

# User authentication for keypad-based devices using keystroke analysis

T.Ord<sup>†</sup> and S.M.Furnell<sup>‡</sup>

<sup>†</sup> Spinnaker International Ltd., Spinnaker House, Budshead Road, Crownhill, Plymouth,  
United Kingdom

<sup>‡</sup> Network Research Group, School of Electronic, Communication & Electrical Engineering,  
University of Plymouth, Plymouth, United Kingdom  
email: sfurnell@plymouth.ac.uk

## Abstract

The use of a Personal Identification Number (PIN) is a common means of ensuring user authentication on numeric keypad devices. However, like passwords and other forms of authentication based upon secret knowledge, PINs have the potential weakness that they may become known to other people.

This paper describes a potential approach for strengthening PIN-based authentication, by incorporating a biometric measurement of the user's typing style when keying in their number. Such keystroke analysis techniques have previously been used in a full keyboard context, but the keypad scenario is considered to represent a more complex problem.

An experimental study is described in which a neural network approach was used to classify and discriminate between 14 test subjects. The main results, using a 6-digit PIN, yielded a False Acceptance Rate (FAR) of 9.9%, with an accompanying False Rejection Rate (FRR) of 30%. Further experiments were able to significantly reduce the error, but at the expense of a longer PIN. The paper also considers potential application areas, in view of the results observed.

## Keywords

Security, Authentication, Biometrics

## 1. Introduction

The accurate authentication of users represents an important issue in a variety of information technology systems, including computers/networks, Automated Teller Machine (ATM) systems and mobile phones. There are various techniques and technologies that can be used to achieve this, the most common being the use of passwords or Personal Identification Numbers (PINs). However, these share the weakness that they are based upon a foundation of secret knowledge. If this information is ever shared or discovered, the system becomes vulnerable to attack (Jobusch and Oldehoeft, 1989). This paper examines the use of keystroke analysis, which recognises that a person's typing pattern on a keyboard or keypad may exhibit unique characteristics. Keystroke analysis is based upon utilising these characteristics to differentiate one user from another. The pattern in keystroke analysis is formed from the different inter-keystroke latencies.

Keystroke analysis is an example of a biometric. Biometric-based authentication systems aim to verify a user's claimed identity by measuring physiological or behavioural characteristics (i.e. something that the user *is* as opposed to something that they *have* or *know*). There are

numerous other biometric techniques, including fingerprints analysis, facial recognition, retinal scanning, iris scanning, vascular patterns, voice dynamics and signature dynamics (Miller, 1994). The first five of these are based on physiological characteristics, whilst the last two, along with keystroke analysis, are based upon behavioural measures. The advantage of keystroke analysis over other biometrics is its low cost (the technique can be implemented entirely in software) and the fact that it can be transparent to the user when they type in their PIN code.

As with other biometric-based systems, the effectiveness of keystroke analysis can be judged on the basis of two types of error:

- False Acceptance Rate (FAR): The extent to which the authentication system will falsely judge an impostor to be the legitimate user. Sometimes referred to as Impostor Pass Rate.
- False Rejection Rate (FRR): The extent to which legitimate users will be incorrectly judged to be impostors and, therefore, denied access by the authentication system. Sometimes referred to as False Alarm Rate.

These errors have a mutually exclusive relationship, such that the decrease of one will generally result in the increase of the other. The level of error must be controlled in the authentication system by the use of a threshold to determine the point at which users will be accepted and rejected. Selecting an appropriate threshold is, therefore, very important: too lax a setting will result in a high level of false acceptance, whereas too strict a threshold will cause legitimate users to be falsely rejected on a frequent basis.

The idea of using keyboard characteristics for authentication is not unique, and there have been a number of previous papers published on this topic, the main results of which are summarised in table 1 below.

Authors	%FAR	% FRR
Gaines et al. (1980)	0%	4%
Legget & Williams (1988)	5%	5.5%
Joyce & Gupta (1990)	0.25%	16.67%
Bleha et al. (1990)	2.8%	8.1%

**Table 1: Summary of previous keystroke analysis studies**

In these previous experiments, the full keyboard has been utilised in examining user's typing patterns. In this study, however, a numerical keypad approach has been taken. This provides a more complex problem than the previous experiments in that only one finger is normally used to enter codes on a numerical keypad, as opposed to the two hands when typing on a full keyboard. When typing with two hands, more information about the user can be obtained, as not only is there a pattern from the typing of each individual hand, but also in the interaction of the two hands (Gentner, 1983). Gentner states that one-finger digraphs have lower variability, which makes the classification of users more difficult than with two finger or two hand digraphs (because the 'signatures' are closer together).

