyser.

The fuel rig comprised of production automotive components (tank, regulator, and pump) integrated within a standalone unit with provision for fuel cooling. The low pressure part of the circuit was used for the Port Fuel Injection (PFI) injector. The fuel rail was fabricated from a modified Bosch production fuel rail. The fuel used throughout was pump grade BP 95 RON unleaded gasoline.

An AVL INDISET 620 data acquisition system and INDICOM V1.5 software were used to record data for combustion analysis over 400 consecutive cycles with a resolution of 0.5°CA. The in-cylinder gauge pressure was not pegged to the intake manifold absolute pressure conditions. The AVL thermodynamic correction was applied to in-cylinder pressure. In addition to in-cylinder pressure, the data acquisition system was used to record AFR, intake manifold absolute pressure, air mass flow rate, ignition timing signal, injection timing signal and engine coolant, air, exhaust and oil temperatures. The greatest error in the engine load condition was ± 0.04 bar recorded at the lowest speed condition.

All instrumentation was calibrated prior to engine testing and periodically throughout the programme. Before each test run, a hot, motored TDC determination was performed. The engine test installation is shown in Fig. 3.

#### 3.2.2. Programme of work

The operating points for engine speed and load were selected to be representative of the conditions typically encountered during city driving between 1st and 4th gear. The engine was tested using the baseline production operating parameters for three speed and part load conditions: 1000 rpm and 1.0 bar GIMEP, 1500 rpm and 1.5 bar GIMEP, and 1800 rpm and 1.8 bar GIMEP. At each part load condition, a mixture response swing was completed using MBT ignition timings (minimum ignition advance for best torque). Fuel was injected at 90°CA BTDC firing (CVI – Closed-valve injection).
The spark plug gap was optimised for the ignition system used.

### 3.3. Engine mapping

The data obtained from the instruments were manipulated and analysed. An engine control map was generated using a Neuro-Fuzzy modelling approach whereby two input and one output parameters were configured in the FIS. This Neuro-Fuzzy system automatically adjusted the parameters of the basic fuzzy logic system very efficiently and identified the unknown process mapping from input to output data.

Engine parameters were collected from 1000 rpm to 1800 rpm; ignition timing was recorded at each operating point where the AFR was monitored and maintained at 1.0 throughout, ensuring optimal combustion efficiency [9]. Small reductions in AFR can optimise power output but may lead to sizeable increases in emissions. Figure 4 depicts a three-dimensional plot that represents the resulting Neuro-Fuzzy mapping of engine speed, GIMEP and ignition timing. It can be seen that as the engine speed and GIMEP increases, the ignition timing increased in a non-linear piecewise manner, this being largely due to non-linearity of the characteristic of the input vector matrix derived from the raw engine data. This assumed that these raw engine data are fully representative of the features of the data that this Neuro-Fuzzy FIS was intended to model.

By combining the optimised engine operational map with Eq. (3), the primary aim of the controller was to maintain the ignition timing with respect to the predictive gradient of the road, \( \theta \). This has formed a major part of this engine sub-system in the simulation.

### 4. Fuzzy predictive model

#### 4.1. Techniques

A challenge of the project was how to meet higher safety standard requirements and obtain reduced fuel consumption through the use of live GPS road information for vehicle control. The approach was to use GPS to track the vehicle, and also to create the base map. At all other times GPS readings were used to validate or correct the base map when a reliable signal was available, and of sufficient accuracy. The correct vehicle position was achieved by tracing this GPS signal received at a predetermined time interval.

The conceptual overview of this model is shown in Fig. 5. The control system acquired the position data through the serial interface, so that it could be used to improve the operation of several sub-systems in the vehicle, e.g. controlling a series of actuators or settings. Distance travelled and vehicle speed were recorded along the way. A 10 m distance span was used initially and this was maintained by on-board timers. The Neuro-Fuzzy technique was used to derive a relief
map of the test track, and here the position data was translated and represented by two input and one output membership functions together with twelve rules as part of the optimisation routine. The relief map that was devised under the scheme was used for future gradient prediction. The chosen intelligent technique involved extraction of necessary representative features from a series of data points. Previous work using a similar Neuro-Fuzzy derived technique for modelling fuel spray penetration was described in [10] and achieved good results.

4.2. Experimental setup

The experiments were performed on a small passenger vehicle. A test route was established on the outskirts of Eastbourne in East Sussex, UK. A stand-alone laptop with a handheld GPS device was used throughout the experiment. The devised system was not connected to the vehicle control system. Therefore, an external GPS receiver was used and data were logged together with the time from the on-board clock through a serial Bluetooth interface. The aim of this investigation was to
use a GPS receiver in conjunction with custom-written Matlab software to collect and store three-dimensional vehicle position data. The incoming stream of data was used to estimate the future elevation of the vehicle; this data was expected to be of use further for dynamically influencing the control of an engine. The programme flowchart in Fig. 6 showed how these algorithms tied together to form a fuzzy predictive control system.

The main software was divided into several functional modules, each of which performed its own set of calculations. The optimisation was performed by the Neuro-Fuzzy module, the FIS generated was stored in the computer memory, whilst the timing of all these activities was governed by the on-board clock and timers.

4.3. Road gradient estimation

The fuzzy predictive control scheme is shown in Fig. 7. The operation was triggered by a start signal and the status of the GPS data A few given set points were needed i.e. predictive distance and sampling rate. It began with the first position data of the vehicle and this was registered and as a result a reference trajectory could be designed. From the reference trajectory, the next reference position was obtained according to the preset distance span. Meanwhile, the predictive algorithm calculated the next position of the vehicle using the current speed gathered from the GPS receiver. Based on the difference between the current and the predicted position, the fuzzy controller deduced the height at a set distance ahead and subsequently calculated the gradient.

The calculation of distance between two points on the Earth required the use of the Great Circle Distance Formula [11].

\[
\text{Distance}(m) = r \times \arccos \left( \frac{\sin(lat1) \times \sin(lat2) + \cos(lat1) \times \cos(lat2) \times \cos(long2 - long1)}{2} \right)
\]

where \( r \) was the radius of the earth, 6378.7 km. The variables \( lat1, long1 \) and \( lat2, long2 \) were the current position and predicted position, respectively.

The software was capable of handling double-precision floating point as this formula required a high level of floating point mathematical accuracy. The future location deduced using the described algorithms was of particular interest since it provided information about the condition of the road ahead, in order to realise the appropriate control signal.

4.4. Validation

A set of measurements is shown in Fig. 8. The solid line represents the height data; this was used to train the Neuro-Fuzzy network to produce a base relief map of the route, a second run was performed at a variable sampling rate i.e. speed dependent sampling. These data were used as test data, the predictive algorithm was applied where future gradient estimation was computed. The result was compared and automatically logged for off-line analysis.
Fig. 8. Measurement of road elevations.

Fig. 9. An overview of the simulation model.

5. Simulation and results

Simulation was carried out using Simulink, the status window was presented by an interface shown in Fig. 9. The simulation model can be divided into four sub-systems, and each sub-system contained its own algorithms and mathematical functions. The vehicle GPS block essentially was a collection of vehicle running data, this was recorded and consolidated to be used and linked to the rest of the sub-systems. The predictive
system shown in Fig. 10 was the core function of the fuzzy predictive model derived in the previous section. This sub-system primarily used a Neuro-Fuzzy technique to learn and model the route and this was continuously updated and replaced by new data as the vehicle moved along. A fuzzy logic derived relief map was generated and this was stored and used as a reference for the gradient prediction.

The gradient block shown in Fig. 11 was a series of mathematical functions. These were employed to convert predictive height into road gradient, taking into account the instantaneous vehicle speed and altitude. The engine/vehicle block consisted of real engine running data and the dynamic vehicle model previously explained in Section 3.1. Each sub-block was associated with variables such as vehicle mass, frontal area, drag coefficient where each individual value could be changed if required. Figure 12 shows the connections...
Fig. 12. Engine/vehicle model.

Fig. 13. Simulation results.
and the mathematical relationship between parameters. This simulation provided an effective platform by adding the complexity of the system under control to the test platform.

At each simulation time step, the controller action was to retrieve data from the GPS database, predict the height, compute the gradient with reference to the relief map and apply it through to the engine and vehicle dynamic model. The adjustable two-second time-step was selected to match the sampling speed of the GPS receiver and was utilised by the rest of the blocks. A series of plots and graphs was generated alongside, to monitor and validate the results. Figure 13 shows the results from a simulation run.

Each of these graphs indicates the corresponding results from each block and/or combination of blocks in the simulation. There were 754 GPS data in the set, covering 7.8 miles of test track. The whole journey took just under 25 minutes. The elevation plot shows the height profile of the test route. The predictive gradient plot was based on the height profile of a training run. The controller obtained this predictive information and adjusted the ignition timing according to the loads. Notice that the change of engine torque is affected by the level of loads acting on the vehicle, i.e. mass, gradient and speed. Consequently, as speed increases, the drag resistance increases exponentially. These graphs demonstrate the effectiveness of the system and how it responded to different loading conditions and road gradient, derived from the fuzzy logic relief map. The ignition timing plot illustrates the optimal value at each time step, when the AFR was maintained in order to achieve minimum fuel consumption and emissions.

The chosen test route included a few successive up and down hill sections, a magnified graph in Fig. 14 depicting one of them. In general, the predictive algorithm was able to distinguish the trend of a section except a glitch between 1150 to 1170 seconds. This is mainly due to the altitude resolution of the GPS receiver. Considering that the relief map information is continuously updated and replaced by new data as the vehicle moves along, thus this error might be dimin-
ished. However, the vehicle dynamic and the engine model worked well in response to the changes in vehicle speed and predictive gradient. The downhill section from 1172 seconds demonstrated the effect of decreasing loads due to both decreasing speed and gradient. The reduced loads on the vehicle reflected a reduction in engine torque which was shown in the ignition timing plot.

6. Conclusions

This paper has demonstrated that intelligent systems, and in particular neuro-fuzzy techniques, can be used for predictive vehicle control. The technique represented a convenient and robust method of achieving road prediction, to form a fuzzy system that ‘looks ahead’ leading to improved fuel consumption and a consequent reduction in exhaust emissions. A new algorithm was demonstrated, which integrates live GPS data with the existing fuzzy logic derived relief map; matching software was developed and successfully implemented. This Neuro-Fuzzy paradigm utilised simple map matching criteria, determining the gradient ahead based on current GPS position, and subsequently influenced the control of an engine. The GPS data observations were combined with fuzzy logic derived position to provide vehicle height information every two seconds.

Experimental results demonstrated the feasibility and advantages of this predictive fuzzy control on the trajectory tracking of a vehicle. Over 900 vehicle positions were generated and computed on each 7.8 mile test run using the newly devised algorithms. A similar number of test data were collected and compared to the height information generated by the predictive algorithm. The results showed that a good agreement was achieved between the predictive and the actual position data. The correlation coefficient of the elevation estimated by the Neuro-fuzzy technique is 0.996, indicating good correlation.

The technique developed in road height estimation performs well and has been simulated using Simulink. This was combined with the technique of HIL in the development and test of a complex real-time system. The results showed how each system responded and a predictive height algorithm was used throughout the simulation. The work also demonstrated the potential effectiveness of the system for use in developing a simplistic vehicle control system for reduced fuel consumption and emissions. Due to the fact that the method is tested and used on known and repeated routes, the system is intended and ideal on buses or fleet vehicles.

6.1. Further work

The system can be further improved with more low-cost GPS receiver technology and integration with in-vehicle sensors as well as with the engine operating parameters. An inclinometer or accelerometer can be found in many current vehicles; this essential fitment not only provides active safety in vehicles, i.e. the stability control, the braking system and automatic gearbox, but could also improve the accuracy of the height predictive algorithm.

Acknowledgements

This work has been co-funded by the European Union under the Interreg IIIa Programme of the European Regional Development Fund, Intelligent Vehicle On-board Systems ‘VBIS’. The authors gratefully acknowledge the contribution of Tiago Peres Lourenco Cardosa and Maxime Van Halst in providing data for this paper.

References


A hybrid quantum evolutionary algorithm for solving engineering optimization problems

Ashish Mani and C. Patvardhan*
Department of Electrical Engineering, Faculty of Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra, India

Abstract. Evolutionary Algorithms (EA) have been successfully employed for solving difficult constrained engineering optimization problems. However, EA implementations often suffer from premature convergence due to the lack of proper balance between exploration and exploitation in the search process. This paper proposes a Hybrid Quantum inspired EA, which balances the exploration and exploitation in the search process by adaptively evolving the populations. It employs an adaptive quantum rotation based crossover operator designed by hybridizing a conventional crossover operator with the principles of Quantum Mechanics. The degree of rotation in this operator is determined adaptively. The proposed algorithm does not require either a mutation operator, to avoid premature convergence, or a local heuristic to improve convergence rate. Further, a parameter-tuning free hybrid technique is employed for handling constraints, which overcomes some limitations in the traditional techniques like penalty factor methods, by hybridizing Feasibility Rules method with Adaptive Penalty Factor method. It is implemented by using two populations, each evolving by applying one of the constraints handling techniques and swapping a part of the populations. A standard set of six diverse benchmark engineering design optimization problems have been used for testing the proposed algorithm. The algorithm exhibits superior performance than the existing state-of-the-art approaches.

Keywords: Quantum evolutionary algorithm, constrained optimization, hybrid constraint handling, Adaptive

1. Introduction

Constrained Optimization is performed for minimizing the cost incurred in the real world engineering applications while meeting the design goals with maximum quality. Such Constrained Optimization Problems (COP) in continuous variables are generally formulated as follows:

\[
\text{Optimize } f(x), \text{ where } x = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^N, \text{ such that:}
\]

(A) \( x_{il} < x_i < x_{iu} \); \( x_i \) is the \( i \)th variable with \( x_{il} \) and \( x_{iu} \) as its lower and upper limits, respectively.

\( g_j(x) < 0; \ g_j \) is the \( j \)th inequality constraint and \( j = 1 \ldots p. \)

\( h_k(x) = 0; \ h_k \) is the \( k \)th equality constraint and \( k = 1 \ldots q. \)

*Corresponding author. Tel.: +91 9358380811; E-mail: cpatvardhan@gmail.com.

The objective function \( f(x) \) as well as the inequality and the equality constraints are often nonlinear, non-convex, non-differentiable, multimodal and of high dimension. Traditional deterministic optimization techniques like calculus based methods and enumerative strategies are often incapable of effectively solving such problems [15]. Calculus based methods typically perform local optimization by using search space domain information like local gradient [21]. Enumerative Strategies suffer from the curse of dimensionality [22]. Evolutionary Algorithms (EA) have been successfully applied to solve such complex constrained optimization problems, which are typically in the NP Hard domain.

EAs are population based stochastic search and optimization techniques inspired by the Nature’s Laws of evolution. They are popular due to their simplicity and ease of implementation. There are many variants of EAs, such as Evolutionary Strategies [18], Genetic Algorithms [7], Particle Swarm Optimization [13], etc.

EAs evolve a population of solutions by employing operators like selection, crossover, mutation and
local heuristic to find global optima. Selection operators are generally guided by the principle of survival of the fittest. Crossover operators generate solutions for the next iteration by recombining solutions from the current iteration. Mutation operators are employed as means for escaping from local minima, i.e., they improve exploration. Local heuristics are used for increasing the convergence rate i.e. they improve exploitation [5]. However, EAs still suffer from issues like premature convergence, slow convergence, stagnation and sensitivity to the choice of the crossover and mutation operators and parameters.

Many efforts have been made to overcome such limitations, by establishing a better balance between exploration and exploitation in the search process of the EA. Quantum inspired Evolutionary Algorithms (QEA) [9] have been also proposed to improve the balance between exploration and exploitation. QEAs are evolutionary algorithms inspired by the principles of Quantum Mechanics. They are developed by drawing some ideas from quantum mechanics and hybridizing them in the current framework of EAs. Thus, QEAs are not pure quantum algorithms, which can be implemented only on the Quantum Computers, but a new type of EAs. They are believed to require smaller population size in searching for global optima in less time as compared to the traditional EAs [9].

The important principles of Quantum Mechanics are Superposition, Entanglement, Interference and Measurement [17]. These principles have been mostly used in inspiring the representation of solution vector, crossover and mutation operators [19]. The principles that have been mostly utilized in designing QEAs are superposition and measurement [9]. These have been primarily used for probabilistic representation of search strings for improving the diversity. Another interesting observation is the use of single qubit (quantum analog of classical bit, and is governed by the principles of quantum mechanics) representation in almost all the efforts, thus ruling out the use of any other principles of quantum mechanics such as entanglement. Thus, there is a need to explore the possibility of designing QEAs that go beyond using Quantum principles, merely for representation of solutions and attempt to improve search by judiciously using these principles throughout the algorithm.

This paper proposes two qubits representation to utilize the principles of entanglement as well as superposition from quantum computing to achieve a better balance between exploration and exploitation for improving the search process. A parameter free adaptive quantum rotation based crossover operator has been designed by leveraging on the entanglement principle and the phase rotation gate for generating new population. Thus, the proposed QEA does not require mutation operator for avoiding premature convergence and local heuristic for improving convergence rate. It is also free from setting of any parameters which are dependent on the user’s judgment as parameters are derived from the fitness landscape of a problem. These features make the proposed algorithm unique, simple in concept and implementation, and at the same time very effective. It also propels the research in QEA in a new direction. EAs essentially perform unconstrained search and so handling of constraints is an important design decision, as no inherent mechanism is available in EAs to incorporate the constraints seamlessly without consuming too much extra computational effort. Many approaches have been suggested in literature [2] for handling constraints like repair algorithms, penalty factor based methods, Feasibility Rules, multi-objective optimization and Co-Evolution methods, etc.

The penalty factor method and its derivatives like dynamic and adaptive penalty factor methods and Feasibility Rules and its derivatives like Stochastic Ranking [20] are most commonly used for handling constraints. With the exception of Death penalty factor method and Feasibility Rules, they require parameter tuning, which itself is an optimization problem solved by the developers, mostly through trial and error and is often problem specific. In case of generic static penalty factor method, if the penalty factor is small, the resulting solution may be infeasible and if it is too large, the solution found is usually too far from the optimum. Death penalty factor method and Feasibility Rules do not require any parameter tuning and are straightforward in implementation. Further, their primary feature is preference of the feasible solutions over infeasible ones. However, they often fail in complex cases to find near optimum solutions [2].

Researchers have tried to solve the above-mentioned problems by introducing modifications in the generic penalty factor method and Feasibility Rules. The dynamic and adaptive penalty factor methods are examples of the modifications in the penalty factor method to solve some of the problems associated with the static penalty. However, they often require some parameters to be fine-tuned. Stochastic Ranking and other ranking based methods are modifications of Feasibility Rules, which permit more flexibility in searching high quality solutions by allowing some infeasible solutions to be treated as better than the feasible solutions, if
their objective function values are better. They are relatively complicated in implementation and involve additional computation of ranking as well as parameter tuning [20].

Therefore, in order to effectively and efficiently handle constraints, the two population Hybrid Constraint Handling Technique (HCT) is employed as described in [14], which is free from fine-tuning of any penalty parameters. Thus, the proposed real coded QEA has two populations.

The rest of the paper is organized as follows. Section 2 discusses related work available in the literature. Basics of Quantum computing are given in Section 3. The design of the proposed algorithm is presented in Section 4. The testing and results are discussed in Section 5. Section 6 concludes with a brief summary and throws some light on the direction of future work.

2. Related work

Feynman [6] had originally proposed that Quantum Mechanical Systems could be used for computing purposes. Moreover, they can provide a better alternative than classical Von Neumann computers for simulating quantum mechanical phenomena [6]. Quantum computing gained popularity with the development of polynomial time Shor’s factoring algorithm and Grover’s algorithm for quick search in unsorted databases [17]. However, these Quantum Algorithms can be implemented efficiently on the Quantum Computers and not on the classical computers. The development in the hardware of quantum computer is still in its nascent stages with a number of technological challenges to be overcome before a quantum computer of significance can become commercially available. The quantum computing paradigm has been shown to be more powerful than classical computing. It can be also used for improving the classical nature inspired stochastic algorithms. Quantum inspired evolutionary algorithms (QEA) are such an effort at integrating the principles of Quantum Computing and Evolutionary Algorithms.

Narayanan and Moore [16] proposed quantum-inspired genetic algorithm by integrating Quantum Parallel versus Interference with genetic crossover operator to solve Travelling Salesman problem. Han and Kim introduced probabilistic qubit representation of the solution vector and also used inspirations from the quantum measurement principle and phase rotation quantum gates [10]. They further improved the phase rotation quantum gate and introduced a new termination criterion [9]. It was claimed that QEA maintained better balance between exploration and exploitation due to the probabilistic nature of the qubit. However, these attempts used binary string representation and designer specified parameters whose choice is governed by the nature of the problem.

An interesting point of such implementations is no direct correspondence between the solution vector and the qubits especially in case of the real coded QEA [1, 9]. The quantum rotation gates/operators also behave independent of the information from the problem and solution domains assuming that the quantum behavior would help in reaching near the optimal solution. However, it should not be forgotten that such algorithms are to be run on a classical computer without simulating any quantum phenomena. Further, it can be argued that increasing the diversity by collapsing the solution qubit may affect the exploitation of the solution as the solution found in the next iteration even of a good candidate solution may end up being far worse due to the probabilistic implementation.

This paper takes a different approach for designing the QEA by using not only qubit representation and the associated superposition principle but also the entanglement principle. Entanglement is one of the fundamental principles and if any two qubits are entangled then performing any quantum operation on one of the qubits would affect the state of the other qubit i.e. there is a relation between the two qubits, which can be utilized for computation purposes. The proposed QEA uses two qubits per solution vector to utilize the entanglement principle. The entanglement principle is integrated in a classical algorithm, so its implementation is classical. The proposed QEA tries to overcome some of the limitations associated typically with the EA implementations. EAs have only objective function value as the domain information regarding a specific problem. This feedback is mostly used in the selection phase and not for directly controlling the crossover or mutation operators or even the local heuristics. Therefore, the feedback through the objective function value is not being utilized properly. The proposed algorithm uses the second qubit to store the information regarding the objective function value of the solution vector. This provides information regarding the solution domain as well as the problem function domain made available simultaneously. The information stored in the first and the second qubit are entangled to harness the power of the important entanglement principle. The first qubit influences the second qubit as the amplitude of the first qubit determines the objective function value and hence
the amplitude of the second qubit. The second qubit influences the first qubit as the parameter free adaptive quantum rotation crossover operator used for evolving the first qubit uses the amplitude of the second qubit to determine the degree of rotation. Any operation performed on either of the two qubits affects the other and so they are entangled in a classical implementation.

The real coded QEA (RQEA) proposed in [1] has used static and dynamic penalty factor methods for handling constraints for different problems, thus exemplifying that static and dynamic penalty factors are problem specific and require fine-tuning, which in itself is an optimization problem. Dominance Based Tournament selection has been used in [3] for constraint handling, which is the application of multi-objective strategy coupled with the probabilistic version of Feasibility Rules. However, it introduces a new parameter Sr. ECPSO employs an adaptive co-evolutionary penalty for handling constraints [11] in which two swarms are employed to handle the constraints, however, three more strategies had to be further incorporated for improving the performance.

3. Quantum inspired elements

3.1. Representation

The smallest information element in a quantum computer is a qubit, which is the quantum analog of a classical bit. The classical bit can be either in state ‘zero’ or in state ‘one’ whereas a quantum bit can be in a superposition of the basis states in a quantum system. A qubit is represented by a vector in the Hilbert space with \(|0\rangle\) and \(|1\rangle\) as the basis states. The vector \(|\psi\rangle\) represents a qubit and defined as:

\[ |\psi\rangle = \alpha|0\rangle + \beta|1\rangle \quad (1) \]

where \(|\alpha|^2\) and \(|\beta|^2\) are probability amplitudes of the qubit to be in state \(|0\rangle\) and \(|1\rangle\) respectively and should satisfy the condition:

\[ |\alpha|^2 + |\beta|^2 = 1. \quad (2) \]

The qubits can store, in principal exponentially more information than the classical bits. However, these qubits exist in a quantum computing system and are constrained by several limitations, like they collapse to one of the basis states upon measurement and evolve by using the unitary transformations only. The simulation of qubits is inefficient on a classical computer.

In QEA, the probabilistic nature of qubits has been widely used for maintaining the diversity [10]. A single qubit is attached to each solution vector and the solution is obtained by taking measurement or collapsing in binary coded as well as real coded QEA. However, due to the reasons explained in the previous section, this paper uses real coded representation of the solution with two sets of qubits. The first set of Qubits \(|\psi_1\rangle\) stores the current value of the ith variable as amplitude \(\alpha_{1i}\), whose value is \([0, 1]\). The upper and lower limits of the variables are scaled between 0 and 1. The amplitude \(\beta\) is not stored as it can be computed from the equation 2. Therefore, the number of qubits per quantum register \(QR_1\) is equal to the number of variables. A quantum register is used for storing qubits. The number of quantum registers has also been made a function of the number of variables in a specific problem. Thus, even the selection of the number of quantum registers, which constitute the population of solutions, has a problem bias. The number of \(QR_1\) is 100 times the number of variables. The number of \(QR_1\) has been maintained large in order to utilize the benefit of superposition and quantum rotation inspiration in the EA implementation.

The structure of \(QR_1\) is shown below:

\[ QR_{11} = [\alpha_{111}, \alpha_{112} \ldots \alpha_{11n}] \]

\[ \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \]

\[ QR_{1100N} = [\alpha_{1100n1}, \alpha_{1100n2} \ldots \alpha_{1100nn}] \]

The second set of Qubits in quantum register \(QR_2\) store ranked and scaled objective function values of the corresponding solution vectors in the \(QR_1\). The ith qubit in the \(QR_2\), \(|\psi_{2i}\rangle\) stores the ranked and scaled objective function value of the ith solution vector \((QR_{1i})\) as amplitude \(\alpha_{2i}\), whose value \([0, 1]\). The fittest vector’s objective function value is assigned 1 and the worst vector is assigned value 0. The rest of the solution vectors’ objective function value is ranked and assigned a value between 0 and 1.

3.2. Entanglement and superposition

Quantum Entanglement is a property of the quantum mechanical system with two or more qubits in which the quantum state of one qubit influences the other even if they are spatially separated [17]. The mathematical representation of the classical implementation of the superposition and entanglement principles is as given below:

\[ |\psi_{2i}(t)\rangle = f_1(|\psi_{1i}(t)\rangle) \quad (3) \]
\[ |\psi_{1i}(t + 1) > = f_2(|\psi_{2i}(t) >, |\psi_{2j}(t) >, |\psi_{1i}(t) >, |\psi_{1j}(t) >) \]

where \( |\psi(t)_{1i} > \) and \( |\psi(t)_{1j} > \) are the first set of qubits associated with \( i \)th and \( j \)th solution vectors respectively, \( |\psi_{2i} > \) and \( |\psi_{2j}(t) > \) are the second set of qubit associated with the \( i \)th and \( j \)th solution vectors, \( t \) is the current iteration number, \( f_1 \) and \( f_2 \) are the functions through which both the qubits are classically entangled. The superposition principle is harnessed classically by ranking and scaling the second qubit so the second qubit is in superposition of the solution vector’s i.e. first qubit. Similarly, the first qubit is also scaled uniformly between zero and one. It is further enhanced by increasing the number of quantum registers.

3.3. Adaptive Quantum rotation based Crossover Operator (AQCO)

Quantum gates are used for evolving qubits in a quantum computing system. Quantum phase rotation gate is used in Grover’s algorithm for amplitude amplification in searching the marked element stored in an unsorted database. Most of the efforts have used rotation gates for evolving the qubits. The operator’s parameters are mostly problem specific and assigned by the designer. Further, the past efforts have also used the mutation operators and the local heuristics [19].

A quantum rotation based adaptive and parameter tuning free crossover operator is designed by hybridizing quantum entanglement with modified BLX-\( \alpha \) crossover operator [12]. The second qubits’ amplitude is used for determining the degree of rotation for evolving the first qubit. The following equation is used for evolving \( |\psi_{1i}(t) > \):

\[ |\psi_{1i}(t + 1) > = |\psi_{1i}(t) > + f_2(|\psi_{2i}(t) >, |\psi_{2j}(t) >) * (|\psi_{2j}(t) > - |\psi_{1i}(t) >) \]

where \( t \) is the iteration number, \( |\psi(t)_{1j} > \) can be any other solution vector in the population.

The two solution vectors, \( |\psi(t)_{1i} > \) and \( |\psi(t)_{1j} > \), involved in the adaptive crossover operator can be selected either deterministically or randomly. If \( |\psi_{1i}(t) > \) has inferior fitness than \( |\psi_{1j}(t) > \) then \( |\psi_{1i}(t) > \) is rotated towards \( |\psi_{1j}(t) > \). If \( |\psi_{1i}(t) > \) is the best solution vector, it is rotated away from \( |\psi_{1j}(t) > \). Therefore, this operator balances the exploration and exploitation and converges the population adaptively towards global optima by using three rotation strategies:

- Rotation towards the Best Strategy (R-I) is implemented by selecting each solution vector (individual) sequentially except the best solution vector. The best individual is used as the pivot towards which all the other individuals are rotated. The main motivation behind this strategy is the conjecture that there is a high probability of improving all the other individuals by searching towards the best individual. Thus, in the initial stages of the search, this strategy helps in exploration and during the later stage, when the individuals are mostly located in a smaller region, helps in exploitation.
- Rotation away from the Worse Strategy (R-II) is implemented by selecting each individual sequentially except the best individual. The best individual is rotated away from all other individuals. This strategy is primarily used for exploitation purpose, as there is a high probability of improving the best individual by searching in its vicinity. The search takes place in all dimensions around the best individual due to its rotation away from the other individuals in the population.
- Rotation Towards the Better Strategy (R-III) is implemented by selecting any two individuals randomly and rotating the inferior one towards the better individual. This strategy is primarily used for the exploration purpose, as there is a uniform guided search in the entire solution space for reaching the regions closer to the optimum. It can be appreciated that, it is not a blind search as is the case in the mutation operators of the canonical EAs.

The function \( f(|\psi_{2i}(t) >, |\psi_{2j}(t) >) \) controls gross and fine searches. Presently, \( f(|\psi_{2i}(t) >, |\psi_{2j}(t) >) \) generates a random number either between \( \alpha_{2i} \) and \( \alpha_{2j} \) or \( |\alpha_{2j}|^2 - |\alpha_{2i}|^2 \). The value \( |\alpha_{2j}|^2 - |\alpha_{2i}|^2 \) is generally smaller than \( \alpha_{2j} - \alpha_{2i} \), thus the later is used for the gross search and the former for the fine search. The salient feature of the new quantum rotation based crossover operator is that it adaptively changes each variable in the solution vector and at the same time is problem driven rather than being an arbitrary choice of the user. Fixed rotation was also attempted by using \( |\alpha_{2j}|^2 - |\alpha_{2i}|^2 \) and \( \alpha_{2j} - \alpha_{2i} \), but failed as it reduces the diversity and spoils the balance between exploration and exploitation in the search process.

4. Proposed quantum evolutionary algorithm

The proposed algorithm has two populations A and B as it is employing HCT [14] for handling constraints.
Feasibility Rules are used for handling constraints in population A whereas Adaptive Penalty Factor method is employed for handling constraints in population B. In this implementation, population A has been given primary preference and is evolved in every iteration, whereas population B is evolved only when population A fails to improve its best solution in any iteration. This introduces diversity in the search process of population A only when required and also reduces the overall computational burden. The pseudo code for the proposed algorithm is given in Fig. 1, which is explained as follows:

- a) Initialization of the first set of qubits in quantum registers $QR_{1A}$ and $QR_{1B}$ is performed randomly. $QR_{1A}$ and $QR_{1B}$ store alphas ($\alpha$) corresponding to the solution vectors scaled between [0, 1] for both the populations.

- b) Computation of the fitness of each solution vector of $QR_{1A}$ is performed by using Feasibility Rules, which is as follows [4]:
  
  i. If the two solution vectors being compared are both feasible, the one with better objective function value $f(x)$ is considered fitter.
  
  ii. If one solution vector is feasible and the other is infeasible, the feasible one is considered fitter.
  
  iii. If both the solution vectors are infeasible, the one with the lower level of constraint violation or degree of infeasibility is considered fitter.

- c) $QR_{2A}$ stores the ranked and scaled objective function values of the solution vectors in $QR_{1A}$. The ranked and scaled objective function value of the $i$th solution vector ($QR_{1A_i}$) is stored in $QR_{2A_i}$ as amplitude $\alpha_{2i}$ whose value [0, 1]. The fittest vector’s objective function value is assigned 1 and the worst vector is assigned value 0. The rest of the solution vectors’ objective function value is ranked and assigned a value between 0 and 1.

- d) Adaptive Quantum Rotation based Crossover is performed by using all the three strategies R-I, R-II and R-III. R-I is used with 50% percent probability, R-II with 25% probability and R-III with 75% probability to strike a proper balance between exploration and exploitation during the search, while increasing the efficiency by 50% when all the three rotation strategies are used with 100% probability.

- e) The population for the next generation are selected by comparing individual parent’s fitness with their best child’s fitness and applying tournament selection i.e. the fitter one would make it to the next generation.

- f) If there is no improvement in the best solution vector of Population A in any generation, then Adaptive Penalty Factor is computed [14]. It is...
designed to make the objective function value of the fittest solution vector in population A and the modified objective function value (equation 7 given in (g)) of the Individual with the best objective function value in population B equal in the current generation. This automatically chooses a new value of the adaptive penalty factor for the generation to guide the search in entire solution domain. The equation for computing adaptive penalty factor ‘S’ is as follows:

\[
I f(CV_{Bi} > 0) \{ \\
S = \left( f_{BAi}(x) - f_{BBi}(x) \right) / CV_{Bi} \text{forMin} \\
S = \left( f_{BBi}(x) - f_{BAi}(x) \right) / CV_{Bi} \text{forMax} \} \\
I f(S < 0) \quad S = 0
\]

where \( f_{BAi}(x) \) = Objective function value of the Best Individual of Population A in the ith generation.

\( f_{BBi}(x) \) = Objective function value of the Individual with best objective function value of Population B in the ith generation.

\( CV_{Bi} \) = Constraint Violation of the Individual with best objective function value of Population B in the ith generation.

Adaptive penalty factor methods have been known to work better provided the penalty factor is adaptively made smaller if the feasible solutions are being found and made larger otherwise [2]. This is done so that there is an indirect pressure for finding feasible solutions. However, these methods do not provide any assistance in actually finding the feasible solutions. In the two population method, population A has a greater propensity towards feasible solutions because it utilizes the Feasibility Rules and population B ensures that the domain on search is not restricted.

\[ g \) Population B’s fitness is evaluated by using modified objective function value, which is determined by using adaptive penalty factor.

Objective function = \( f(x) \)

Constraint Violation = Degree of infeasibility
\[ = \{ \Sigma(g_i(x)) + \Sigma(|h_i(x)|) \}. \]

Modified Objective function \( \Phi = f(x) \)
\[ + S \times \{ \Sigma(g_i(x)) + \Sigma(h_i(x)) \}, \] (7)

where \( g_i + (x) = \max \{ 0, g_i(x) \} \) and \( h_i \) = \( 0, if \left | h_i(x) \right | - \delta < 0 \) else \( |h_i(x)|, \delta = 10^{-10}. \)

h) \( QR_{2B} \) stores ranked and scaled objective function values of the solution vectors in \( QR_{1B}. \) The ranked and scaled objective function value of the \( i^\text{th} \) solution vector \( QR_{1Bi} \) is stored in \( QR_{2Bi} \) as amplitude \( \alpha_{2i} \), whose value \([0, 1] \). The fittest vector’s objective function value is assigned 1 and the worst vector is assigned value 0. The rest of the solution vectors’ objective function value is ranked and assigned a value in between 0 and 1.

i) Adaptive quantum rotation based crossover is performed only 50% time using R-III as this population is primarily used for exploration.

j) Same as in e).

k) Swap [14] exchanges a part of population A and B using Greedy and Random strategies so that the search is guided towards global optima while maintaining the diversity in both the populations. Greedy strategy is implemented through the replacement of the least-fit individual of both the populations A and B by the fittest individual in the other population. Random strategy is implemented by selecting and exchanging part of the population randomly, which precludes individuals used in the Greedy strategy. The number of individuals being exchanged in the Random Strategy is a strategy parameter, named \( swap_{per} \) and is 10% of the population size in this implementation for all the problems.

5. Testing and results

The proposed algorithm has been tested on six benchmark engineering constrained optimization problems, which have been widely used for testing similar optimization algorithms.

Thirty independent runs are performed for each problem using the proposed algorithm, which is implemented in ‘C’ programming language on an IBM Workstation with Pentium-IV 2.4 GHz processor, 2 GB RAM and Windows XP platform. The testing has been performed for determining the stability of the proposed algorithm. The stability and robustness have been determined by analyzing statistically the quality of the solutions produced for each problem in thirty independent runs. The efficiency has been determined by the number of generations and the maximum number of generations is limited to 500 for all the problems.

The problems are arranged in the order of their increasing difficulty for the proposed algorithm. The experiments have been designed to gain better insight in-
the hybrid constraint handling technique. The fourth configuration (SC-IV) uses standard configuration but without Swap to validate the selection of parameters. This indicates the importance of random rotation and exploration by R-III in the search process for this problem. SC-IV has performed worst of all the four configurations, thus highlighting the importance of HCT for this problem, possibly due to the presence of an integer variable.

The third problem (P-3) is minimization of the volume of material required for construction of a three bar truss while satisfying stress limitations under a load condition [8]. The stress in each member is limited to 2KN/cm$^2$ and the load is limited to 2KN/cm$^2$. The results for all the four configurations for P-3 are presented in Table 3. The overall performance of SC-I is better than all the other configurations. SC-II and SC-III have reached optimal in their best run but are less robust than SC-I as indicated by other statistical parameters. This indicates the importance of random rotation and exploration by R-III for robust performance in this problem. SC-II has performed better than SC-III, thus again, highlighting the importance of exploration by R-III in the search process for this problem. SC-IV has performed worst of all the four configurations, thus highlighting the importance of HCT for this problem, possibly due to the presence of an integer variable.

The fourth problem (P-4) is minimizing the weight of the speed reducer [11] with face width, module of teeth, number of teeth on pinion, length of the first shaft between bearings, length of the second shaft between bearings, diameter of the first shaft, and diameter of the second shaft as variables (all of them are continuous except number of teeth on pinion, which is integer). The weight of the speed reducer is to be minimized subject to the constraints on bending stress of gear teeth, surface stress, transverse deflections of the shafts and stresses in the shaft. The results for all the four configurations for P-4 are presented in Table 4. The overall performance of SC-I is better than all the other configurations. SC-II and SC-III have reached optimal in their best run but are less robust than SC-I as indicated by other statistical parameters. This indicates the importance of random rotation and exploration by R-III for robust performance in this problem. SC-II has performed better than SC-III, thus again, highlighting the importance of examination by R-III in the search process for this problem. SC-IV has performed worst of all the four configurations, thus highlighting the importance of HCT for this problem, possibly due to the presence of an integer variable.

The fourth problem (P-4) is minimizing the weight of a tension/compression spring, subject to the constraints to working of the proposed algorithm by analyzing its Adaptive Quantum rotation based Crossover Operator (AQCO) in four different configurations. The first experiment is performed with the standard configuration (SC-I) as described in Section 4 i.e. all three types of the rotation strategies have been used. The second configuration (SC-II) uses deterministic rotation instead of random rotation in the standard configuration to determine whether random rotation is a better design option than deterministic rotation. The third configuration (SC-III) uses R-I and R-II of the standard configuration but not R-III, i.e. random exploration is turned off to find the effectiveness of the exploration mechanism of AQCO. The fourth configuration (SC-IV) uses standard configuration but without Swap to validate the selection of the hybrid constraint handling technique.

The first problem (P-1) is originally proposed by Himmelblau [11] with five design variables, six nonlinear inequality constraints and 10 boundary conditions. The results for all the four configurations for P-1 are presented in Table 1. The performance of SC-I, SC-II and SC-IV is optimal in all the runs. This indicates that for certain class of problems even the deterministic rotation and use of the Feasibility rules can be effective. SC-III reached optimal in its best run but is less robust than the other three configurations as indicated by the statistical parameters. This indicates that limiting exploration by turning off R-III adversely affects searching ability of the algorithm for this problem.

The second problem (P-2) is designing of the speed reducer [11] with face width, module of teeth, number of teeth on pinion, length of the first shaft between bearings, length of the second shaft between bearings, diameter of the first shaft, and diameter of the second shaft as variables (all of them are continuous except number of teeth on pinion, which is integer). The weight of the speed reducer is to be minimized subject to the constraints on bending stress of gear teeth, surface stress, transverse deflections of the shafts and stresses in the shaft. The results for all the four configurations for P-2 are presented in Table 2. The overall performance of SC-I is better than all the other configurations. SC-II and SC-III have reached optimal in their best run but are less robust than SC-I as indicated by other statistical parameters. This indicates the importance of random rotation and exploration by R-III for robust performance in this problem. SC-II has performed better than SC-III, thus again, highlighting the importance of exploration by R-III in the search process for this problem. SC-IV has performed worst of all the four configurations, thus highlighting the importance of HCT for this problem, possibly due to the presence of an integer variable.
of minimum deflection, shear stress, surge frequency, and limits on the outside diameter and on the design variables. There are three design variables: the wire diameter, the mean coil diameter, and the number of active coils [3]. The results for all the four configurations for P-4 are presented in Table 4. The overall performance of SC-I is better than all the configurations. SC-II, SC-III and SC-IV have reached near optimal in their best run but are less robust than SC-I as indicated by other statistical parameters. This indicates that random rotation, exploration by R-III and HCT are important for robust performance of the proposed algorithm in this problem. SC-II has performed better than SC-III and SC-IV. SC-IV has performed worst than all the configurations. The relative performance of SC-II, SC-III and SC-IV indicates the relative importance of random rotation, exploration by R-III and HCT for this problem i.e. HCT is the most important followed by exploration through R-III and random rotation.