The paper presents an experimental study of keystroke analysis in a numeric keypad context, in order to obtain a practical measure of its effectiveness. It is considered that the successful

implementation of such an approach would have value in contexts where traditional secret knowledge PINs are currently the only form of protection.

## **2. Methods and Procedure**

In order to examine the inter-keystroke times of users entering numerical codes, a data acquisition system had to be designed. The inter-keystroke times were required to be measured in milliseconds and stored together with the code of the keys that were pressed. The keypad on a standard PC AT-101 keyboard was used for this purpose, with appropriate modifications to the PC timer and key action interrupt routines to enable the required information to be collected at the appropriate resolution. In order to make the PC keyboard layout correspond more closely to that found on ATM machines and telephony devices, the arrangement of the numeric keys was reversed (i.e. so that the keys 1,2,3 appear at the top of the keypad rather than the bottom).

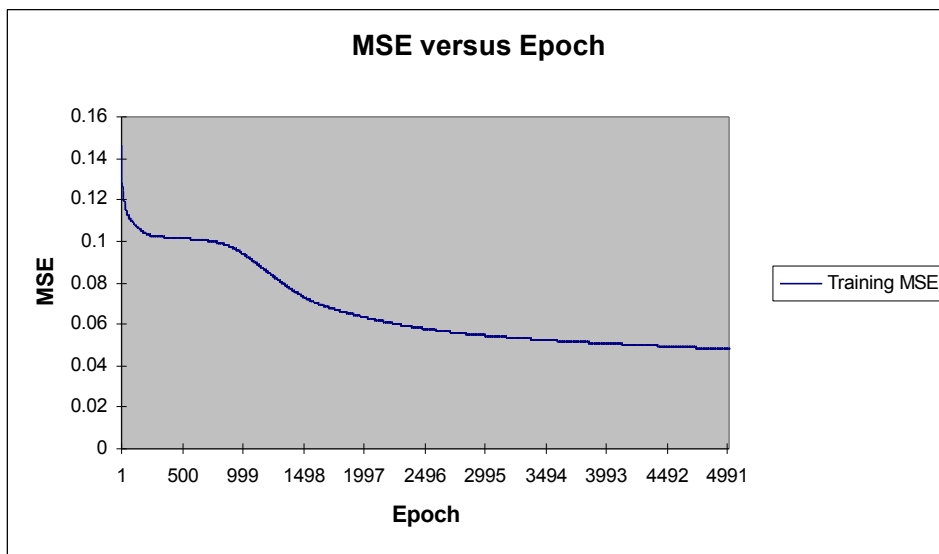
The experiments were conducted within the Research Department of Spinnaker International, Plymouth. The data acquisition system was operating for 6 months, from 0900 to 1730 each day. Co-operation was encouraged by making the system serve as a fire register, which the test subjects were required to use whenever entering and leaving the building. The participants were all experienced with using PC's and, therefore, had basic typing skills, but none were touch typists. In fact, as was expected, all the participants only used one finger to type in the numerical password. All users were asked to type in the same numerical code, 288970, which was selected as it was a number that they were already familiar with. The ASCII code of every key pressed and the corresponding inter-keystroke latency were stored to a text file as the users typed in the code.

A total of 50 samples were collected for each user, which were then used as inputs to a neural network to assess the effectiveness of the authentication technique. The first 30 samples were used as reference data for training the neural network, while the last 20 were used for testing purposes. As the data collection system was running for a total of 6 months and from 0900 to 1730, any effects from the uncorrelated sources of noise from the state of the user or from the equipment itself, is expected to be averaged out. These sources of noise could include minor illnesses, the time of day the entry is made, stress, and tiredness. The analysis stage of the experiments took place in non-real time. Any latency less than 40 milliseconds and greater than 1 second were not used. This was because a latency of less than 40ms could arise if a user hit two keys together, and it was assumed that latencies over 1 second arose from the user being distracted from an external source (and, hence, not part of their natural rhythm).