The fifth problem (P-5) is designing a welded beam for minimum cost, subject to some constraints [3]. The objective is to find the minimum fabrication cost, considering four design variables and the constraints of shear stress, bending stress in the beam, buckling load on the bar, and end deflection on the beam. The results for all the four configurations for P-5 are presented in Table 5. The overall performance of SC-I is better than all the configurations. This indicates that random rotation, exploration by R-III and HCT are important for robust performance of the proposed algorithm in this problem. SC-IV performed better than SC-II and SC-III. SC-III performed worst than all the configurations. The relative performance of SC-II, SC-III and SC-IV indicates the relative importance of random rotation, exploration by R-III and HCT for this problem i.e. HCT is the most important followed by random rotation and R-III.

The sixth problem (P-6) is designing a compressed air storage tank with a working pressure of 3,000 psi and a minimum volume of 750 ft³. A cylindrical vessel is capped at both ends by hemispherical heads. Using rolled steel plate, the shell is made in two halves that are joined by two longitudinal welds to form a cylinder.
rithms like RQiEA [1], ECPSO [11] and DTS [3] on their best solution for problems P-4, P-5 and P-6. The results of column SC-I in Table 1 to Table 6 along with Table 7 show that the proposed algorithm named HAR-QiEA is far better than the earlier proposed RQiEA [1], which is a QEA utilizing probabilistic representation of qubit for maintaining diversity. It is also better than the traditional state of art EAs like ECPSO [11] and DTS [3].

The first three problems P-1, P-2 and P-3 have not been considered in the comparative study as these have been comprehensively solved and so offer no further insights. However, it is important to use them along with other problems in testing for two reasons. Primarily, it has facilitated in a better understanding of the impact of design alternatives and the role of various rotation strategies used in the proposed algorithm. Further, it has also showed that the proposed algorithm is not only good for the problems, which appear difficult to other algorithms, but is equally good for the problems, which appear easy to other algorithms. As otherwise, it can be argued that the performance improvement over one set of the problems is made possible by sacrificing the same over the other set.

6. Conclusions and future work

Constrained Optimization is an important problem in engineering domain for which a new hybrid adaptive quantum evolutionary algorithm is proposed. The algorithm hybridizes the quantum entanglement and superposition principles with a conventional crossover operator to adaptively evolve the population using three different rotation strategies, viz., Rotation towards the best, Rotation away from the worse and Rotation towards the better. The degree of rotation in each strategy is adaptively determined by the feedback from the fitness space. Thus, there is no need for fine tuning parameters in the quantum rotation based crossover operator. It has been implemented by using two qubits representation instead of one, which enables the utilization of the quantum entanglement and superposition principles hitherto not tapped.

The proposed algorithm uses the latest hybrid constraint handling technique based on the hybridization of Feasibility Rules and Adaptive Penalty Factor method. This technique leverages best characteristics from both the methods in a simple and effective way by using two populations, each evolving by employing one of the methods of handling constraints and swapping part of the population at the end of the generation. Further, it is free from fine-tuning of the penalty parameters, as the adaptive penalty factor is calculated from the fitness information available about the best individuals in both populations.

The proposed algorithm has been tested on a standard set of six benchmark engineering optimization problems, for validating design decisions and perform comparative study with the state-of-art algorithms. The results showed that the proposed algorithm performed better than the existing algorithms without even employing a mutation operator or a local heuristic for improving the quality of the solution.

Future work would involve more in-depth analysis to understand the working of the proposed algorithm. Moreover, an effort would be made to study its applicability in other areas requiring optimization, such as parameter estimation, dynamic economic dispatch problems, etc.

Acknowledgments

We are extremely grateful to the Most Revered Chairman, Advisory Committee on Education, Dayalbagh, for continued guidance and support in our every endeavor. We are also thankful to the Reviewers for giving positive feedback towards improving the presentation of the paper.

References


Power system security enhancement using fuzzy logic composite criteria and particle swarm optimization algorithm

K. Vaisakh\textsuperscript{a,∗}, Ch. Padmanaba Raju\textsuperscript{b} and S. Sivanaga Raju\textsuperscript{c}

\textsuperscript{a}Department of Electrical Engineering, AU College of Engineering, Andhra University, Visakhapatnam-530003, AP, India
\textsuperscript{b}Department of Electrical and Electronics Engineering, P.V.P. Siddhartha Institute of Technology, Kanuru, Vijayawada, AP-520007, India
\textsuperscript{c}Department of Electrical and Electronics Engineering, J.N.T.U. College of Engineering, Kakinada-533003, AP, India

Abstract. The network contingencies in power system often contribute to over-loading of network branches, unsatisfactory voltages and also to voltage collapse. To maintain security against voltage collapse, it is desirable to estimate the effect of contingencies. For security enhancement, remedial action against possible network overloads in the system following the occurrence of contingencies is required. Line overloads can be removed by means of generation re-dispatching and by adjustment of reactive power control variables such as tap-changing transformers, generator excitations, and shunt reactive power compensating devices. This paper presents fuzzy logic composite criteria to evaluate the degree of severity of the network contingency. A Particle Swarm Optimization (PSO) based algorithm was also proposed for solving the Optimal Power Flow (OPF) problem to alleviate line overloads. The proposed PSO algorithm identifies the optimal values of generator active-power output and the adjustment of reactive power control devices. Simulation results on IEEE 14, 30, and 118-bus test systems are presented and compared with the results of other approaches reported in the literature.

Keywords: Optimal power flow, security enhancement, composite criteria, fuzzy logic, contingency ranking

1. Introduction

With the continued increase in demand for electrical energy with little addition to transmission capacity, security assessment and control have become important issues in power-system operation. Security assessment deals with determining whether or not the system operating in a normal state can withstand contingencies (such as outage of transmission lines, generators, etc.) without any limit violation. Contingency screening and ranking is one of the important components of on-line system security assessment. The contingency ranking methods, generally, ranks the contingencies in an approximate order of severity with respect to a scalar performance index (PI), which quantifies the system stress [1]. The common disadvantages of several PI-based ranking methods are the masking phenomenon.

Moreover, with increased loading of existing power transmission systems, the problem of voltage stability and voltage collapse has also become a major concern in power system planning and operation. It has been observed that voltage magnitudes do not give a good indicator of proximity to a voltage stability limit and voltage collapse [2,3]. Therefore, it is necessary to consider voltage stability indices as pre/post-contingency quantities in the evaluation of severity of network contingency.
To measure the severity level of voltage stability problems, a lot of performance indices have been proposed [4]. They could be used on-line or off-line to help operators determine how close the system is to collapse. In general, these indices aim at defining a scalar magnitude that can be monitored as system parameters change, with fast computation speed. They include the sensitivity factors [5], second order performance index [6], voltage instability proximity index (VIPI) [7], singular values and eigen values [8], and so on.

For secure operation of the system without any limit violation, complete modeling of the system through load flow equations and operational constraints is necessary. The solution of formulated Optimal Power Flow (OPF) model gives the optimal operating state of a power system and the corresponding settings of control variables for economic and secure operation, while at the same time satisfying various equality and inequality constraints. The equality constraints are the power flow equations, while the inequality constraints are the limits on control variables and the operating limits of power system dependent variables. Amongst a number of different operational objectives that an OPF problem may be formulated, a widely considered objective is to minimize the fuel cost subject to equality and inequality constraints.

As to the possibility of including stability constraints into standard OPF formulations [9], discusses the issue and [10] develops a conceptual framework. Several hybrid OPF formulations incorporating voltage stability constraints are presented in [11,12]. The method puts requirements on evaluating the critical point of voltage stability, so the problem size and computation burden are enhanced. Ref [13] reports an optimal dispatch with voltage stability constraints, using the bifurcation technique to calculate the voltage stability margin. Ref [14] proposes a voltage stability constrained OPF with the modified form of L-index as voltage stability constraints.

In the recent past, a new evolutionary computation technique, called particle swarm optimization (PSO) has been proposed by Kennedy and Eberhart in 1995 [15]. PSO has been motivated by the behavior of bird flocks and fish schools. PSO is one of the latest versions of nature inspired algorithms with characteristics of high performance and easy implementation. This method has been found to be robust for solving problems featuring non-linearity, non-differentiability, multiple optima and high dimensionality. PSO has been successfully applied to various fields of power system optimization such as economic dispatch [16], reactive and voltage control [17], etc.

The paper presents a new approach to the assessment of power system security. Using fuzzy membership functions of post-contingent quantities, it quantifies the security state of a power system, which uses off-line screening for the most vulnerable system states. It introduces fuzzy logic composite criteria (FLCC) of the power system using system’s variables characterized by fuzzy sets of a trapezoidal form. The FLCC uses the voltage stability indices at the load buses and reactive power outputs of generators as post-contingent quantities in addition to real power loadings and bus voltage violations to evaluate the network contingency.

An optimal power flow formulation with voltage stability constraint for overload alleviation by the optimal settings of all controllable variables for a static powersystem-loading condition was also presented in this paper. The proposed approach is illustrated through a corrective action plan for a few harmful contingencies in the IEEE 14-bus, IEEE 30-bus and IEEE 118-bus systems. Simulation results demonstrate that the PSO algorithm along with the fuzzy logic composite criteria provides very remarkable results compared to those reported in the literature.

The remainder of the paper is organized as follows: Section 2 describes the voltage stability L-index computation, while Section 3 describes the fuzzy logic composite criteria. Section 4 explains the computation of overall severity index based on fuzzy logic composite criteria for a network contingency ranking. Section 5 describes the severity index based on line loading. Section 6 describes the formulation of an optimal power flow problem, while Section 7 explains the standard particle swarm optimization. Section 8 then details the procedure of proposed PSO based algorithm for solving optimal power problem for alleviating network overloads, and Section 9 presents the results of the proposed algorithm on IEEE 14, 30 and 118–bus systems. Lastly, Section 10 outlines the conclusions.

### 2. Voltage stability index (L-index) computation

The voltage stability L-index is a good voltage stability indicator with its value change between zero (no load) and one (voltage collapse) [18]. Moreover, it can be used as a quantitative measure to estimate the voltage stability margin against the operating point.

For a given system operating condition, using the load flow (state estimation) results, the voltage stability $L$-index is computed as [18].
\[ L_j = \left| 1 - \sum_{i=1}^{g} F_{ji} \frac{V_i}{V_j} \right| \quad j = g + 1, \ldots, n \quad (1) \]

All the terms within the sigma on the RHS of Eq. (1) are complex quantities. The values of \( F_{ji} \) are obtained from the network Y-bus matrix. For stability, the index \( L_j \) must not be violated (maximum limit = 1) for any of the nodes \( j \). Hence, the global indicator \( L_j \) describing the stability of the complete subsystem is given by maximum of \( L_j \) for all \( j \) (load buses). An \( L_j \)-index value away from 1 and close to 0 indicates an improved system security. The advantage of this \( L_j \)-index lies in the simplicity of the numerical calculation and expressiveness of the results.

3. Fuzzy logic composite criteria

Fuzzy set theory is a mathematical concept proposed by Prof. L. A. Zadeh in 1965. Fuzzy logic is a kind of logic using graded or quantified statements rather than ones that are strictly true or false. The fuzzy sets representing linguistic variables allow objects to have grades of membership from 0 to 1. The intervals for fuzzification of linguistic variables are determined based on the pre/post-contingent quantities. The input and output membership functions are selected based on the classification of pre/post-contingent quantities. The pre/post-contingent quantities are first expressed in fuzzy set notation before they can be processed by the fuzzy reasoning rules.

3.1. Line loadings

Each pre/post-contingent percentage line loading is divided into four categories using fuzzy set notations: Lightly Loaded (LL), 0–50%, Normally Loaded (NL), 50–85%, Fully loaded (FL), 85–100%, Over Loaded (OL), above 100%. Figure 1 shows the correspondence between line loading and the four linguistic variables.

The output membership functions to evaluate the severity of a pre/post -contingent quantity are also divided into four categories using fuzzy set notations: Less Severe (LS), Below Severe (BS), Above Severe (AS) and More Severe (MS), as shown in Fig. 2.

After obtaining the severity indices of all the lines the Overall Severity Index (\( OSI_{LL} \)) of the line loading for a particular line outage is obtained using the wing expression.

\[ OSI_{LL} = \sum w_{LL} SI_{LL} \quad (2) \]

where \( w_{LL} = \) Weighting coefficient for a severity index,
\( SI_{LL} = \) Severity Index of a pre/post -contingent quantity

The weighting coefficients used for the severity indices are \( w = 0.25 \) for LS, 0.50 for BS, 0.75 for AS and 1.00 for MS.

The effect of these weighting coefficients is that the overall severity index is first dominated by fourth category of severity index (MS) next by third, second and first category of severity index, respectively.

3.2. Bus voltage profiles

In this case each pre/post-contingent bus voltage profile is divided into three categories using fuzzy set notations: Low Voltage (LV), below 0.9pu, Normal Voltage (NV), 0.9–1.02 pu and Over Voltage (OV), above 1.02pu. Figure 3 shows the correspondence between bus voltages (in per unit) and the four linguistic variables.

The output membership functions used to evaluate the severity of a post -contingent quantity are also divided into three categories using fuzzy set notations: Below Severe (BS), Above Severe (AS) and MS (More Severe) are shown in Fig. 4.
After obtaining the severity indices of all the voltage profiles, the Overall Severity Index ($OSI_{VP}$) of the bus voltage profiles for a particular line outage is obtained using the expression

$$ OSI_{VP} = \sum w_{VP} SI_{VP} $$ \hfill (3)

The weighting coefficients used for the severity indices are $w_{VP} = 0.30$ for BS, 0.60 for AS and 1.00 for MS.

### 3.3. Voltage stability indices

Each pre/post-contingent voltage stability index is divided into five categories using fuzzy set notations: Very Low Index (VLI), 0–0.2, Low Index (LI), 0.2–0.4, Medium Index (MI), 0.4–0.6, High Index (HI), 0.6–0.8 and Very High Index (VHI), 0.8 above. Figure 5 shows the correspondence between voltage stability L-index and the four linguistic variables.

The output membership functions to evaluate the severity of a pre/post-contingent quantity are also divided into five categories using fuzzy set notations: Very Less Severe (VLS), Less Severe (LS), Below Severe (BS), Above Severe (AS) and More Severe (MS) and are shown in Fig. 6.

After obtaining the severity indices of all the voltage stability indices the Overall Severity Index ($OSI_{VSI}$) of the bus voltage stability index for a particular line outage is obtained using the expression

$$ OSI_{VSI} = \sum w_{VSI} SI_{VSI} $$ \hfill (4)

The weighting coefficients used for the severity indices are $w_{VSI} = 0.20$ for VLS, 0.40 for LS, 0.60 for BS, 0.80 for AS and 1.00 for MS where $w_{VSI} = $ Weighting coefficient for a severity index, $SI_{VSI} =$ Severity Index of a pre/post -contingent quantity. The fuzzy rule base used to evaluate the severity of a network contingency is given in Table 1.

### 4. Severity index based on fuzzy logic composite criteria

The overall fuzzy logic composite criteria based severity index is obtained by using the parallel operated fuzzy inference systems, as shown in Fig. 7, for the pre/post contingency operating conditions. The overall severity index for line loading, voltage profiles, and voltage stability indices are added and the sum is used for contingency ranking which is designated as the Fuzzy Logic Composite Criteria (FLCC) based severity.
Table 1

<table>
<thead>
<tr>
<th>Line loadings</th>
<th>Voltage profiles</th>
<th>Voltage stability indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Output</td>
<td></td>
</tr>
<tr>
<td>LL</td>
<td>LS</td>
<td>LS</td>
</tr>
<tr>
<td>NL</td>
<td>BS</td>
<td>BS</td>
</tr>
<tr>
<td>FL</td>
<td>AS</td>
<td>AS</td>
</tr>
<tr>
<td>OL</td>
<td>MS</td>
<td>MS</td>
</tr>
<tr>
<td>LV</td>
<td>MS</td>
<td>MS</td>
</tr>
<tr>
<td>NV</td>
<td>BS</td>
<td>BS</td>
</tr>
<tr>
<td>OV</td>
<td>VLI</td>
<td>VLI</td>
</tr>
<tr>
<td>VLI</td>
<td>LI</td>
<td>LI</td>
</tr>
<tr>
<td>LI</td>
<td>HI</td>
<td>HI</td>
</tr>
<tr>
<td>MI</td>
<td>VHI</td>
<td>VHI</td>
</tr>
</tbody>
</table>

Fig. 7. Parallel operated fuzzy inference systems.

index. When the overall FLCC severity index for each contingency in the contingency list has been figured out, the overall severity indices for those contingency cases with a severity index exceeding a pre-specified value are listed out and ranked according to the fuzzy logic composite criteria based severity index.

5. Severity index based on Line Loading

The severity of a contingency to line overload may be expressed in terms of the following severity index, which express the stress on the power system in the post contingency period [22]:

\[
\text{Severity index} L F_{SI} = \sum_{l \in L_0} \left( \frac{S_l}{S_{l \text{max}}} \right)^{2m} \tag{5}
\]

where

- \( S_l \) = flow in line \( l \) (MVA)
- \( S_{l \text{max}} \) = rating of the line \( l \) (MVA)
- \( L_0 \) = set of overloaded lines
- \( m \) = integer exponent

The line flows in Eq. (5) are obtained from Newton–Raphson load-flow calculations. While using the above severity index for security assessment, only the overloaded lines are considered. For IEEE 14-bus, IEEE30-bus and IEEE 118-bus systems considered in this work, the value of \( m \) has been fixed as 1.

6. Mathematical Formulation of the OPF Problem

The conventional formulation of the optimal-power-flow problem determines the optimal settings of control variables such as real power generations, generator terminal voltages, transformer tap settings and shunt compensation. The OPF minimizes an objective function such as generation cost, active power loss and total severity index, subject to several equality and inequality constraints. The problem can be mathematically modeled as follows:

\[
\begin{align*}
\text{Min} & \quad F(x, u) \\
\text{subject to} & \quad g(x, u) = 0 \\
& \quad h_{\text{min}} \leq h(x, u) \leq h_{\text{max}}
\end{align*} \tag{6}
\]

where vector \( x \) denotes the state variables of a power system network that contains the slack bus real power output \( (P_{G1}) \), voltage magnitudes and phase angles of the load buses \( (V_i, \delta_i) \), and generator reactive power outputs \( (Q_{G}) \). Vector \( u \) represents control variables that consist of real power generation levels \( (P_{GN}) \) and generator voltages magnitudes \( (|V_{GN}|) \), transformer tap setting \( (T_K) \), and reactive power injections \( (Q_{CK}) \) due to volt-ampere reactive (VAR) compensations; i.e.,

\[
u = [P_{G2}..., P_{GN}, V_{G1}..., V_{GN}, T_1..., T_{NT}, Q_{C1}..., Q_{CS}]
\tag{9}\]

where \( N \) = number of generator buses,
\( NT \) = number of tap changing transformers
\( CS \) = number of shunt reactive power injections.

6.1. Objective functions

The OPF problem has three types of objective function:
Objective Function I: \( \text{Min} \) 
\[ F_T = \sum (a_i P_{gi}^2 + b_i P_{gi} + c_i) \]  

Objective Function II: \( \text{Min} \) 
\[ LF_{SI} = \sum_{i \in L_o} \left( \frac{S_i}{S_i^{\text{max}}} \right)^{2m} \]  

Objective Function III: 
\[ \text{Min} FLCC = \text{Min} \ (OSI_{LL} + OSI_{VP} + OSI_{SI}) \]  

6.2. Constraints

The OPF problem has two categories of constraints:

**Equality Constraints**: These are the sets of nonlinear power flow equations that govern the power system, i.e.,

\[ P_{Gi} - P_{Di} = \sum_{j=1}^{n} |V_i||V_j||Y_{ij}| \cos(\theta_{ij} - \delta_i + \delta_j) = 0 \]  

\[ Q_{Gi} - Q_{Di} + \sum_{j=1}^{n} |V_i||V_j||Y_{ij}| \sin(\theta_{ij} - \delta_i + \delta_j) = 0 \]  

where \( P_{Gi} \) and \( Q_{Gi} \) are the real and reactive power outputs injected at bus respectively, the load demand at the same bus is represented by \( P_{Di} \) and \( Q_{Di} \), and elements of the bus admittance matrix are represented by \( |Y_{ij}| \) and \( \theta_{ij} \).

**Inequality Constraints**: These are the set of constraints that represent the system operational and security limits like the bounds on the following:

1) generators real and reactive power outputs
\[ P_{G}^{\text{min}} \leq P_{Gi} \leq P_{G}^{\text{max}}, i = 1, \ldots, N \]  
\[ Q_{G}^{\text{min}} \leq Q_{Gi} \leq Q_{G}^{\text{max}}, i = 1, \ldots, N \]  

2) voltage magnitudes at each bus in the network
\[ V_i^{\text{min}} \leq V_i \leq V_i^{\text{max}}, i = 1, \ldots, NL \]  

3) transformer tap settings
\[ T_i^{\text{min}} \leq T_i \leq T_i^{\text{max}}, i = 1, \ldots, NT \]  

4) reactive power injections due to capacitor banks
\[ Q_{C}^{\text{min}} \leq Q_{Ci} \leq Q_{C}^{\text{max}}, i = 1, \ldots, CS \]  

5) transmission lines loading
\[ S_i \leq S_i^{\text{max}}, i = 1, \ldots, nl \]  

6) voltage stability index
\[ L_{ji} \leq L_{ji}^{\text{max}}, i = 1, \ldots, NL \]  

**Handling of Constraints**: There are different ways to handle constraints in evolutionary computation optimization algorithms. In this paper, the constraints are incorporated into fitness function by means of penalty function method, which is a penalty factor multiplied with the square of the violated value of variable is added to the objective function and any infeasible solution obtained is rejected.

To handle the inequality constraints of state variables including load bus voltage magnitudes and output variables with real power generation output at slack bus, reactive power generation output, and line loading, the extended objective function can be defined as:

\[ OF = \sum_{i=1}^{N} F_i(P_{Gi}) + K_p h(P_{G1}) + K_q \sum_{i=1}^{N} h(Q_{Gi}) + K_v \sum_{i=1}^{NL} h(|V_i|) + K_s \sum_{i=1}^{NL} h(|S_i|) \]  

where \( K_p, K_q, K_v, K_s \) are penalty constants for the real power generation at slack bus, the reactive power generation of all generator buses or PV buses and slack bus, the voltage magnitude of all load buses or PQ buses, and line or transformer loading, respectively. \( h(P_{G1}), h(Q_{Gi}), h(|V_i|), h(|S_i|) \) are the penalty function of the real power generation at slack bus, the reactive power generation of all PV buses and slack bus, the voltage magnitudes of all PQ buses, and line or transformer loading, respectively. NL is the number of PQ buses. The penalty function can be defined as:

\[ h(x) = (x - x_{\text{max}})^2, \text{ if } x > x_{\text{max}} \]  
\[ = (x_{\text{min}} - x)^2, \text{ if } x < x_{\text{min}} \]  
\[ = 0, \text{ if } x_{\text{min}} \leq x \leq x_{\text{max}} \]  

where \( h(x) \) is the penalty function of variable \( x, x_{\text{max}} \) and \( x_{\text{min}} \) are the upper limit and lower limit of variable \( x \), respectively.

7. Overview of particle swarm optimization

Particle swarm optimization is an effective evolutionary computation techniques based on swarm intelligence. In 1995, Kennedy and Eberhart first introduced
the PSO method, which is motivated by the social behavior of organisms, such as fish schooling and birds flocking [15]. PSO, as an optimization tool, provides a population based search procedure in which individuals called particles change their positions with time. In a PSO system, particles fly around a d-dimensional problem space. During flight, each particle adjusts its position according to its own experience as well as by the best experiences of other neighboring particles. The basic elements of PSO technique as shown in Fig. 8 are briefly defined as follows:

7.1. Particle position, $X_i$

Each individual represents a candidate solution within the population and it is represented by a d-dimensional vector. Let us consider $X_i = (X_{i1}, X_{i2}, \ldots, X_{id})$ be the position of the $i^{th}$ particle.

7.2. Particle velocity, $V_i$

It is the velocity of the moving particles represented by a d-dimensional vector. The velocity of the $i^{th}$ particle is given by $V_i = (V_{i1}, V_{i2}, \ldots, V_{id})$ and it is bounded between the limits $V_{id}^{\min} \leq V_{id} \leq V_{id}^{\max}$.

7.3. Individual best, $P_{best_i}$

When a particle moves through the search space, it compares its fitness value at the current position to the best previous fitness value. The best position of the $i^{th}$ particle that is associated with the best fitness encountered so far is called $P_{best_i}$ and its vector representation is given by $P_{best_i} = (P_{i1}, P_{i2}, \ldots, P_{id})$. The fitness of the objective function for the $P_{best}$ of the $i^{th}$ particle is determined by the following relation.

![Flow chart of PSO method.](image)
\[ F(P_i) \leq F(X_i), \quad i = 1, 2, \ldots, d. \] (24)

7.4. Global best, \( g_{best} \)

It is the best position among all individual best positions achieved so far and is given by \( g_{best} = (P_{g1}, P_{g2}, \ldots, P_{gid}) \). The global best can be determined by
\[ F(P_{gi}) \leq F(P_i), \quad i = 1, 2, \ldots, d. \] (25)

7.5. Velocity updation

Using the global best and individual best of each particle, the \( i^{th} \) particle velocity in the \( d^{th} \) dimension is updated according to the following equation.
\[ V_{id}^{k+1} = W \ast V_{id}^k + C_1 \ast rand1 \ast (P_{id} - X_{id}) + C_2 \ast rand2 \ast (P_{gid} - X_{id}) \] (26)

where \( C_1 \) and \( C_2 \) are the acceleration constants, which represent the weighting of stochastic acceleration terms that pull each particle towards \( P_{best} \) and \( g_{best} \) positions. \( k \) represents the current iteration and rand1 and rand2 are two random numbers in the range \([0,1]\). The second term of Eq. (26) represents the cognitive part of PSO where the particle changes its velocity based on its own thinking and memory. The third term represents the social part of PSO where the particle changes its velocity based on the social-psychological adaptation of knowledge. If a particle violates the velocity limits, its velocity equal to the limit. Inertia weight \( W \) is a control parameter that is used to control the impact of the previous velocities on the current one. Hence, it influences the trade-off between the global and local exploration abilities of the particles.

7.6. Position updation

Based on the updated velocities, each particle changes its position according to the following equation:
\[ X_{id}^{k+1} = X_{id}^k + V_{id}^{k+1} \] (27)

7.7. Stopping criteria

It is the condition under which the search process will terminate. The search will terminate if the number of iterations reaches the maximum allowable number.

7.8. Step by step procedure for PSO-based OPF algorithm

\textbf{Step 1:} Assume the population size and maximum number of generations.

\textbf{Step 2:} Initialize the particle position vector between the limits of each control variable and depending on population size. The velocity vector of each particle is also initialized.

\textbf{Step 3:} Calculate the value of objective function of each generation.

\textbf{Step 4:} Obtain the values of \( P_{best} \) (the control variables corresponding to the minimum objective function value in each generation), and \( g_{best} \) (minimum objective function value among all generations), respectively.

\textbf{Step 5:} Update the velocity vector according to Eq. (26) and check the control variable’s limits violation. If there is any violation, set the value of the velocity vector corresponding to their limits.

\textbf{Step 6:} Update the particle position vector according to Eq. (27).

\textbf{Step 7:} If the value of control variables (new \( P_{best} \)) corresponding to minimum value objective function of current generation is less than the previous \( P_{best} \), then the current value is set to \( P_{best} \). If the \( P_{best} \) is better than \( g_{best} \), that value is set to be \( g_{best} \).

\textbf{Step 8:} If the number of generations reaches the maximum value, then go to the next step. Otherwise, go to step 3.

\textbf{Step 9:} The individual that generates the latest \( g_{best} \) is the optimal set of control variables with the global minimum value of the objective function.

8. Results and discussions

8.1. Results for IEEE 14-bus test system

This Section present the details of the study carried out on IEEE 14-test system for security enhancement. The network and load data for this system are taken from [19]. A modified IEEE 14-bus test system as shown in Fig. 9 is used in this work. It is comprised of 14 buses, 17 lines, 5 generators, 0 synchronous condensers, 2 shunt capacitors, 11 loads and 3 transformers. The generator cost coefficients along with the real and reactive power generation upper and lower limits
8.1.1. Case 1: Base case condition

The optimal values of control variables along with the real-power generation of the slack busbar generator are given in Table 2. The minimum cost obtained with the proposed PSO algorithm is $839.2236/h, which is less than the minimum generation cost of $839.97 obtained with MATPOWER [19] which uses the sequential quadratic programming(SQP) method. Also, it was found that all the state variables satisfy the lower and upper limits.

For comparison, the OPF problem was solved using an Evolutionary Programming (EP) method [20] with population size of 20 and 250 generations. All the solutions satisfy the constraints on reactive power generation limits and line flow limits. The convergence of generation cost is shown in Fig. 10. From the Fig. 10, it can be observed that the PSO took approximately 25 iterations to reach the same production cost reached by the EP. This shows that the proposed algorithm occupies less computer space and takes less time to reach the optimal solution.

8.1.2. Case 2: Contingency ranking using fuzzy approach

In this case, the proposed fuzzy logic approach is applied for the network contingency ranking with the optimal base case control variables under base-load conditions to identify the harmful contingencies. From the contingency analysis, it was found that line outages 2–3, 1–2, and 3–6 have resulted in heavy overload on other lines and are ranked as top 3 contingencies and are given in Table 3.

The severity indices of line loadings, voltage profiles, and voltage stability indices for the top 10 network contingencies are obtained and are given in Table 4.
Table 4
Contingency Ranking of the 14-bus system

<table>
<thead>
<tr>
<th>Contingency</th>
<th>OSI_{LL}</th>
<th>OSI_{VP}</th>
<th>OSI_{VS}</th>
<th>FLCC ((OSI_{LL} + OSI_{VP} + OSI_{VS}))</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2–3</td>
<td>419.1160</td>
<td>631.4070</td>
<td>36.1539</td>
<td>1086.700</td>
<td>1</td>
</tr>
<tr>
<td>1–2</td>
<td>320.7676</td>
<td>631.6508</td>
<td>36.1539</td>
<td>988.5723</td>
<td>2</td>
</tr>
<tr>
<td>3–6</td>
<td>218.5911</td>
<td>685.5359</td>
<td>36.1539</td>
<td>940.2810</td>
<td>3</td>
</tr>
<tr>
<td>13–14</td>
<td>227.3063</td>
<td>672.0200</td>
<td>36.1539</td>
<td>935.4804</td>
<td>4</td>
</tr>
<tr>
<td>10–11</td>
<td>246.0711</td>
<td>645.7325</td>
<td>36.1539</td>
<td>927.9576</td>
<td>5</td>
</tr>
<tr>
<td>1–8</td>
<td>268.1070</td>
<td>619.3989</td>
<td>36.1539</td>
<td>923.6598</td>
<td>6</td>
</tr>
<tr>
<td>12–13</td>
<td>239.5745</td>
<td>644.5590</td>
<td>36.1539</td>
<td>920.2875</td>
<td>7</td>
</tr>
<tr>
<td>7–9</td>
<td>259.7757</td>
<td>623.2713</td>
<td>36.1539</td>
<td>919.2090</td>
<td>8</td>
</tr>
<tr>
<td>9–14</td>
<td>228.0047</td>
<td>639.9695</td>
<td>36.1539</td>
<td>904.1282</td>
<td>9</td>
</tr>
<tr>
<td>2–8</td>
<td>205.6971</td>
<td>660.9521</td>
<td>36.1539</td>
<td>902.8031</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 5
Number of lines/buses under each severity category

<table>
<thead>
<tr>
<th>Contingency</th>
<th>Line loading</th>
<th>Voltage profile</th>
<th>Voltage stability indices</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LS BS AS MS</td>
<td>BS AS MS</td>
<td>VLS BS AS MS</td>
<td></td>
</tr>
<tr>
<td>2–3</td>
<td>12 3 2 2</td>
<td>0 3 6</td>
<td>9 0 0 0 0 0</td>
<td>1</td>
</tr>
<tr>
<td>1–2</td>
<td>14 2 2 1</td>
<td>0 3 6</td>
<td>9 0 0 0 0 0</td>
<td>2</td>
</tr>
<tr>
<td>3–6</td>
<td>14 4 1 0</td>
<td>0 2 7</td>
<td>9 0 0 0 0 0</td>
<td>3</td>
</tr>
<tr>
<td>13–14</td>
<td>14 4 1 0</td>
<td>0 2 7</td>
<td>9 0 0 0 0 0</td>
<td>4</td>
</tr>
<tr>
<td>10–11</td>
<td>13 5 1 0</td>
<td>0 3 6</td>
<td>9 0 0 0 0 0</td>
<td>5</td>
</tr>
<tr>
<td>1–8</td>
<td>13 4 2 0</td>
<td>0 3 6</td>
<td>9 0 0 0 0 0</td>
<td>6</td>
</tr>
<tr>
<td>12–13</td>
<td>13 5 1 0</td>
<td>0 3 6</td>
<td>9 0 0 0 0 0</td>
<td>7</td>
</tr>
<tr>
<td>7–9</td>
<td>12 6 1 0</td>
<td>0 3 6</td>
<td>9 0 0 0 0 0</td>
<td>8</td>
</tr>
<tr>
<td>9–14</td>
<td>14 4 1 0</td>
<td>0 3 6</td>
<td>9 0 0 0 0 0</td>
<td>9</td>
</tr>
<tr>
<td>2–8</td>
<td>15 3 1 0</td>
<td>0 2 7</td>
<td>9 0 0 0 0 0</td>
<td>10</td>
</tr>
</tbody>
</table>

Fig. 10. Variation of generation cost.

The number of lines, the number of load buses under different severity categories for the top 10 network contingencies are given in Table 5. From Tables 4 and 5, it is observed that line outage 2–3 is the most severe one, and results in maximum total severity index, i.e., FLCC based severity index compared to other line outages.

8.1.3. Case 3: OPF for overload alleviation

To test the ability of the proposed PSO algorithm for solving optimal power flow problem, it was applied un-
under the selected three most severe network contingencies. Three objective functions are considered for the minimization using the proposed PSO algorithm. The algorithm was run for a maximum of 250 generations with a population size of 20 and was made to stop at the end of the total generations.

Tables 6–8 present the result for IEEE 14-bus test system. Table 6 presents the optimal settings of the control-variables such as real power generations, tap changing transformers, generator voltages, and shunt compensations with the three objective functions under the rank 1 network contingency. Table 7 presents various performance parameters under the three most severe network contingencies and with the three objective functions. Table 8 presents the number of lines and number of load buses under different severity categories for the three most severe network contingencies and with the three objective functions. From the Table 8 it is evident that the overloading of the transmission lines has been completely alleviated under all the selected network contingencies. This shows the effectiveness of the proposed algorithm for overload alleviation.

<table>
<thead>
<tr>
<th>Generator bus no.</th>
<th>Limits</th>
<th>Objective functions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Real power</td>
<td>P_{G1}</td>
<td>0</td>
</tr>
<tr>
<td>generation</td>
<td>P_{G2}</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>P_{G3}</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>P_{G4}</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>P_{G5}</td>
<td>0.7</td>
</tr>
<tr>
<td>Generator voltages</td>
<td>V_{G1}</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>V_{G2}</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>V_{G3}</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>V_{G4}</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>V_{G5}</td>
<td>0.95</td>
</tr>
<tr>
<td>Transformer tap</td>
<td>Tap-1</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Tap-2</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Tap-3</td>
<td>0.95</td>
</tr>
<tr>
<td>Shunt compensation</td>
<td>Qsvc6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Qsvc8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Qsvc9</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Qsvc10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Qsvc14</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 7</th>
<th>Performance parameters of IEEE 14-bus system under the rank 1, 2 and 3 network contingencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contingency</td>
<td>Parameters</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>2–3</td>
<td>P-Loss</td>
</tr>
<tr>
<td></td>
<td>Cost ($/hr)</td>
</tr>
<tr>
<td></td>
<td>OSI_{LL}</td>
</tr>
<tr>
<td></td>
<td>OSI_{V}</td>
</tr>
<tr>
<td></td>
<td>OSI_{VST}</td>
</tr>
<tr>
<td></td>
<td>FLCC</td>
</tr>
<tr>
<td>1–2</td>
<td>P-Loss</td>
</tr>
<tr>
<td></td>
<td>Cost ($/hr)</td>
</tr>
<tr>
<td></td>
<td>OSI_{LL}</td>
</tr>
<tr>
<td></td>
<td>OSI_{V}</td>
</tr>
<tr>
<td></td>
<td>OSI_{VST}</td>
</tr>
<tr>
<td></td>
<td>FLCC</td>
</tr>
<tr>
<td>3–6</td>
<td>P-Loss</td>
</tr>
<tr>
<td></td>
<td>Cost ($/hr)</td>
</tr>
<tr>
<td></td>
<td>OSI_{LL}</td>
</tr>
<tr>
<td></td>
<td>OSI_{V}</td>
</tr>
<tr>
<td></td>
<td>OSI_{VST}</td>
</tr>
<tr>
<td></td>
<td>FLCC</td>
</tr>
</tbody>
</table>
Table 8

<table>
<thead>
<tr>
<th>Contingency</th>
<th>Objective functions</th>
<th>Method</th>
<th>Line loadings</th>
<th>Voltage profiles</th>
<th>Bus voltage stability indices</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS</td>
<td>BS</td>
<td>AS</td>
</tr>
<tr>
<td>2-3</td>
<td>Objective Function 1</td>
<td>EP</td>
<td>12</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO</td>
<td>12</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Objective Function 2</td>
<td>EP</td>
<td>15</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO</td>
<td>16</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Objective Function 3</td>
<td>EP</td>
<td>16</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO</td>
<td>16</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1-2</td>
<td>Objective Function 1</td>
<td>EP</td>
<td>14</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO</td>
<td>14</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Objective Function 2</td>
<td>EP</td>
<td>17</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO</td>
<td>17</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Objective Function 3</td>
<td>EP</td>
<td>17</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO</td>
<td>17</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3-6</td>
<td>Objective Function 1</td>
<td>EP</td>
<td>13</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO</td>
<td>14</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Objective Function 2</td>
<td>EP</td>
<td>17</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO</td>
<td>17</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Objective Function 3</td>
<td>EP</td>
<td>18</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO</td>
<td>18</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO</td>
<td>17</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 11. Comparison of Line Loadings for three objective functions (cost, LF_{SI}, FLCC) by EP method under line contingency of 2-3.

From Table 8, it can also be observed that, the PSO and EP methods based OPF are able to alleviate the overloads effectively with the FLCC objective function compared to the other two objective functions. Figures 11 and 12 show the percentage line loadings after the optimization by EP and PSO methods with the three objective functions.

8.2. Results for IEEE 30-bus test system

This Section presents the details of the study carried out on IEEE-30 bus test system for security en-
It is comprised of 30 buses, 41 lines, 6 generators, 0 synchronous condensers, 2 shunt capacitors, 21 loads and 4 transformers. The proposed algorithm was implemented in MATLAB computing environment with Pentium-IV, 2.66 GHz computer with 512 MB RAM.

Three different cases were considered for the study. In the first case, the proposed PSO-based algorithm was applied to obtain the optimal-control variables in the IEEE 30-bus system under base case load conditions. In the second case, the proposed fuzzy logic algorithm was applied to network contingency ranking. In the third case, the proposed fuzzy logic composite criteria and PSO algorithm was applied to alleviate overloads under selected set of severe network contingency through the solution of optimal power flow.

8.2.1. Case 1: Base case condition

The optimal values of control variables along with the real-power generation of the slack bus-bar generator are given in Table 9. The minimum cost obtained with the proposed PSO algorithm is $800.966/h, which is less than the minimum generation cost of $803.1916/h obtained with interior point method. Also, it was found that all the state variables satisfy the lower and upper limits.

For comparison, the OPF problem was solved using an evolutionary programming method [20] with the
population size of 20 and 250 generations. All the solutions satisfy the constraints on reactive power generation limits and line flow limits. The convergence of generation cost is shown in Fig. 14. From the Fig. 14, it can be observed that the PSO took approximately 60 generations to reach the same production cost reached by EP. This shows that the proposed PSO algorithm occupies less computer space and takes less time to reach the optimal solution.