The neural network was constructed on a simulation package called NeuroSolutions. The neural network used was a Multi-Layer Perceptron (MLP) with the Back-Propagation Learning Rule (Bishop, 1995). The training of the network can be split into three sub-stages: the feed-forward of the reference samples through the network, the back-propagation of the error, and then the weight update. The MLP, when used with the Back-propagation learning rule, is an example of supervised learning (Looney, 1997). Each feature vector (in this case represented by a typing sample) is fed into the system, along with its known class identifier as the desired output vector (0 or 1 in this case), and the network learns to map the input feature vector into the desired class identifier. For each user, the network is trained on recognising that user's 30 reference samples, whilst at the same time recognising that the other 13 users' samples (390 in total) are not from the same user. To facilitate this the desired output for the

target user is set to '1', whilst the desired output for the other 13 impostors is set at '0'. This process is repeated for each user acting as a target and the other 13 users as impostors.

One complete training iteration is referred to as an epoch (i.e. when all training samples have been presented to the network once). Batch learning was applied in this experiment, where after each epoch all the weights for each sample were stored, and the weights updated with the average weight update. The appropriate number of epochs for training the network for each user was found to be 5000 by Cross Validation. On average there was no advantage in increasing the number of epochs. Cross Validation shows if the network is being over-trained, which results in the network being unable to recognise the general case of that pattern, only the specific patterns it was trained with.



**Figure 1: Example of the Neural Network learning curve**

Figure 1 illustrates how the Mean Square Error at the output of the Neural Network is reduced as the number of iterations of presenting the set of training patterns into the Multi-Layer Perceptron Network increases. This is essentially the learning curve of the neural network. As the curve does not reach 0, it has not learned the tasks of separating each user class exactly. As such, the error of this classifier is greater than 0. This was expected, as there are certain limitations to this classification, which are caused by the data itself. The features may be inadequate to distinguish the different user classes no matter how well a discriminant function can separate the classes.

The network was trained for each user, resulting in each user having their own sets of weights. The network topology, epochs, and all other variables were the same for each user. This method was chosen so as to minimise the potential hardware and software that would be involved in a practical implementation. One neural network could be used, with a particular users' weights loaded when identified, which could then be used to authenticate that user.

For these experiments, the maximum acceptable FRR was set at 30% as it is considered to be the highest level that could be tolerated by users. Currently, PIN-based authentication systems, such as ATM's and mobile phones, typically permit the user three attempts to enter the code correctly. As such, a window of opportunity would still exist for them to recover from a false rejection. The probability of a valid user being denied access after three attempts

is therefore 2.7%. Given the mutually exclusive relationship between the FAR and FRR that was described earlier, the advantage of allowing a relatively high FRR is that it enables a corresponding reduction of the FAR – which is considered to be the more important measure from the security perspective. The goal of these experiments was, therefore, to determine the FAR that could be achieved with a FRR of 30%.

### 3. Results

This section of the report details the actual results obtained from the neural network experiment. Following this, additional experiments were undertaken to further analyse and improve the results.