8.2.2. Case 2: Contingency ranking using fuzzy approach

In this case, the proposed fuzzy logic approach for network contingency ranking process was applied with the optimal base case control variables under base case load conditions to identify the harmful contingencies. From the contingency analysis, it was found that line outages 2–5, 11–13, 8–11 have resulted in heavy overload on other lines and are ranked as top 3 contingencies and are given in Table 10. The severity indices of line loadings, voltage profiles, and voltage stability indices for the top 10 network contingencies are calculated and are given in Table 11. The number of lines, load bus voltage profiles, and load bus voltage stability indices under each severity category for the top 10 network contingencies are given in Table 12. From the
Table 11  
Contingency ranking of IEEE 30-bus system

<table>
<thead>
<tr>
<th>Contingency</th>
<th>OSI_LL</th>
<th>OSI_LP</th>
<th>OSI_VI</th>
<th>FLCC_Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-5</td>
<td>649.399</td>
<td>1650.7</td>
<td>96.4105</td>
<td>2396.5</td>
</tr>
<tr>
<td>11-13</td>
<td>547.2608</td>
<td>1668.5</td>
<td>96.4105</td>
<td>2312.2</td>
</tr>
<tr>
<td>8-11</td>
<td>531.8282</td>
<td>1654.0</td>
<td>96.4105</td>
<td>2282.2</td>
</tr>
<tr>
<td>13-7</td>
<td>376.3138</td>
<td>1798.4</td>
<td>96.4105</td>
<td>2271.2</td>
</tr>
<tr>
<td>1-8</td>
<td>536.9544</td>
<td>1632.9</td>
<td>96.4105</td>
<td>2266.3</td>
</tr>
<tr>
<td>27-30</td>
<td>439.7578</td>
<td>1720.5</td>
<td>103.3392</td>
<td>2263.6</td>
</tr>
<tr>
<td>13-3</td>
<td>391.6967</td>
<td>1774.8</td>
<td>96.4105</td>
<td>2262.9</td>
</tr>
<tr>
<td>24-25</td>
<td>371.9588</td>
<td>1784.3</td>
<td>96.4105</td>
<td>2252.7</td>
</tr>
<tr>
<td>1-2</td>
<td>586.5918</td>
<td>1563.3</td>
<td>96.4105</td>
<td>2246.3</td>
</tr>
<tr>
<td>27-29</td>
<td>416.3521</td>
<td>1721.2</td>
<td>96.4105</td>
<td>2233.9</td>
</tr>
</tbody>
</table>

Table 12  
Number of lines/buses under different severity category before optimization

<table>
<thead>
<tr>
<th>Contingency</th>
<th>Line loading</th>
<th>Voltage profile</th>
<th>Voltage stability indices</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS BS AS MS</td>
<td>LS BS AS MS</td>
<td>VLS BS AS MS</td>
<td>VLS BS AS MS</td>
<td></td>
</tr>
<tr>
<td>2-5</td>
<td>28 8 2 2</td>
<td>0 8 16</td>
<td>24 0 0 0 0 0</td>
<td>1</td>
</tr>
<tr>
<td>11-13</td>
<td>31 7 0 2</td>
<td>0 8 16</td>
<td>24 0 0 0 0 0</td>
<td>2</td>
</tr>
<tr>
<td>8-11</td>
<td>33 4 1 2</td>
<td>0 8 16</td>
<td>24 0 0 0 0 0</td>
<td>3</td>
</tr>
<tr>
<td>13-7</td>
<td>34 5 1 0</td>
<td>0 6 18</td>
<td>24 0 0 0 0 0</td>
<td>4</td>
</tr>
<tr>
<td>1-8</td>
<td>33 4 1 2</td>
<td>0 8 16</td>
<td>24 0 0 0 0 0</td>
<td>5</td>
</tr>
<tr>
<td>27-30</td>
<td>32 6 2 0</td>
<td>0 7 17</td>
<td>23 1 0 0 0 0</td>
<td>6</td>
</tr>
<tr>
<td>13-3</td>
<td>33 6 1 0</td>
<td>0 7 17</td>
<td>24 0 0 0 0 0</td>
<td>7</td>
</tr>
<tr>
<td>24-25</td>
<td>34 5 1 0</td>
<td>0 6 18</td>
<td>24 0 0 0 0 0</td>
<td>8</td>
</tr>
<tr>
<td>1-2</td>
<td>33 4 0 3</td>
<td>0 9 15</td>
<td>24 0 0 0 0 0</td>
<td>9</td>
</tr>
<tr>
<td>27-29</td>
<td>33 5 2 0</td>
<td>0 7 17</td>
<td>24 0 0 0 0 0</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 13  
Optimal settings of control variables for IEEE 30-bus system under rank 1 contingency

<table>
<thead>
<tr>
<th>Generator</th>
<th>Limits</th>
<th>Objective functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_G1</td>
<td>0.5</td>
<td>EP PSO</td>
</tr>
<tr>
<td>P_G2</td>
<td>0.2</td>
<td>EP PSO</td>
</tr>
<tr>
<td>P_G3</td>
<td>0.1</td>
<td>EP PSO</td>
</tr>
<tr>
<td>P_G4</td>
<td>0.1</td>
<td>EP PSO</td>
</tr>
<tr>
<td>P_G5</td>
<td>0.15</td>
<td>EP PSO</td>
</tr>
<tr>
<td>P_G6</td>
<td>0.12</td>
<td>EP PSO</td>
</tr>
<tr>
<td>V_G2</td>
<td>0.95</td>
<td>EP PSO</td>
</tr>
<tr>
<td>V_G3</td>
<td>0.95</td>
<td>EP PSO</td>
</tr>
<tr>
<td>V_G4</td>
<td>0.95</td>
<td>EP PSO</td>
</tr>
<tr>
<td>V_G5</td>
<td>0.95</td>
<td>EP PSO</td>
</tr>
<tr>
<td>V_G6</td>
<td>0.95</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Tap-1</td>
<td>0.95</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Tap-2</td>
<td>0.95</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Tap-3</td>
<td>0.95</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Tap-4</td>
<td>0.95</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Qsvc10</td>
<td>0.00</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Qsvc12</td>
<td>0.00</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Qsvc15</td>
<td>0.00</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Qsvc17</td>
<td>0.00</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Qsvc20</td>
<td>0.00</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Qsvc21</td>
<td>0.00</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Qsvc23</td>
<td>0.00</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Qsvc24</td>
<td>0.00</td>
<td>EP PSO</td>
</tr>
<tr>
<td>Qsvc25</td>
<td>0.00</td>
<td>EP PSO</td>
</tr>
</tbody>
</table>

K. Vaisakh et al. / Power system security enhancement using fuzzy logic composite criteria and particle swarm optimization algorithm 217
Tables 11 and 12 it is found that line outage 2–5 is the most severe one, and results in maximum total severity index compared to other three lines.

8.2.3. Case 3: OPF for overload alleviation

To test the ability of the proposed PSO algorithm for solving optimal power flow problem, to alleviate overloads, it was applied under the selected three most severe network contingencies. The same three objective functions are considered for the minimization using the proposed PSO algorithm. The algorithm was run for a maximum of 250 generations with a population size of 20 and was forced to stop after 250 generations.

Tables 14–16 present the results for IEEE 30-bus test system. Table 14 presents the optimal settings of the control-variables under the rank 1 network contingency with the three objective functions. Table 15 presents various performance parameters under the three most severe network contingencies and with the three objective functions. Table 16 presents the number of lines and number of load buses under different severity categories for the three most severe network contingencies and for all the three objective functions. From the

<table>
<thead>
<tr>
<th>Contingency</th>
<th>Objective function</th>
<th>Linear loadings</th>
<th>Bus voltage profiles</th>
<th>Bus voltage stability indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-5</td>
<td>Objective Function 1</td>
<td>EP 29 9 1 1</td>
<td>0 20 4</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO 27 11 0 2</td>
<td>0 22 0</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>Objective Function 2</td>
<td>EP 32 8 0 0</td>
<td>0 24 0</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO 32 8 0 0</td>
<td>0 24 0</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>Objective Function 3</td>
<td>EP 34 6 0 0</td>
<td>0 24 0</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO 34 6 0 0</td>
<td>0 24 0</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td>11-13</td>
<td>Objective Function 1</td>
<td>EP 31 7 2 0</td>
<td>0 21 3</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO 31 7 1 1</td>
<td>0 19 5</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>Objective Function 2</td>
<td>EP 35 5 0 0</td>
<td>0 24 0</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO 36 4 0 0</td>
<td>0 24 0</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>Objective Function 3</td>
<td>EP 36 4 0 0</td>
<td>0 24 0</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO 37 3 0 0</td>
<td>0 24 0</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td>8-11</td>
<td>Objective Function 1</td>
<td>EP 33 6 1 0</td>
<td>0 21 3</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO 32 7 0 1</td>
<td>0 23 1</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>Objective Function 2</td>
<td>EP 34 6 0 0</td>
<td>0 23 1</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO 35 5 0 0</td>
<td>0 23 1</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>Objective Function 3</td>
<td>EP 36 4 0 0</td>
<td>0 23 1</td>
<td>24 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSO 36 4 0 0</td>
<td>0 23 1</td>
<td>24 0 0 0 0</td>
</tr>
</tbody>
</table>
Fig. 15. Comparison of Line Loadings for three objective functions (cost, LF_{SI}, FLCC) by EP method under line contingency of 2-5.

Fig. 16. Comparison of Line Loadings for three objective functions (cost, LF_{SI}, FLCC) by PSO method under line contingency of 2-5.
Fig. 17. Comparison of convergence characteristics of objective function-1 under the selected three network contingencies.

Fig. 18. Comparison of convergence characteristics of objective function-2 under the selected three network contingencies.

Fig. 19. Comparison of convergence characteristics of objective function-3 under the selected three network contingencies.

Table 16 it is evident that the overloading of the transmission lines has been completely alleviated under all the selected network contingencies. This shows the effectiveness of the proposed algorithm for overload alleviation.

From Table 16, it can also be observed that, in solving the OPF with the FLCC objective function, the PSO and EP methods are able to alleviate the overloads effectively compared to the other two objective functions.

Figures 15–16 show the percentage line loading after the optimization by EP and PSO methods with the three objective functions. Figures 17–19 show the convergence characteristics of the three objective functions under the selected three network contingencies.

8.3. Results of IEEE-118 bus test system

In this case, the proposed algorithm was also applied to alleviate the line overload under contingency condition in the IEEE 118-bus system. The test system has 54 generator buses and 186 transmission lines. The real-power generation $P_g$ of the generators and the line rating of the transmission line is taken from ref. [22]. All other data are the same as the standard IEEE 118-bus data [23]. Contingency analysis was conducted on this system using the proposed fuzzy logic approach to identify the harmful contingencies. The network contingencies which caused overloading of lines are given in Table 17. From the contingency analysis, it was found that line outages 8–59, 92–91, and 64-11 have resulted in heavy overload on other lines and are ranked as top 3 contingencies and are given in Table 18.

The severity indices of line loadings, voltage profiles, and voltage stability indices for the top 3 network contingencies are calculated for each contingency and are given in Table 19. The number of lines, voltage profiles, and voltage stability indices of load buses under each severity category for the top 3 network contingencies are given in Table 20. From the Tables 19 and 20 it is found that line outage 8–59 is the most severe one, and results in maximum total severity index compared to other three lines.

The proposed PSO algorithm for solving optimal power flow problem, to alleviate overloads, was applied
under the selected top 1 network contingency. The variation of fitness function with respect to number of iterations is shown in Fig. 20. Tables 17–19 present the results for IEEE 118-bus test system. Table 17 presents the summary of the overloaded lines under top 6 network contingencies. Table 18 presents various severity indices under the top 3 network contingencies. Table 19 presents the number of lines and number of load

<table>
<thead>
<tr>
<th>Contingency</th>
<th>OSI_{LL}</th>
<th>OSI_{VP}</th>
<th>OSI_{VI}</th>
<th>FLCC</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-59</td>
<td>2053.4</td>
<td>2611.0</td>
<td>257.0947</td>
<td>4921.5</td>
<td>1</td>
</tr>
<tr>
<td>92-91</td>
<td>1686.0</td>
<td>2611.0</td>
<td>257.0947</td>
<td>4554.1</td>
<td>2</td>
</tr>
<tr>
<td>64-11</td>
<td>1680.8</td>
<td>2610.9</td>
<td>257.0947</td>
<td>4548.8</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contingency</th>
<th>Line loading</th>
<th>Voltage profile</th>
<th>Voltage stability indices</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LS</td>
<td>BS</td>
<td>AS</td>
<td>MS</td>
</tr>
<tr>
<td>8-59</td>
<td>147</td>
<td>25</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>92-91</td>
<td>147</td>
<td>30</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>64-11</td>
<td>151</td>
<td>25</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Fig. 20. Variation of fitness value.

Fig. 21. Line Loadings of IEEE 118-bus system.
buses under different severity categories for the three most severe network contingencies and for all the three objective functions. From the Table 19 it is evident that the overloading of the transmission lines has been completely alleviated under all the selected network contingencies. This shows the effectiveness of the proposed algorithm for overload alleviation. From the Table 19, it can also be observed that, in solving the OPF with the FLCC objective function, the PSO and EP methods are able to alleviate the overloads effectively compared to the other two objective functions. Figures 21–23 show the percentage line loadings, voltage profiles, voltage stability indices after the optimization.

8.4. Summary and recommendations

Based on the observed result for the three test systems it can be concluded that the PSO algorithm with the FLCC objective function gives more promising result than line flow based and cost based objective functions. The results obtained by the PSO algorithm with three objective functions are better than the results obtained by evolutionary programming method. The proposed fuzzy logic composite criteria based method of contingency ranking is able to distinguish clearly the actual severity of the system considering line loading, voltage profiles and voltage stability index from one contingency to other. Hence the proposed method eliminates the problem of the masking effect. Therefore, the PSO algorithm with the FLCC based severity index is recommended to be used to the extent possible.

9. Conclusions

This paper has proposed a PSO method and fuzzy logic composite criteria to aid power-system operators. In this paper, in addition to real power loadings and bus voltages, the voltage stability indices at the load buses were also used as the post-contingent quantities to evaluate the network composite contingency ranking. These post-contingent quantities are expressed in fuzzy set notation. Then the fuzzy rules employed in contingency ranking are compiled to reach the overall
system severity index. The proposed contingency ranking method eliminates the masking effect effectively. The proposed fuzzy approach has been tested on large networks and results indicate a clear ranking of contingencies. The application of this method for solving optimal power flow problems during the normal operation and the power system controls during contingencies have been presented. The line overloads were relieved through adjustment of generator outputs, generator voltages, tap changing transformers, and shunt compensation. The simulation results on IEEE 14, 30, and 118-bus systems have been presented for illustration purpose.

**List of symbols**

\( N_j \): Total number of busbars

\( N_{PQ} \): Number of load busbars

\( N_G \): Number of generators

\( S \): Slack bus

\( P_i, Q_i \): Real and reactive powers injected into network at bus i

\( V_i, V_j \): Voltage magnitude at busbars i and j

\( G_{ij}, B_{ij} \): Self conductance and susceptance of busbars i and j

\( P_{ij} \): Voltage-angle difference between bus i and j

\( P_{gi} \): Real power generation at bus i

\( Q_{gi} \): Reactive power at bus i

**References**


Defect detection in flat surface products using log-Gabor filters

A.S. Tolba\textsuperscript{a,∗}, Ahmad Atwan\textsuperscript{b}, N. Amanneddine\textsuperscript{a}, A. M. Mutawa\textsuperscript{c} and H. A. Khan\textsuperscript{a}

\textsuperscript{a}Arab Open University
\textsuperscript{b}Faculty of Computer and Information Sciences, Mansoura University, Egypt
\textsuperscript{c}Faculty of Engineering, Kuwait University, Kuwait

Abstract. This paper introduces a novel method for defect detection in homogeneous flat surface products. The coefficient of variation is used as a homogeneity measure for approximate defect localization and features extracted from the Log – Gabor filter bank response are used to accurately localize and detect the defect while reducing the complexity of Gabor based inspection approaches. The scanning window size and threshold parameter are the two major factors that affect the system performance. An adaptive technique is proposed for selecting the size of the scanning window and automating the selection of the threshold level. Compared to the Log-Gabor filters, the proposed combination resulted in speeding up the defect detection process by about ten times. The experimental results show that the proposed system gives promising results and is applicable for defect detection in homogeneous surfaces like paper, steel plates, ceramic tiles and foils.

Keywords: Automated visual inspection, defect detection, feature extraction, log-gabor filter bank, homogeneity measures

1. Introduction

Automated Visual Inspection (AVI) is the process of detecting, analyzing and classifying abnormal structures in a surface using machine vision techniques. The increasing competition in the industrial sector imposes high requirements on controlling the quality of flat surface products such as textiles, paper, steel slabs, glass, plastic films, foils, parquet slabs, ceramics, etc. [2]. Automation of the visual inspection process saves companies a lot of time and raises the quality of their products by avoiding the subjectivity, boredom, and slow speed of the traditional human based inspection process. When a defective product reaches the consumer, the company's reputation will suffer. AVI makes 100% quality control and documentation possible.

This paper presents a new method to speed up the defect detection process using Log-Gabor filters. Comparison with the results of Gabor filters shows the strength of the proposed method and its suitability for real-time product inspection. By introducing the use of homogeneity measure, our system focuses on the identified regions of interest, therefore reducing the computation cost. Also, by utilizing an adaptive sliding window, we have achieved faster implementation of the AVI system. In this work, we have implemented a new methodology for solving the problem of high computational overhead of Gabor-filters based defect detection.

This paper is organized as follows: Section 2 gives an overview of the most recent AVI approaches. Section 3 describes the role of homogeneity for faster defect detection. Section 4 presents the approximate defect localization method using the coefficient of variation. Section 5 describes Log-Gabor based feature sets used for accurate defect detection and localization. Section 6 presents the experimental results of defect detection. Finally, concluding remarks are given in Section 7.

2. Overview of existing automated visual inspection approaches

Automated Visual Inspection (AVI) systems are in great demand in the industry, since Human Visual In-
specification (HVI) systems suffer from various drawbacks such as fatigue and boredom and as such humans can detect only about 60% to 75% of the major defects in products [1]. Excellent reviews of recent advances in surface defect detection using texture analysis are provided [3–5]. These reviews have discussed statistical, structural, filter- and model-based approaches to AVI.

Mak et al. [4] have proposed a semi-supervised scheme for defect detection in textiles using the Optimal Gabor Filters. This scheme has resulted in an overall correct detection rate of 92.31%. Meihong Shi et al. [7], have presented a Simplified Pulse Coupled Neural Network (SPCNN) based defect with for defect segmentation of fabric defects. Detection results indicated better performance compared to defect segmentation based on OTSU thresholding and a Pulse Coupled Neural Network (PCNN) based method. Cuenca et al. [8], have presented a new texture descriptor for extraction of Local Semicover Pattern (LSP) features for defect detection and localization in fabrics. Experiments on samples from the Textile texture database (TILDA) [12] have shown that the computational costs of Local Semicover Pattern (LSP) are two orders of magnitude less than those of other more complex methods such as the Fractal Dimension, Wavelet Transform and Gabor Filters, and one order less than those of Gray-Level Co-occurrence (GLC) (only 32 grey levels were considered). Only the Difference Histogram method has a similar cost but with slightly poorer performance.

The results summarized in Table 1, clearly indicate that most of the comparative studies [3,8,14–19,21,22,28] related to texture classification, texture segmentation and defect detection cannot identify an optimal set of features to perform these tasks. Some features such as those based on GLCM and LBP are effective in texture characterization. Most of the comparative studies [3,8,13,14,19,21,22] have stressed the necessity of fusing multiple feature sets with different parameters to increase the accuracy and reduce the computational overhead. Although Gabor based defect detection approach is the most popular one, its computational complexity has limited its practical value. The aim of this paper is to enhance the performance of Log Gabor based approaches by processing only the areas of interest as selected by using the coefficient of variation for approximate defect localization. The proposed approach aims at addressing the following problems:

1. Reducing the computational cost drastically by inspecting only those parts with the highest probability of including a defect. Preprocessing is used to specify the Region of Interest (ROI) based on a fast homogeneity metric
2. Using a fast implementation of the adaptive sizing of the sliding window. Texel size varies from one product to the other. This problem has been tackled by automating the sliding window sizing process in our work.
3. Gray level image processing results also in low computational complexity compared to color image processing.
4. Localization of defects with the scanning moving window helps to classify the defects if needed.
5. Application dependency has been avoided using a trainable system that is trained on the nature of the normal product (textile, paper, steel slabs, foils) surface characteristics.

### 3. Adaptive sizing of the scanning sliding windows

Research results on defect detection have shown great variation in system performance with the size of the scanning window. To solve this problem the following new algorithm has been implemented for automatic determination of the sliding window size. The proposed algorithm is based on calculation of the coefficient of variation $c_v$ for successive blocks as given by:

$$c_v = \frac{\sigma}{\mu}$$

where, $\mu$ is the mean gray level of the block, $\sigma$ is the standard deviation of gray levels in a given block.