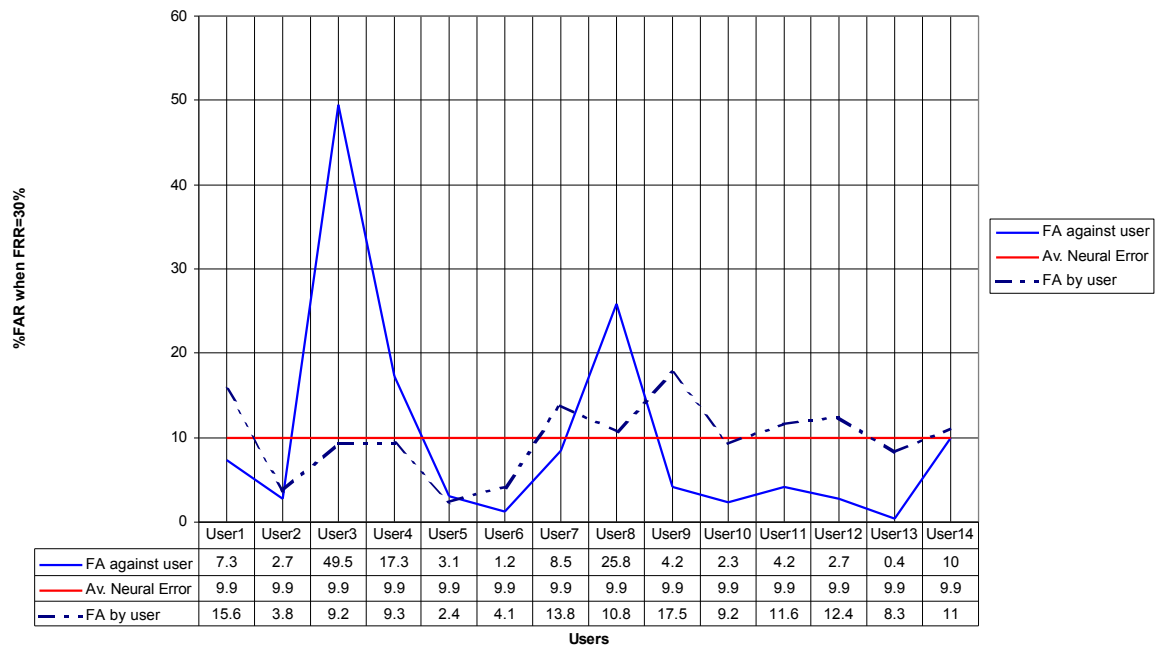
User	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Average
1		0	83	15	0	0	0	85	15	0	0	0	5	0	15.6
2	0		5	10	0	0	0	5	0	30	0	0	0	0	3.8
3	20	5		55	0	5	5	15	0	0	0	0	0	15	9.2
4	10	5	36		0	5	5	15	0	0	10	5	0	30	9.3
5	0	5	21	0		5	0	0	0	0	0	0	0	0	2.4
6	0	0	13	0	20		0	20	0	0	0	0	0	0	4.1
7	0	0	90	25	10	0		0	0	0	0	0	0	55	13.8
8	25	0	55	35	0	0	5		5	0	0	15	0	0	10.8
9	20	0	88	0	0	0	5	75		0	40	0	0	0	17.5
10	20	20	10	45	0	0	0	20	0		5	0	0	0	9.2
11	0	0	86	10	0	0	0	10	30	0		15	0	0	11.6
12	0	0	36	5	0	0	20	90	5	0	0		0	5	12.4
13	0	0	43	5	10	0	25	10	0	0	0	0		25	8.3
14	0	0	78	20	0	0	45	10	0	0	0	0	0		11
Average	7.3	2.7	49.5	17.3	3.1	1.2	8.5	25.8	4.2	2.3	4.2	2.7	0.4	10	9.9

**Table 2: Neural Network Results**

Table 2 presents the overall FAR results observed for the comparison of user reference profiles against test samples (with the FRR of 30%). Each column indicates the false acceptance rate that was observed against the profile of a particular user by other user's test samples. The rows indicate the level of false acceptance achieved by each user when their samples were used to represent impostor cases (e.g. User 4 was able to pass as User 6 in 5% of cases). The bottom row gives the average FAR against each user, for an FRR of 30%. The final column gives the average impostor performance of each user (e.g. User 9 appears to be the most likely candidate to be able to masquerade as another user, with 17.5% chance of false acceptance). The average FAR of the classifier was determined as 9.9%.

Figure 2 illustrates the overall classification performance in graphical terms. Most significantly, it shows the average percentage FAR that was observed against each user's reference profile. These are the results used to measure the overall authentication performance of the classifier, indicating what percentage of impostors would be allowed access to a user's protected resources when that user themselves were rejected 30% of the

time. The dashed line indicates the average percentage of each user's test samples that were falsely accepted by the other user's references (i.e. how much each user contributed to the overall FAR that was observed).