**Adaptive Sliding-Window Sizing Algorithm**

1. Determine the size $M \times N$ pixels of the whole image acquired by the imaging system
2. Initialize the sliding window size to $w_h \times w_w$ pixels, where $w_h$ is the window height and $w_w$ is the window width.
3. Set the iteration counter to 1
4. Iterate over the acquired image of a normal product starting the scan from the top-left corner to the bottom-right corner.
5. Compute the coefficient of variation $c_v(i)$ for the $i^{th}$ iteration for all scanned image windows.
6. Increase the window size to $(w_h + 20) \times (w_w + 20)$ pixels, and set counter = counter + 1
7. Repeat steps 4 to 6 for the next window size and calculate $c_v(i + 1)$ until the end of the scanned normal image is reached.
8. Stop when counter > 6
9. Find the window size corresponding to the minimum coefficient of variation and select that window as the most suitable window size.
Table 1

<table>
<thead>
<tr>
<th>Feature Extraction Algorithm</th>
<th>PCC</th>
<th>Rank</th>
<th>Reference</th>
<th>Feature Extraction Algorithm</th>
<th>PCC</th>
<th>Rank</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabor</td>
<td>87.89%</td>
<td>3</td>
<td>[13]</td>
<td>Multiresolution Histograms</td>
<td>N/A</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>Wavelet97</td>
<td>88.15%</td>
<td>1</td>
<td>Fourier power spectrum annuli</td>
<td>N/A</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelet54</td>
<td>88.02%</td>
<td>2</td>
<td>Gabor wavelet features</td>
<td>N/A</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daub4</td>
<td>75.65%</td>
<td>4</td>
<td>Daubechies wavelet packets features</td>
<td>N/A</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haar</td>
<td>73.57%</td>
<td>5</td>
<td>Auto-co-occurrence matrices $11 \times 11$ pixels</td>
<td>N/A</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordinary Histogram</td>
<td>77.54%</td>
<td>8</td>
<td>[14] Markov random field parameters $3 \times 3$ window</td>
<td>N/A</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBP</td>
<td>84.18%</td>
<td>7</td>
<td>GLCM</td>
<td>33.3% 5</td>
<td>27</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>GLCM</td>
<td>97.09%</td>
<td>1</td>
<td>FFT</td>
<td>48.3% 4</td>
<td>27</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>Gabor Filter</td>
<td>91.22%</td>
<td>6</td>
<td>Wavelet Transform</td>
<td>80%</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DWHT</td>
<td>95.8%</td>
<td>4</td>
<td>Gabor Transform</td>
<td>85%</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDCT</td>
<td>95.14%</td>
<td>5</td>
<td>Clustering</td>
<td>91.6%</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eigenfilter N $7 \times 7$</td>
<td>95.70%</td>
<td>3</td>
<td>Gabor</td>
<td>72.3% 2</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLCM</td>
<td>72.5%</td>
<td>3</td>
<td>Steerable Pyramids</td>
<td>69.06%</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBP</td>
<td>85.83%</td>
<td>1</td>
<td>Contour</td>
<td>71.53%</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texture Spectrum</td>
<td>82.17%</td>
<td>2</td>
<td>Gabor</td>
<td>77.2% 3</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLCM</td>
<td>95.05%</td>
<td>2</td>
<td>GLCM</td>
<td>96.9% 1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>Sample cv</th>
<th>Population MLE ($C_v$)</th>
<th>Confidence Interval of $C_v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI 90%</td>
<td>CI 95%</td>
<td>CI 99%</td>
<td>($LB = \text{Lower bound}, UB = \text{upper Bound}$)</td>
<td></td>
</tr>
<tr>
<td>90.7</td>
<td>2.73</td>
<td>0.0300</td>
<td>0.0725</td>
<td>LB = 0.01949</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UB = 0.01799</td>
</tr>
</tbody>
</table>

It has been found in our experiments that the most suitable window size is one that encompasses at least a complete Texel (a primitive texture element). Further research is in progress by the first author to investigate the interaction between the window size and defect size. Figure 2, shows the relation between the block size and the coefficient of variation for the sample given in Fig. 1.

The coefficient of variation is used to measure the relative sensitivity of the defect detection scheme to the variation in data, or feature sets. It is defined by the ratio of the standard deviation ($\sigma$) over the average ($\mu$). Since the sample coefficient of variation ($c_v$) could be a poor estimator [62] for the population coefficient of variation, $c_v$ is used here to specify the most stable defect detection scheme. For an accurate estimation of the population $c_v$ based on a sample of size $n \geq 5$, the maximum likelihood estimate is given by

$$c_v = \frac{\sigma}{\sqrt{n}\mu}$$  \hspace{1cm} (2)

The 95% confidence interval (CI) [63] for the Maximum Likelihood Estimate (MLE) of $c_v$ indicates the significance and reliability of the estimation of the coefficient of variation (Table 2).

4. Approximate defect localization using the coefficient of variation

Homogeneity reflects the uniformity of gray level distribution within a region of an image. It can be used both for selecting the training set in defect detection applications and guiding the defect detection process in the regions of interest (ROIs). In [27], homogeneity is defined as a composition of both the standard deviation and discontinuity of gray level intensities. Standard deviation $\sigma_i$ describes the contrast within a local region $i$ at image subdivision level $l$. Discontinuity is a measure of abrupt changes in gray levels and can be obtained by applying edge detectors to the corresponding region. Performing analysis of the homogeneity measure according to Fig. 3 and Table 3, the computational complexity of the processing could be drastically reduced, especially in the case of defect detection in homogeneous product surfaces like raw textiles.

Figure 3, shows the image pyramid for a texture extracted from the TILDA database [12]. Table 3 gives the standard deviation $\sigma_i$ of the image windows at three successive image division levels L0, L1, and L2. At level L1 it is observed that the largest gray level variation occurs in the fourth quarter of the original
Fig. 1. Homogeneous textile sample.

Fig. 2. Block size versus coefficient of variation.

image $l_4$ at level L1. Further division at level 2 shows that the least homogeneous image section exists in the second quarter $l_2$ at level L2.

5. Log-gabor filter based feature extraction

Log-Gabor based feature extraction represents the core of the defect detection and localization in the proposed AVI System as shown in Fig. 4.

Field [58] presented the Log-Gabor function as an alternative to the Gabor function because log Gabor functions have extended tails that renders them more efficient in encoding natural images than the ordinary Gabor functions [57,58]. Log-Gabor filters shown in Fig. 5 capture the important high frequency components that characterize the defects in flat surfaces. Gabor functions emphasize the low frequency components and deemphasize the high frequency components in any texture to be encoded. Features extracted from blocks convolved with Gabor filters have been used to detect responses to global textural variation. The ability of Gabor filters to approximate certain characteristics of
how information is processed in the primary visual cortex [59] in addition, their optimal localization properties in both spatial and frequency domains render them suitable for the characterization of flat surface textures. Although Gabor filters have been used widely in defect detection, they are very far from being implemented in real time. In this paper, a new method for speeding up the detection process using the Log-Gabor filters is presented.

To extract features that characterize a texture, the surface of the product is scanned with a sliding window. The gray level image of each sliding window is convolved with 24 Log-Gabor filters (4 scales and 6 orientations). Standard deviation is then computed for each filter response resulting in a 24-dimensional feature vector. The Log-Gabor Filter Bank is given in Field [58] by:

\[
G(r_0, \theta_0) = G(r_0).G(\theta_0) = \exp\left(-\frac{(\log(r_0))^2}{2(\log(\sigma_R))^2}\right).\exp\left(-\frac{(\theta - \theta_0)^2}{2\sigma_\theta^2}\right)
\]

(3)

Where \(\theta_0\) is the orientation of the filter, \(r_0\) is the central radial frequency, \(\sigma_\theta\) and \(\sigma_R\) are the angular and radial spreads of the Gaussian function respectively. Convolution of image window \(w(x,y)\) with the elementary Gabor function \(G(r_0, \theta_0)\) results in filtered window \(F_{\rho,\theta}\) at specific scale \(\rho\) and direction \(\theta\):

\[
F_{\rho,\theta}(x,y) = w(x,y) * G(r_0, \theta_0)
\]

(4)

A filter bank is constructed by dividing the spatial frequency domain into six different orientations for each of four different image resolutions resulting in a filter bank \(\{F_i\}_{i=1}^{24}\). Each texture window \(w\) will be convolved with the filter bank resulting in 24 filtered windows in Log-Gabor Domain. The texture of the flat surface is characterized by extracting the standard deviation \(\sigma\) of the Gabor-filtered window. Each window \(w_i\) is characterized by a feature vector \(f\) extracted from filters \(\{F_{\rho,\theta}\}_{i=1}^{24}\) responses, where

\[
f_{w_i} = \{\sigma_1, ..., \sigma_{24}\}
\]

(5)

and

\[
\sigma = \sqrt{\frac{\sum_{i=1}^{M} \sum_{j=1}^{N} (F_{\rho,\theta}(i,j) - \bar{F})^2}{N-1}}
\]

(6)

For defect detection, the Euclidean distance \(d\) between the standard deviations of 24 filtered windows and the standard deviation of normal filtered windows \(\sigma_{r_{ij}}\) calculated in a preprocessing phase is compared to a threshold level \(\delta\) to identify the defective image windows as shown in Fig. 4.

\[
d = \sqrt{\sum_{i=1}^{4} \sum_{j=1}^{6} (\sigma_{ij} - \sigma_{r_{ij}})^2} \geq \delta
\]

(7)

where \(\sigma_{r_{ij}}\) is the standard deviation for the Log-Gabor filtered defect free surface at scale \(i\) and orientation \(j\).
6. Experiments and discussion

To evaluate the proposed method, a large set of images from the TILDA database [12] for regularly textured fabrics and other images has been used in our experiments. The images were acquired in gray level with a resolution of 768 × 512 pixels.

Using the coefficient of variation for preliminary approximate localization of defects results in speeding up the process by 9.8 times on average. The gain in speed varies with the extent of the defect with respect to the whole acquired image window. Sample defect detection results are shown in Table 4. The white blocks indicate the segments of the original image for which the distance between the standard deviations of the responses of the 24 log-Gabor filters exceeds the standard deviations of the normal surface by a threshold value specified by a training process. A sample distribution of the standard deviations of Log-Gabor filter Responses is shown in Fig. 6. Table 6 shows the detection results of all samples used in this paper.

6.1. System performance evaluation:

Judging the performance of the proposed defect detection system is a major factor that requires careful consideration. The percentage of correct detection (PCD) generally has been used as a measure of performance in most defect detection systems. However there exists a variety of measures for judging the performance of classifiers. In our work we have considered the following four performance measures as discussed in detail in [53,56]:
Table 4
Sample defect detection results

<table>
<thead>
<tr>
<th>Sample</th>
<th>Detected Defect</th>
<th>Distribution of Log-Gabor Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Sample Image" /></td>
<td><img src="image2.png" alt="Detected Defect Image" /></td>
<td><img src="image3.png" alt="Log-Gabor Image" /></td>
</tr>
<tr>
<td><img src="image4.png" alt="Sample Image" /></td>
<td><img src="image5.png" alt="Detected Defect Image" /></td>
<td><img src="image6.png" alt="Log-Gabor Image" /></td>
</tr>
</tbody>
</table>

Fig. 5. Log-gabor filter bank for four scales and six orientations.

\[
\text{Precision} = \frac{\text{TA}}{\text{TA} + \text{FA}} \quad (8)
\]

\[
\text{Recall (Sensitivity)} = \frac{\text{TA}}{\text{TA} + \text{FN}} \quad (9)
\]

\[
\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FA}} \quad (10)
\]

\[
\text{Accuracy} = \frac{(\text{TA} + \text{TN})}{(\text{TA} + \text{TN} + \text{FA} + \text{FN})} \quad (11)
\]

All of the above quantities normally are expressed as percentages. The various terms appearing in the above equations are: True Abnormal (TA), False Abnormal (FA), False Normal (FN) and True Normal (TN). These terms can be obtained easily from the confusion matrix related to a defect detection or classification task. The meanings associated with the above measures are given below in the context of defect detection tasks:

1. Precision: indicates the percentage of correct normal classifications.
2. Recall (Sensitivity): indicates the percentage of samples that were classified as abnormal and labeled as abnormal.
3. Specificity: indicates the percentage of samples that were classified as normal and labeled as normal.

4. Accuracy: indicates the PCD Marina Sokolova, et. al [60], have shown that “the accuracy measure, does not distinguish between the numbers of correct labels of different classes”. Sensitivity and specificity separately estimate a classifier’s performance on different classes. It has been shown [60], that higher accuracy does not guarantee overall better performance of an algorithm and that a combination of measures gives a balanced evaluation of the algorithm’s performance. In this paper, we have used the Youden’s Index, Likelihoods, and Discriminant Power (DP) given in [60] to evaluate the performance of our system:

Youden’s Index $\gamma =$

$\text{sensitivity} - (1 - \text{specificity})$ (12)

Positive Likelihood: $\rho_+ =$

$(1 - \text{sensitivity})/\text{specificity}$ (13)

and Negative Likelihood $\rho_- =$

$\text{sensitivity}/(1 - \text{specificity})$ (14)

Discriminant Power

$$DP = \sqrt{3/\pi} \left[ \log(\text{sensitivity})(1-\text{sensitivity}) + \log(\text{specificity}/(1 - \text{specificity})) \right]$$ (15)

Youden’s index evaluates the classifier’s performance to a finer degree with respect to both classes. A higher positive value of $\rho_+$ means a better performance on the positive (abnormal) class. A higher negative value of $\rho_-$ means a better performance on the negative (normal) class. The DP evaluates how well a classifier discriminates between normal and abnormal surfaces. The classifier performance is good for PD > 3 [60]. The performance of the proposed system has been tested on 15 samples taken from the TILDA textile database. The test results given in Table 5 and Fig. 7, show the excellent system performance and about ten times increase in processing speed.

7. Conclusions

This paper has presented a new system for fast defect detection in flat surface products using a combination of both the coefficient of variation and the Log-Gabor filters. Using the coefficient of variation as a homogeneity measure, defects are approximately lo-
Table 5

System performance measures

<table>
<thead>
<tr>
<th>Sample</th>
<th>TN</th>
<th>FN</th>
<th>TA</th>
<th>FA</th>
<th>Precision</th>
<th>Recall/ sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Time in seconds for log-gabor</th>
<th>Time in seconds fast system (Log-gabor + CV)</th>
<th>Youden's index</th>
<th>Negative likelihoods $\rho_{-}$</th>
<th>Discriminant power (DP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>177</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>123.207074</td>
<td>4.82709</td>
<td>1</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S2</td>
<td>163</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124.333972</td>
<td>7.3064</td>
<td>1</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S3</td>
<td>156</td>
<td>0</td>
<td>38</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>123.822914</td>
<td>9.67656</td>
<td>1</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S4</td>
<td>161</td>
<td>0</td>
<td>35</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>126.035330</td>
<td>7.89534</td>
<td>1</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S5</td>
<td>154</td>
<td>0</td>
<td>41</td>
<td>1</td>
<td>0.976</td>
<td>1</td>
<td>0.994</td>
<td>0.995</td>
<td>123.756517</td>
<td>8.99356</td>
<td>0.9762</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S6</td>
<td>169</td>
<td>0</td>
<td>25</td>
<td>2</td>
<td>0.926</td>
<td>1</td>
<td>0.988</td>
<td>0.989</td>
<td>123.562316</td>
<td>41.7134</td>
<td>0.9259</td>
<td>0.1</td>
<td>∞</td>
</tr>
<tr>
<td>S7</td>
<td>185</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124.896499</td>
<td>4.4976</td>
<td>1</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S8</td>
<td>188</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124.836818</td>
<td>2.52494</td>
<td>1</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S9</td>
<td>128</td>
<td>0</td>
<td>68</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124.712361</td>
<td>39.3181</td>
<td>1</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S10</td>
<td>175</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124.743997</td>
<td>12.4425</td>
<td>1</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S11</td>
<td>163</td>
<td>0</td>
<td>32</td>
<td>1</td>
<td>0.97</td>
<td>1</td>
<td>0.994</td>
<td>0.995</td>
<td>124.756626</td>
<td>19.5029</td>
<td>1</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S12</td>
<td>155</td>
<td>0</td>
<td>41</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124.494954</td>
<td>15.5128</td>
<td>0.9697</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S13</td>
<td>190</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>125.092220</td>
<td>5.50137</td>
<td>1</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>S14</td>
<td>170</td>
<td>0</td>
<td>19</td>
<td>7</td>
<td>0.73</td>
<td>1</td>
<td>0.961</td>
<td>0.964</td>
<td>124.313181</td>
<td>8.43892</td>
<td>0.7308</td>
<td>0.3</td>
<td>∞</td>
</tr>
<tr>
<td>S15</td>
<td>182</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124.050468</td>
<td>3.076464</td>
<td>1</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>Average</td>
<td>0.9735</td>
<td>1</td>
<td>0.9958</td>
<td>0.9962</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124.4410</td>
<td>12.7485</td>
<td>0.9735</td>
<td>0.0267</td>
<td>∞</td>
</tr>
</tbody>
</table>
### Table 6
Test Results

<table>
<thead>
<tr>
<th>Sample</th>
<th>Detection</th>
<th>LG Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Sample Image" /></td>
<td><img src="image2.png" alt="Detection Image" /></td>
<td><img src="image3.png" alt="LG Features Image" /></td>
</tr>
<tr>
<td><img src="image4.png" alt="Sample Image" /></td>
<td><img src="image5.png" alt="Detection Image" /></td>
<td><img src="image6.png" alt="LG Features Image" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Sample Image" /></td>
<td><img src="image8.png" alt="Detection Image" /></td>
<td><img src="image9.png" alt="LG Features Image" /></td>
</tr>
<tr>
<td><img src="image10.png" alt="Sample Image" /></td>
<td><img src="image11.png" alt="Detection Image" /></td>
<td><img src="image12.png" alt="LG Features Image" /></td>
</tr>
<tr>
<td><img src="image13.png" alt="Sample Image" /></td>
<td><img src="image14.png" alt="Detection Image" /></td>
<td><img src="image15.png" alt="LG Features Image" /></td>
</tr>
<tr>
<td><img src="image16.png" alt="Sample Image" /></td>
<td><img src="image17.png" alt="Detection Image" /></td>
<td><img src="image18.png" alt="LG Features Image" /></td>
</tr>
</tbody>
</table>
Table 6, continued

<table>
<thead>
<tr>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
<td><img src="image3.png" alt="Image 3" /></td>
</tr>
<tr>
<td><img src="image4.png" alt="Image 1" /></td>
<td><img src="image5.png" alt="Image 2" /></td>
<td><img src="image6.png" alt="Image 3" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Image 1" /></td>
<td><img src="image8.png" alt="Image 2" /></td>
<td><img src="image9.png" alt="Image 3" /></td>
</tr>
<tr>
<td><img src="image10.png" alt="Image 1" /></td>
<td><img src="image11.png" alt="Image 2" /></td>
<td><img src="image12.png" alt="Image 3" /></td>
</tr>
<tr>
<td><img src="image13.png" alt="Image 1" /></td>
<td><img src="image14.png" alt="Image 2" /></td>
<td><img src="image15.png" alt="Image 3" /></td>
</tr>
</tbody>
</table>

Table 6, continued

<table>
<thead>
<tr>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image 1" /></td>
<td><img src="image2" alt="Image 2" /></td>
<td><img src="image3" alt="Image 3" /></td>
</tr>
<tr>
<td><img src="image4" alt="Image 1" /></td>
<td><img src="image5" alt="Image 2" /></td>
<td><img src="image6" alt="Image 3" /></td>
</tr>
<tr>
<td><img src="image7" alt="Image 1" /></td>
<td><img src="image8" alt="Image 2" /></td>
<td><img src="image9" alt="Image 3" /></td>
</tr>
<tr>
<td><img src="image10" alt="Image 1" /></td>
<td><img src="image11" alt="Image 2" /></td>
<td><img src="image12" alt="Image 3" /></td>
</tr>
</tbody>
</table>
calized in a preprocessing stage. This helps concentrate the detailed and accurate localization of defects using the computationally demanding Log-Gabor filter banks. To the best of our knowledge, Log-Gabor filter banks are being used for defect detection for the first time. Compared to the Log-Gabor filters, the proposed combination resulted in speeding up the defect detection process by about ten times. A new algorithm has also been proposed for adaptive determination of the scanning window size. Adaptive sizing of the window helps characterize the surface in a more accurate way and reduces the computation overhead. A defect detection accuracy of 99.62% has been achieved with 97.35% precision, 100% recall/sensitivity and 99.58% specificity. The experimental results show that the proposed system gives promising results and is applicable for defect detection in homogeneous surfaces like paper, steel plates, ceramic tiles, and foils.

Acknowledgements

The authors would like to thank the Workgroup on Texture Analysis of the DFG for providing the TILDA Textile Texture Database. We also would like to thank our respective institutions for providing us with the opportunity and facilities to undertake this research. Thanks are also due to Peter Kovesi for making his Log Gabor wavelet filters available [61].

References


Abstract. In this paper we present CASEP2: a hybrid neuro-symbolic system combining case-based reasoning (CBR) and artificial neural networks that aims at clustering and classifying users’ behavior in an e-commerce site. A user behavior is represented by a sequence of visited web pages, in a session. Each registered behavior is associated to one of the following classes: buyer or non-buyer. Our goal is to provide a system that mines the web site access log in order to predict the class of an on-going user navigation. One major challenge to face is to provide scalable algorithms that can handle efficiently the large amount of data to learn from. Predictions should be made in real-time, during the current navigation. In addition, raw data has a sequential nature and are very noisy. In the proposed system, two original neural networks, named M-SOM-ART networks, are applied: one to implement the retrieval phase of a CBR cycle, and the second to implement the reuse phase. This hybrid scheme allows to ensure incremental learning as well as efficient treatment of large-scale sequential data. Experiments on real log data of an e-commerce site show the relevancy of the proposed approach.

Keywords: Sequence processing, case based reasoning, self-organizing map, hybrid neuro-CBR systems

1. Introduction

Automatic adaptation and personalization of web sites has attracted lot of attention in the last few years. Applications include: recommender systems, web navigation assistants, web searching and information retrieval, intelligent tutoring and e-learning systems. One major trend in this area consists on mining web access log in order to infer automatic adaptations or personalization rules for current site visitors [43]. Users profiles can be learned from access log applying different techniques. Learned profiles can then be used to recognize the actual profile of an on-going navigation in the site. This would be the first step towards site personalization. In this work, we are interested in mining users access log in an e-commerce site, in order to learn discriminate between behaviors of buyers and those of non-buyers. Customers’ behavior is simply represented by the sequence of web pages she/he has visited. Past navigations are organized into navigation sessions. Each session is associated with the label buyer (resp. non-buyer) if the customer dose (resp. dose not) purchase an item during the session. No customer identification is needed. Only navigation sessions need to be identified. This is achieved by applying classical time-based heuristics [8].

Taking into account the dynamic nature of web sites, and the dynamic changes of customers behaviors, we propose to apply the case-based reasoning methodology to learn a classification model of customers behaviors. Case-Based Reasoning (CBR hereafter) is a problem solving methodology that is based on reusing adapted solutions applied to solved past similar prob-
Main questions considered in this work are:

- **How to implement the CBR cycle in the context of real-time-constrained data intensive applications** [1]? Actually, the size of log files, used to feed the case base, can be quite big and it increases every day. Accurate label prediction of on-going navigation session should be made before the end of the session. This implies to develop both very efficient case indexing schemes (for enhancing case retrieval), and efficient case-base maintenance strategies that enable to control the growth of the case base [45,56,62,64]. In addition the reuse phase, where new solution for a target case is devised from retrieved past similar cases, should also be implemented in a very efficient way.

- **How to efficiently handle cases that have a sequential structure**? Again, this issue raises a number of interesting problems related to:

  * **Case representation**: Main questions to answer are: What is an adequate sequence representation structure? point-based [33], interval-based [30]? How to edit a source case out of a raw data sequence? (i.e. navigation sessions). Sequences can be very long with varying length. All events forming a sequence might not be of equal importance. Different options can be considered: A source case can represent a whole sequence [57] or a subsequence [33]. In the latter case, templates for extracting relevant subsequences should be defined [47,49].

  * **Case retrieval**: Similarity measurement depends on case representation. When cases have a same length, standard similarity function measurement can be applied. For sequences of different lengths, special similarity measurement must be considered [65].

  * **Case reuse**: the temporal (sequential) aspect of the cases must be taken into account during the problem solving cycle. Successive computing done in sequence processing can be reused in the processing of the current and future states of the same sequence.

In an earlier work [70,72], we have proposed and implemented a CBR system, called CASEP, for sequence classification. Main shortcomings of CASEP are the following:

- Cases in CASEP have a fixed length and represent a part of the sequence history. For instance, only the last \( n \) visited pages are taken into consideration for predicting the next page to visit. \( n \) is a fixed system parameter. There is no clear methodology to fix this parameter. \( n \) is usually fixed in an experimental manner. This is clearly a serious limitation of the CASEP approach.

- The case base maintenance applied in CASEP has obtained a limited efficiency [73]. Experiments conducted on real data show that the proposed policy allows to reduce the case base size by at most 15% of its original size.

- In reuse phase, we have associated a confidence measurement for each solution provided by retrieval phase. This confidence is computed in simple manner. More convenient method can be used in this phase using a neural network.

In order to improve CASEP system, we propose a new hybrid neuro-CBR system called CASEP2 [73,74], in which case length is variable and cases are indexed using an original temporal neural network [67] based on the principle of self-organizing maps [36,40]. Description and evaluation of this new system proposition are the main subject of this paper.

The reminder of this paper is organized as follows. In Section 2 we give a brief description of related-work. The proposed approach is at the intersection of three main research fields: applying CBR to sequence and time series processing [38,42], Hybrid neural-CBR systems [59] and user modeling. Consequently, we focus in Section 2 on works related to these different issues. The CASEP2 approach is detailed in Section 3. Experimentation and evaluation study of the proposed approach are described and commented in Section 4. Finally we conclude in Section 5.

2. Related work

2.1. Hybrid Neural-CBR systems

Following the classification proposed in [22] we classify neuro-symbolic hybrid approaches according to the integration degree (or mode) of neural and symbolic components. Components integration mode represents
the way in which symbolic and neural components are configured in relation to each other and to the overall system. As illustrated on Fig. 1, we can distinguish between four different modes [23,26,50]:

- **Chain-processing**: In this mode the output of a component is used as an input of the other component.

- **Sub-processing**: Following this mode one of the two components is embedded in the second. Frequently, the main processing component is the symbolic one.

- **Meta-processing**: In this mode, one component is used to guide the problem solving procedure applied by the second component. Both neural and symbolic components can be used to implement a meta level. Examples of symbolic meta-processing are described in [24,51]. Few systems implements a neural meta-processing level. An exception is the hybrid natural language parser proposed in [35].

- **Co-processing**: Applying this integration mode, both symbolic and neural networks can interact directly with the environment to solve problems in either a competing or a collaborative way. One example is the PROBIS classification system [46] where a prototype-based network and a CBR component collaborate for classifying objects. An object is first examined by the neural component. If no clear decision can be made, then the CBR component is activated. The case resolution can implies also modifying the configuration of the neural network.

Different systems applying different neural-CBR hybrid approaches are proposed in the scientific literature. Main tasks achieved by hybrid approaches are:

- **Case indexing**: In [66] authors use a three layer back-propagation network in order to learn features to index cases in a case base. In PROBIS [46], the prototype-based network, used for object classification, is also used to index cases in a flat case base. Each prototype in the neural component indexes a section of the flat memory. Associations are learned during the supervised training phase. Cases that are not covered by any prototype are gathered in special section called atypical case section. During a problem solving phase, if no solution is provided by the neural component, for instance when more than one prototype of the neural network has been activated by the current input, only cases contained in sections associated with activated prototypes and atypical cases are searched to retrieve cases similar to the target case to solve.

- **Case selection**: case selection and retrieval is a crucial phase in the CBR cycle [54]. One important task is to learn weights to be applied for matching target case features and source cases. In [14, 52] two different neural approaches are provided to automatically learn relevant weights to apply for retrieving case. Another example is the system described in [53] where a rough self-organizing map is applied to feature selection for case retrieval.

- **Case adaptation**: Many neural network models, such as radial basis function (RBF) and back-propagation networks can be used to learn adaptation knowledge using discrepancy vectors [55]. A back-propagation based approach for case adaptation is given in [55]. A RBF network is applied in [10] to case adaptation in the context of ocean temperature prediction system.

Most of the above cited hybrid neural-CBR systems, do not process sequential data. Exceptions are systems described in [10,18]. However, in both works, the sequential aspect of the data was taken into account only by representing cases by fixed-size temporal windows. This could be restrictive in many applications.

### 2.2. Applying CBR to sequence processing

As stated in Section 1, main questions to handle when applying CBR to sequential and time series data are related to: case representation, case retrieval and case reuse. In next sections, we structure the review of the current state of the art of CBR systems applied to sequential data using these three issues.
2.2.1. Case representation

Two main sequential data coding schemes has been proposed in CBR systems. The first scheme, mostly used in existing systems consists on representing a sequence as a succession of instants (or points). The second scheme is based on coding a raw data sequences in the form of inter-related intervals. The representation based on succession of instants was used in several fields like information retrieval [29] and forest fire forecasting [57]. The interval-based representation is proposed in [30]. The representation of case with relations between temporal intervals is a new approach that is not often used. These relations use the Allen temporal logic [2]. This approach is certainly more complex than the first one, but makes the representation of more precise information in a case possible. The complexity of case structure can represent some inconvenience, especially when we have large amount of data with time constraints, as in our application field. However, we think that this representation is interesting and can be useful in fields where the point representation of a case can fail.

Another issue to fix, is case edition from raw sequences (instant-based or interval-based sequences). Different options have been proposed by different systems. We distinguish two main approaches: works where a case represents an entire sequence and works where a case is a sub-sequence. In later approach, a raw sequence can be used to edit several cases [33]. Examples of systems applying the first approach are: Rebecas, a forest fire forecasting system [57], RADIX [11], a web navigation adviser, and the recommender system PADIM [17].

One first example of a CBR system where a case is represented by a sub-sequence of raw sequence is the BROADWAY recommendation framework [32]. A variety of recommender systems has applied this CBR-based recommendation approach mainly: BROADWAY-V1 web navigation advisor [29] and BeCBKB web query refinement system [37]. The CASEP sequence classification system [70,72] as well as COBRA automatic web adaptation approach [47] use also this representation. We can also cite systems proposed in [10,18], which perform prediction task in a complex field where the data are temporal sequences. This temporal aspect of the data was taken into account by representing the cases by temporal windows with fixed length. In [48], the case is also represented by temporal windows, but the length of these windows is variable and depends on pre-defined threshold. Representing an entire sequence by a single case allows avoiding additional storage of more precise knowledge. However, in most of these systems, no lessons are learned from the different reasoning cycles done during a sequence evolution. For example, in Rebecas the only system knowledge update operation is the addition of complete sequences to the case base. Moreover, the retrieval of a whole sequence is time consuming. Case representation with sub-sequences can require more memory capacity if whole sequences and cases are kept in the memory. The advantage of this representation is that, for each reasoning step, useful knowledge is learned by the system. The different reasoning steps are taken into consideration during the evolution of the sequence as it is the case of BROADWAY-V1 and CASEP. This can improve the system results by reducing source cases retrieval time.

2.2.2. Case retrieval

Different similarity metrics are proposed depending on the adopted case structure. In [30], where a case is represented by a temporal network. Three similarity values are computed: the activation strength based on direct index matching, the explanation strength based on similarity explained in the general domain model (where each relation has a certain explanatory strength), while the matching strength combines both former similarities into a resulting similarity degree. An additional temporal similarity measurement: the temporal path strength is also referred. For each such path the matching degree of corresponding findings is calculated. Comparing the temporal paths is not trivial because there are often many possible temporal paths for a combination of two primitive relations. The search space is reduced by goal-directed search, guided by what to predict. In Rebecas system, an alignment algorithm is used in order to compare sequences. In RADIX system, a similarity measurement combines representativeness and dispersion measurements. It takes into account the different lengths of sequences for similarity computation. In PADIM system, the search space of cases is first divided by the general context of the system and by the focal supervision object. Then a conceptual similarity is used to find similar initial supervision environment and finally an event driven similarity allows finding the closest specific context for the current episode. The conceptual similarity has to answer the question: To which extent the current supervision objects are of the same kind of supervision objects as the reminded ones? Supervision objects are developed according to their relationships and a kind of subsumption process finds the level of matching between the current sub-
Several efforts have been devoted in order to introduce temporal information processing in Kohonen map (SOM networks) [12,13,16,31,34,36,63]. The goal is to allow taking into account temporal and spatial data types. A spatio-temporal sequence (or simply, temporal sequence) is a finite set of time ordered n-dimensional feature vectors. In [34] a phoneme recognition system is described for which two temporal variants of SOM maps are used to represent sequential features of input data. In the first model, called SOM with exponentially weighted decay, the input pattern of SOM is computed in function of former inputs. The second model uses an external delay line model: the last $T$ items of the input sequence are concatenated and presented to the network. This procedure requires high computational efforts, increasing the training time. However, it produces very good results in recognition tasks. Kohonen [39] has proposed the Hypermap architecture, where different time windows centered on a particular time step $t$ are used to construct two types of network input vectors: the context vector and the pattern vector. The context vector is first used to select a subset of nodes in the network. The best-matching neuron is then selected on the basis of the pattern vector from this subset. This approach is used in phoneme recognition, biological sequence processing and speech recognition.

Chappell and Taylor [12] have proposed Temporal Kohonen Map (TKM) for sequence classification. TKM maintains the activation history of each neuron by means of a variable called the leaky integrator potential. This approach succeeded in classifying a word in the same position within a set of sentences having different contexts. Euliano and Principle [16] have proposed the SOTPAR model based on the biologically inspired diffusion of activation through time over the neurons in the map and temporal decay of activation. These couplings are based on the propagation of activation waves starting at each winning neuron and decaying as time goes by.

### 2.4. Applying neural networks to user behavior modeling

In the field of user behavior modeling, neural networks have been mainly used for classification and recommendation by allowing to group users sharing similar profiles. One example is the system described in [21], where a neural network is applied to classify users navigation paths. In [27] a self-organizing map neural network was used to identify groups of bank customers based on repayment behavior and recency, frequency, monetary behavioral scoring predictors. It also classified bank customers into three major profitable groups of customers. The resulting groups of customers were then profiled by customer’s feature attributes determined using an Apriori association rule inducer [4].

Self Organizing Maps (SOM) has also been extensively used for recommendation computation. In [20] authors apply a SOM network for clustering documents based on a subjectively predefined set of clusters in a specific domain. In [58] SOM is applied to create a movie recommendation system. Another SOM-based recommender systems is described in [61].

In [60], a neural network is applied in order to predict the next step for a user trajectory in a virtual environment. Another applications fields where neural-
computed recommendations are studied and experimented are: student behavior modeling in intelligent tutoring systems [5] and personalized TV show recommender [7].

Next, we propose a hybrid neuro-CBR system CASEP2 [73,74] where temporal aspect of data is taken into account by using a new case modeling as well as an adequate neural network. A maintenance strategy is also proposed in order to process the large amount of processed raw noisy data (i.e. web access log data).

3. CASEP2 System Description

For sake of clarity, we start first by introducing some notations used later in the paper. Let \( q \) be an ordered set of states. In our target application, a sequence represents a user navigation in an e-commerce site. Let \( E_q = (v_i)_{1 \leq i \leq n} \) be a state in a sequence \( q \). A state is an \( n \)-dimensional feature vector given by a vector of values \( v_i, 1 \leq i \leq n \) of \( n \) characteristics \( c_i, 1 \leq i \leq n \) and by its position \( j \) in the sequence \( q \) (a state). In our application, a state represents a visited web page. Our target problem consists in providing the values \( s_k \) of a property \( S \) of states succeeding the current state of a sequence. The target property \( S \) can represent, for example, a predicted characteristic of following states of the sequence or a classification of a sequence. In our application, it represents the classification of the web site visitor in one of the two classes \{buyer, non-buyer\}.

In CASEP2 (see Fig. 2), a M-SOM-ART neural network (denoted ANN1) [19,71] indexes cases in the case base and may provide solutions for some target cases without using the CBR component. Another M-SOM-ART neural network (denoted ANN2) performs classifications in the reuse phase of the CBR component. In much a similar way to the PHOBIS system [47], the case base is divided into several sections. Each is indexed by a neuron from ANN1 except for one section: the atypical cases section. The later section contains cases added to the case base during the use mode of the system. The system has two main functioning modes:

- **Off-line construction mode:** In this mode the ANN1 is trained. It builds prototypes and an index of the case base. A reduction of the case base is also performed in this mode. All tasks in this mode are triggered periodically.

- **On-line use mode:** when a target case is presented to the system, a neuron is activated in the ANN1. If the confidence associated to the solution is greater than a given threshold \( \beta \), this solution is returned by the system and the CBR component is not used. Else if the activated neuron is indexing a section of the case base, cases in this section will be searched for retrieving cases similar to the target case, otherwise the search is done in the atypical section of the case base. Cases are added to the atypical part of the case base during the use mode of the system. These cases are used in the ANN1 training during the construction phase.
3.1. CASEP2 components description

3.1.1. Artificial neural network description

The neural network used in CASEP2 is the M-SOM-ART [19,71] network which has the following properties:

- It performs classifications and clustering tasks;
- It processes temporal sequences;
- It has the plasticity and the stability properties. The stability concerns the preservation of previously learned knowledge and the plasticity concerns the adaptation to any change in the inputs.

This network is a temporal growing neural network which integrates a self-organizing map (SOM) [40] in an Adaptive Resonance Theory (ART) paradigm [9]. This paradigm incorporates in SOM stability and plasticity properties. The ART paradigm controls the neural network evolution by introducing a vigilance test, which verifies if the activated neuron is rather close to the input. The temporal aspect of the data is taken into account by modeling sequences using dynamic covariance matrices. Clusters of cases are formed and indexed by prototypes.

The input sequence \( X = (x_i \in \mathbb{R}^n) \ (1 \leq i \leq p) \) is modelled using its associated dynamic covariance matrix \( \text{COV}_X \in \mathbb{R}^n \times \mathbb{R}^n \) defined as follows [19,67, 68]:

\[
\text{COV}_X = \frac{1}{p} \left[ x_1 x_T + \sum_{i=2}^{p} (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T \right]
\]

where \( \bar{x}_t = \frac{1}{t} \sum_{i=1}^{t} x_i \ (t \geq 2) \) is the dynamic mean vector associated to \( x_t \in \mathbb{R}^n \) in the sequence and computed using the precedent and the current vectors \( \{x_i\}, \ (1 \leq i \leq t), \) and \( x^T \) represents the transposed vector of \( x \). This model allows representing the position (because the mean vector is introduced in the computation of the covariance matrix) and the shape of the cloud of points representing the sequence. A dynamic mean vector is introduced in the covariance matrix computation in order to take into account the order of the vectors in the sequence. All sequences models have the same dimension. The distance between a covariance matrix \( \text{COV}_X = (x_{ij}), \ 1 \leq i, j \leq n \) and neuron weights \( W_c = (w_{ij}^c), \ 1 \leq i, j \leq n \) is the Frobenius matricial distance \( d(f) \) given by:

\[
d(f(\text{COV}_X, W_c)) = [\text{tr}(\text{COV}_X - W_c)^T (\text{COV}_X - W_c)]^{1/2}
\]

where \( \text{tr}(M) \) is the trace of the matrix \( M \) and \( X^T \) is the transposed of the vector \( X \).

The M-SOM-ART is first initialized with one neuron. New neurons are added to the map following the simplified ART paradigm using the vigilance test to select the winner. If the vigilance constraint is satisfied, the map is updated; else a new neuron is added to the map. The closest neuron to the input in the perimeter of the map is determined and a new neuron is added in its neighborhood. The corresponding connections are also added to the network. At the last time of the learning process, the neurons are labelled using the labels of the inputs that have activated them. The M-SOM-ART learning algorithm is described in Algorithm 5.1.

\[\text{Algorithm 5.1 M-SOM-ART learning algorithm.}\]

1. Present an input signal \( X = (x_i) (1 \leq i \leq p), x_i \in \mathbb{R}^n \)
2. Model the input \( X \) by its associated dynamic covariance matrix:
\[
\text{COV}_X = \frac{1}{p} \left[ x_1 x_T + \sum_{i=2}^{p} (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T \right]
\]

Initialise the reference vector \( W_{c1} \) of the first unit \( c_1 \) to the dynamic covariance matrix of the first input signal.
3. Determine the winner \( s(X) \in A \) by:
\[
s(X) = \arg\min_{c \in A} \frac{f(d(\text{COV}_X, W_c))}{c}
\]

where \( f(d) \) is the Frobenius distance.
4. Subject the winner \( s(X) \) to the vigilance test:
\[
\Delta W_r = \epsilon(t) h_{rs} [\text{COV}_X - W_{cr}]
\]

\( \epsilon(t) \) is the learning rate, \( \sigma \) is the standard deviation of the Gaussian, \( \sigma \) and \( h_{rs} \) is the neighborhood function.
5. Increase the time parameter \( t = t + 1 \) if \( t = t_{\text{max}} \) label each unit using a majority vote on the labels of the inputs that have activated these units.

The choice of this network is based on its properties. It performs sequence clustering task for indexing the case base and classification task for providing target cases’ solutions. The temporal aspect of data is taken into account and the stability-plasticity properties are very important for a long time use of the system. This
model gives better results for user navigation classification compared to other existing models [19, 67].

3.1.2. Case based reasoning component

3.1.2.1. Case structure

In CASEP2, any sequence \( q(m) = (E_j^q), 1 \leq j \leq m \) is modelled by a dynamic covariance matrix \( COV_q \) given by:

\[
COV_q(m) = \frac{1}{m} \left[ E_q^1 (E_q^1)^T + \sum_{j=2}^{m} (E_j^q - \bar{q}(j)) (E_j^q - \bar{q}(j))^T \right]
\]

where \( \bar{q}(j) \) represents the dynamic mean vector of the sequence states and \( X^T \) represents the transposed of the vector \( X \). The problem part of the case is defined by the dynamic covariance matrix associated to the sequence and the solution part is the sequence class. The target case problem part represents the current sequence, which is also modeled by a dynamic covariance matrix. A sequence is formed at each presentation of a new state, the current sequence is a sub-sequence (a part of the site visitor navigation) of the whole final sequence (see Fig. 3).

The case model is given by the ANN1. It allows representing each sequence and its sub-sequences by matrices with the same dimension. This avoids sub-sequences extraction. In addition, in the current sequence \( q(m+1) \) corresponding to the target case, the new covariance matrix \( COV_q(m+1) \) can be computed using the previous one \( COV_q(m) \) associated to the same sequence \( q(m) \) at each presentation of a new state as follows:

\[
COV_q(m+1) = \frac{m}{m+1} COV_q(m) + \frac{1}{m+1} \left[ (E_{q(m+1)}^1 - \bar{q}(m+1)) (E_{q(m+1)}^1 - \bar{q}(m+1))^T \right]
\]

3.1.2.2. Memory organization

CASEP2 contains a simple indexing system that contains two levels of memory (see Fig. 4):

- Memory that contains prototypical cases (prototypes): it is used during the retrieval phase as an indexing system in order to decrease retrieval time. Each prototype is represented by one neuron, which can index a set of cases (a part of the case base). Only the activated neurons during the last step of the training are linked to the case base.
- Memory that contains real cases (the case base): it is a simple flat memory in which cases are orga-
nized into parts of similar cases. It is partitioned by the neural network. Each part, except one, is linked to a neuron. The part, which is not linked to the neural network, contains the cases added to the case base during the use mode of the system.

The use of the neural network improves the retrieval efficiency. The use of the case base can allow obtaining more precise results than those obtained using only the neural network, which retains just the representative cases (the prototypes).

3.1.2.3. Case base maintenance

Additional new measurements are associated to the cases as suggested in [25,32,56]. These measurements are used to improve the prediction quality and to reduce the case base size. They are presented below:

- **Positive contribution (CP)** represents the number of successful retrievals. It takes into account the number of cases, which provide the target case solution.
- **Negative contribution (CN)** represents the number of unsuccessful retrievals. It also takes into account the number of cases, which provide the target case solution.
- **Case quality (QC)** represents the rate of the case success in solving target cases:
  \[
  QC = \frac{CP}{CN + CP}.
  \]
- **Extent (ET)** determines a neighborhood of the case, which contains the cases covered by this case. This neighborhood can contain cases with different solutions. ET defines a variable similarity threshold for each source case \(case_i\):
  \[
  \alpha(case_i) = 1 - ET(case_i).
  \]

3.1.2.4. Measurements initialization

When a case is added to the case base, its measurements are initialized as follows: \(CN = 0, CP = \frac{1}{m}\), \(QC = 1\) and \(ET = 1 - \alpha_0\), where \(m\) is the number of selected representative cases and \(\alpha_0\) is the similarity threshold.

3.1.2.5. Measurements update

Updates are performed for the retrieved cases (Algorithm 5.2) as follows:

- If the solution is correct, then the \(CP\) of the cases which have contributed to provide the solution increases.
- If the solution is not correct, then the \(CN\) of the cases which have contributed to provide the solution increases. If the \(QC\) of these cases is smaller than a certain threshold \(d'\), their \(ET\) decreases.

### Algorithm 5.2 Case measurements update.

```plaintext
For each target case \(case_i\)
   For each case \(case_j\) retrieved from BC //BC is the case base
      If \(valSol = sol(case_j)\) then //valSol is the solution value in S
         If \(sol(case_j) = sol(case_i)\) then \(CP(case_j) = CP(case_j) + \frac{1}{nb}\)
         \(nb\) is the number of the extracted cases, which have contributed to provide the solution */
      Else
         \(CN(case_j) = CN(case_j) + \frac{1}{nb}\)
         If \(QC(case_i) < d'\) then \(ET(case_i) = ET(case_i) - pr(ET)\)
         \(lbd'\) is a system parameter and \(pr(ET)\) is a function of \(ET\)
      endIf
   endFor
endFor
```

Our maintenance strategy consists in:

- Dividing the case base into several parts using M-SOM-ART neural network in order to reduce the case search space.
- Reducing the use of noise cases, which deteriorate the system competence by decreasing their extent (ET). This leads to increase their similarity threshold.
- Reducing the case base size, off line [44]. The notion of coverage is used to construct a reduced case base. The coverage set of a case \(c\) is the set of cases which are similar to \(c\) and have the same solution as \(c\).

A case \(case_i\) is considered as noise if its quality is bad \((QC(case_i) < d'\) where \(d'\) is a threshold defined in the system).

The case base size reduction (see Algorithm 5.3) consists in constructing, from the case base, a reduced case base (initially empty) by gradually adding cases (selective addition strategy). First, these cases are ordered by their decreasing extent. If two cases have the same extent, they are ordered by their decreasing quality, and then put in the reduced case base. A case \(c\) is not added to the reduced case base (not useful case) if \(c\) is covered by an other case \(c'\) in this base and if the neighborhood of \(c\) is included in the neighborhood of \(c'\).
Algorithm 5.3 Case base reduction.

```
list ← () //list is a set of ordered cases
Order the BC cases in list according to their decreasing ET, then QC
// Construct the reduced case base RBC
RBC ← \{first_element \in list\}

For all case case_i \in list
  For all case case_j \in RBC
    If \[(1 - \text{Similarity}(case_i, case_j) + ET(case_i) < ET(case_j)) \land \text{sol}(case_i) = \text{sol}(case_j)\]
      then update the measurement associated to case_i
    Else RBC ← RBC \cup \{case_i\}
  endFor
endFor
```

3.2. CASEP2 modes

3.2.1. Construction mode

In this phase, the neural network is trained, the construction of the case base and its reduction are done.

3.2.1.1. CASEP2 initialization

Initially, the case base is empty. A training dataset is used for adding cases to the case base and for forming clusters linked to the ANN1. This training base contains a set of sequences. These sequences are modeled using dynamic covariance matrices. At the last step of the ANN1 training phase, the case base is formed. For each presentation of a sequence, one neuron is activated. Each activated neuron is associated to one part of the case base. Each case that activates a neuron is added to a part linked to this neuron. Maintenance measurements are initialized and associated to the cases added to the case base as described previously.

3.2.1.2. CASEP2 updating

After the system initialization, it is used for providing classifications in the use mode. During this mode, cases are added to the atypical part of the case base. This allows their use to provide target cases’ solutions just after their addition to the system (this makes the system incremental in real time).

When the size of this part is greater than a given threshold \(\gamma\) (\(\gamma\) is a system parameter), the system is updated in the construction mode where the ANN1 is trained using the cases contained in the atypical part. Some cases are added to the existing parts and other parts are formed and linked to the neurons.

3.2.1.3. Reduction phase

In this phase (see Algorithm 5.3), a case base reduction is done in the same way that in CASEP system [70, 72]. Not useful cases are removed from the case base. This allows controlling the contents of the case base. The removal of obsolete cases can also be considered when the system is used for a long time.

3.2.2. Use mode

For each presentation of a new state in the current sequence, a new case is formed and modeled by a covariance matrix. One neuron is then activated in the neural network. If the solution confidence associated to the activated neuron is greater than a threshold \(\beta\), the solution is returned by the system. Otherwise the CBR cycle is triggered. Next we describe main phases of the CBR cycle applied in CASEP2.

3.2.2.1. Retrieve phase

- If the activated neuron is linked to a part of the case base, the search process is restricted to this part.
- Otherwise, the search is done in the atypical part of the case base.

The similarity measurement between two cases modeled by the covariance matrices \(COV_X\) and \(COV_Y\) is inversely proportional to the Frobenius distance, it is given by:

\[
similarity(COV_X, COV_Y) = \frac{1}{1 + fd(COV_X, COV_Y)}
\]

The \(\alpha\) cases the most similar\(^1\) to the target case are retrieved. The similarity threshold is variable and depends on the extent of each case [72]. This allows taking into account the noise contained in the data. For the classification task, if one part of the case base contains only cases belonging to the same class, the search in this part can be avoided and the solution is given by the ANN1.

3.2.2.2. Reuse phase

In this phase, the ANN2 is used in order to perform classification task. This network uses the \(\alpha\) retrieved

---

\(^1\)If less than \(\alpha\) cases are retrieved, these cases are used in the reuse phase. If no case is retrieved, the solution is provided by the ANN1.
cases as training dataset. Then a class is provided for the target case. In order to obtain a neural network with reasonable size (this is necessary for satisfying the real time constraint), the ANN2 is initialized for each presentation of a new sequence (corresponding to new site visitor navigation in our application). In the same sequence, the same ANN2 is used until the end of the sequence. This allows reusing learned knowledge from previous processing done in the same sequence.

### 3.2.2.3. Learning phase

In this phase, the cases are added to the system when the solution is not correct.\(^2\) The update of the maintenance measurements associated to the retrieved cases and the initialization of these measurements for the added cases are done as described previously, in the same way that in CASEP [72]. The atypical part contains the learned cases, which can be used in the system just after their addition to the case base, this makes the system incremental in real time.

### 4. Experimental results

We have performed several experiments on log files of an e-commerce Web site, where thousands navigations are registered every day. More precisely, the behaviour of each site visitor is described by the information about the succession of pages that he/she has visited (time, user’s IP address, URL of the requested page etc.). This succession of pages represents the temporal aspect of the data. Since the log file contains noise, some pre-processing is done before the use of M-SOM-ART.

The data is first filtered in order to remove a part of noise (see Fig. 5). An e-commerce web site is a dynamic site. Site pages are not characterized by fixed variables such as: hierarchical address (URL), content, . . . , etc. Pages are represented by numerical identifiers which have no meaning. For that reason, authors of [69] have elaborated a method for coding sessions from the Log file which consists in characterizing a page by its passage importance, i.e. by its weight of precedence and succession relative to the other pages of the site which appear in the log file. The principle is to calculate for each page its frequency of precedence and succession over all the other pages and to regroup these frequencies in one matrix: Pages x (previous Pages + next Pages) that we call: quasi-behavioral matrix.

Data is then processed using a SOM network in order to obtain an evolution space: user profiles are extracted by regrouping similar pages. The navigation pages are then represented by the weights of neurons in the obtained SOM. The navigation is represented by the succession of these vectors (corresponding to the succession of pages). Notice that navigations (i.e. sequences) can have varying lengths.

The goal of our experiments is to classify site visitors in one of the two classes \{buyer, non-buyer\} using CASEP2. In our experimentations, two bases are used. The first one contains 3000 sequences, it is used during the construction mode in order to initialize the case base. The second one contains 10000 sequences, it is used in the use mode. The buyer sequences represent less than 10% of all the sequences in the two bases. The system is evaluated using the following criteria:

\(^2\)In our application the solution is not correct if the provided class is not the same as the real class.
Table 1
CASEP2 Experimented parameter values

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value type</th>
<th>Value interval</th>
<th>‘Appropriate’ values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>similarity threshold</td>
<td>real</td>
<td>[0,1]</td>
<td>0.6</td>
</tr>
<tr>
<td>$\beta$</td>
<td>solution confidence threshold associated to activated neuron</td>
<td>real</td>
<td>[0,1]</td>
<td>0.6, 0.7, 0.8</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>atypical part size threshold</td>
<td>integer</td>
<td>[1,...]</td>
<td>500</td>
</tr>
<tr>
<td>$d'$</td>
<td>case quality threshold</td>
<td>real</td>
<td>[0,1]</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Table 2
Total results comparison

<table>
<thead>
<tr>
<th>CASEP</th>
<th>M-SOM-ART</th>
<th>CASEP2-V2(0.6)</th>
<th>CASEP2-V2(0.7)</th>
<th>CASEP2-V2(0.8)</th>
<th>CASEP2-V1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification rate</td>
<td>76.62%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Recall</td>
<td>61%</td>
<td>86.49%</td>
<td>87.3%</td>
<td>86.82%</td>
<td>84.02%</td>
</tr>
<tr>
<td>Precision</td>
<td>80%</td>
<td>86.49%</td>
<td>87.3%</td>
<td>86.82%</td>
<td>84.02%</td>
</tr>
</tbody>
</table>

Table 3
Buyer class results comparison

<table>
<thead>
<tr>
<th>CASEP</th>
<th>M-SOM-ART</th>
<th>CASEP2-V2(0.6)</th>
<th>CASEP2-V2(0.7)</th>
<th>CASEP2-V2(0.8)</th>
<th>CASEP2-V1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>43.69%</td>
<td>70.73%</td>
<td>69.05%</td>
<td>70.33%</td>
<td>75.39%</td>
</tr>
<tr>
<td>Precision</td>
<td>53.73%</td>
<td>80.76%</td>
<td>84.77%</td>
<td>82.23%</td>
<td>71.57%</td>
</tr>
</tbody>
</table>

Table 4
Non-buyer class results comparison

<table>
<thead>
<tr>
<th>CASEP</th>
<th>M-SOM-ART</th>
<th>CASEP2-V2(0.6)</th>
<th>CASEP2-V2(0.7)</th>
<th>CASEP2-V2(0.8)</th>
<th>CASEP2-V1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>70.61%</td>
<td>93.02%</td>
<td>94.86%</td>
<td>93.70%</td>
<td>87.59%</td>
</tr>
<tr>
<td>Precision</td>
<td>95.38%</td>
<td>88.46%</td>
<td>88.99%</td>
<td>88.40%</td>
<td>89.58%</td>
</tr>
</tbody>
</table>

– **Recall**: represents the ratio of the number of correct classifications to the number of queries.
– **Precision**: represents the ratio of the number of correct classifications to the number of all predictions.
– **Classification rate**: represents the ratio of the number of classifications to the number of queries.

We have tested several values of the parameters for determining the ‘appropriate’ ones (trial and error). Table 1 shows the value interval and the ‘appropriate’ values of each parameter. The similarity threshold $\alpha_0$ is initialized to 0.6, this value decreases if the case contributes with success to resolve new cases and increases when the case fails to get case solutions. The atypical part size threshold associated to activated neurons is defined in order to increase the system efficiency. This value depends on the number of navigations processed by the system and the time required to get solutions. If the size of atypical part is large, the solution search in this part will be time consuming and the system efficiency is deteriorated. The case quality threshold allows to reduce the use of cases that deteriorate the competence of the system.

4.1. Results

The addition of the maintenance measurement improves system results. This is tested using the same databases in CASEP [70,72]. In order to evaluate CASEP2 system, we have compared its results in use phase to those obtained using CASEP and M-SOM-ART neural network. In the first experiments, CASEP2 (CASEP2V1 in the Tables 2, 3, 4) uses the CBR module to provide the classifications (the neural network is used to index the case base and in the retrieval phase of the CBR system). In the second ones, CASEP2 uses the M-SOM-ART network for providing target cases’ solutions without using the CBR component when the confidence associated to the provided class is greater than a threshold $\beta$. If this confidence is lower than $\beta$, the CBR component is used (the values of $\beta$ are 0.6, 0.7, and 0.8 corresponding to CASEP2V2(0.6), CASEP2V2(0.7), and CASEP2V2(0.8) in the Tables 2, 3, 4). Results are shown in the Tables 2, 3, 4.

Table 2 shows that CASEP2V2 (CASEP2V2(0.6) and CASEP2V2(0.7)) gives better global results than M-SOM-ART and CASEP because it uses the neural net-
work and the CBR components to provide the target cases’ solutions. In CASEP2, the global result improves with the neural network use frequency (when the threshold $\beta$ increases, the neural network use frequency decreases). CASEP2 and M-SOM-ART provide solutions for all target cases, this is not the case for CASEP. Their recall results are better than those of CASEP. Tables 2 and 3 show that neural network use increasing in CASEP2 improves the detection of the most frequent class (the non-buyer class) and CBR use increasing improves the rare class (buyer class) detection. Concerning system efficiency, CASEP2 processes sequences in less time than CASEP (because the case base is divided into several parts and there is no cases extraction) and the neural network M-SOM-ART is more efficient than CASEP2. In CASEP2, neural network frequency use improves system efficiency. These results show that the CBR component can recognize the rare class better than the neural network alone which recognizes the most frequent class. In these experiments, one of the CBR or the neural network is used to provide target cases’ solutions.

5. Conclusion

We have presented a hybrid neuro-CBR system for sequence classification. A new case modeling and different interactions schemes between adequate neural networks and CBR component are presented. The hybrid system improves several aspects of its two components. The neural network indexes the case base and allows reducing the search space during retrieval phase, this improves system efficiency. In addition, the neural network adapts retrieved cases’ solutions. Case based reasoning makes the system incremental in real time, this is not the case for the neural network. The chosen neural network takes into account the dynamic arrival of data, but learning phases are not done in real time. In addition, the CBR can give more precise results than the neural network since it keeps all the cases while the neural network keeps just means of these cases.

In CASEP2, the modules are tightly integrated since both CBR and ANN modules communicate using shared memory. Moreover the principal processing is ensured by the CBR component and the two M-SOM-ART neural networks are sub-processors. The first M-SOM-ART network indexes the case base while the second one provides sequence classification in the reuse phase. We can also view the first M-SOM-ART and the CBR component as co-processors because they both participate to provide the target cases’ solutions.

In CASEP2, we use the same memory structure as the one proposed in PROBIS system [46], but this one does not process the temporal data and the used neural network is different from those used in CASEP2. Moreover the addition of cases is not done in the same manner. In [10,18], authors use the neural networks in the different CBR cycle phases to process temporal data, but this temporal aspect is taken into account by defining temporal windows with fixed length to represent the cases. This can influence the quality of system results in several applications. In CASEP2, no restriction is imposed for the sequence length in the case representation. Temporal data processing is done by an adequate neural network that have, moreover, the stability and the plasticity properties which are important for a long-term use of the system. In addition, this network is used in case adaptation and takes into account the precedent processing done in the same sequence. In [10,18], the memory organization is not the same as in CASEP2 (this concerns mainly the atypical part). The ANN used for case indexing is the GCS, which is a growing SOM that does not take into account the temporal aspect of the data and that does not assure the preservation of old knowledge. In these systems an RBF neural network is used in the reuse phase but does not take into account the temporal aspect of the data. Concerning case representation, in CASEP2, cases are represented by successions of instants. They represent sub-sequences with different lengths. We propose a new modeling of sequences, which takes into account the points (states) distribution and their order in the sequence. In CASEP2, the maintenance measurements are associated with each case like in CASEP [70,72]. These measurements determine case relevance according to its use for solving target cases.

In addition to this maintenance strategy, the case base is divided into several parts. This leads to improve system efficiency. In CASEP, the reuse task involves computing a confidence associated to each class. More adequate method is used in CASEP2 since we use M-SOM-ART neural network, which classifies sequences and preserves the preceding processing done in the same sequence. In CASEP2, the CBR system provides training data sets for both neural networks; allowing the use of the learned cases just after their addition to the case base (this cannot be done if the neural network is used alone); and can provide more precise results since concrete cases are used for giving the target cases’ solutions (in the M-SOM-ART network only prototypes are stored).
We have performed some experiments on an e-commerce web site. More experiments will be done in different applications and for the prediction task. The similarity measurement used in CASEP2 is based on the Frobenius matricial distance described above. More suitable distances can be used to compare covariance matrices. They use the sphericity measure \[6\] and are especially conceived for covariance matrices comparison. In future work, we will use these distances to define new similarity measurements between cases. We will also represent the sequence similarities using kernel matrices \[28\]. The kernel is a special similarity, which processes the property to provide the computation for a scalar product in some associated Hilbert space. It can be defined between complex objects like time series, sequences, graphs, etc.

References


[28] N. Hiroyuki, S. Yasumasa and H. Akira, Learning a kernel matrix for time series data from DTW distances. Neural Info-


Rushed Kanawati, Maria Malek, and Sylvie Salotti, editors, Proceedings of the first international workshop on applying Case-based reasoning to time series prediction, Trondheim, June 2003.