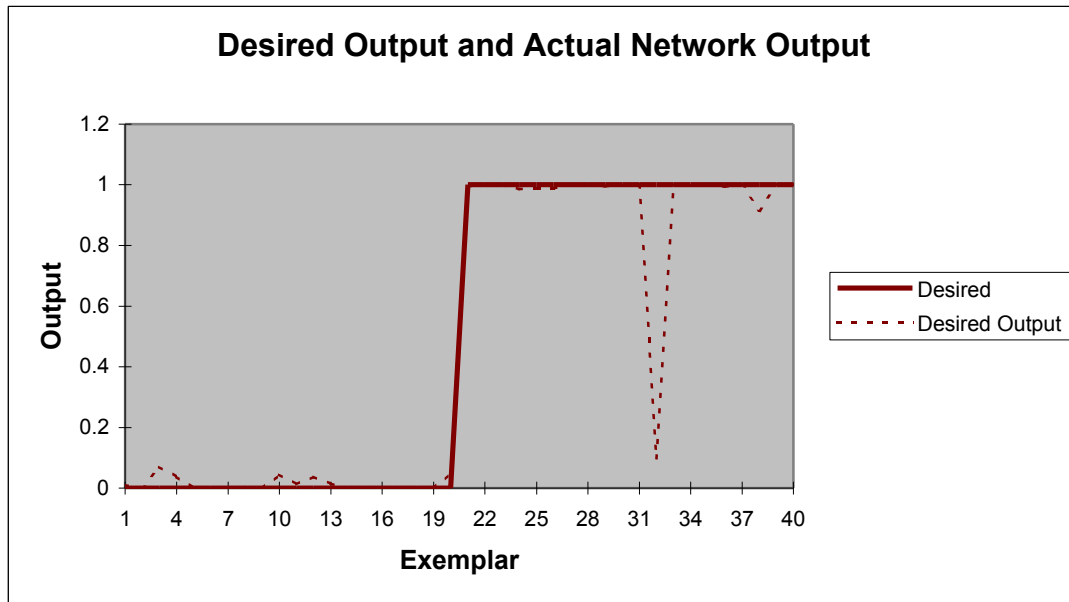
**Figure 2: Overall False Acceptance (FA) results with a 30% FRR**

The overall FAR of 9.9% was considered to be high (particularly in view of the accompanying high FRR). However, as indicated in table 2 and figure 2, this overall result was adversely affected by a minority of cases in which particular users scored extremely high levels of false acceptance (thus distorting the overall average obtained). As such, additional experiments were conducted in an attempt to further investigate and improve the results.

In the full study conducted by the authors, the neural network classifier was one of three classification techniques evaluated (the others being the Minimum Distance and Mahalanobis Distance classifiers). Across the three sets of results, the worst overall False Acceptance Rate observed was that of User 3's samples against User 4's reference (55%). As such, this was selected as the target for further investigation (note: it can be seen from table 2 that when considering only the results of the neural network approach, the comparison of User 3 against User 4 was *not* the worst case result). Further training of the network with the existing samples and additional layers did not yield any worthwhile improvement, so it was decided that the experiment should be repeated with longer signatures.

The data acquisition stage was repeated, but only using Users 3 and 4. It was decided that a 10 inter-keystroke latency signature should be tested, and the code chosen for both users to type was 01752237101. This was Spinnaker International's fax number with the area code. It was important to use a code that both users could remember easily. If this were not the case, additional error sources would have been added to the problem. Both users entered the same number of samples as before, i.e. 50. The first 30 were used as reference samples to train the network, and the next 20 as test samples.

The results showed that, for a threshold of between 0.073 and 0.098, the FAR and FRR were both 0, thus giving perfect classification. As illustrated in figure 3, only one of User 4's samples was less than 0.95, but as it was greater than all of User 3's output values, the threshold could be set between the highest User 3 output and the lowest User 4 output, to give the no error.



**Figure 3: Output of network with longer samples**

These results indicate that even the most difficult classification tasks can be alleviated by increasing the length of the feature vectors. A longer feature vector provides more information and thus it was expected that the results would improve. However, increasing the length of the input code has practical implications for the end user (i.e. limiting their ability to easily remember it). As such, it would not represent a solution in all contexts.

The full results of the study, which also included the investigation of the Minimum Distance and Mahalanobis Distance classifiers, can be found in Ord (1999). These other classifiers were found to produce inferior results to the neural network approach reported in this paper.

#### **4. Practical applications**

The techniques used in this study can be implemented in any application that involves user authentication with a numeric keypad. The most common application that involves the user authenticating themselves via a numerical keypad is in ATM systems. The authentication in this case is currently provided by the user's knowledge of their PIN code. If an impostor obtains this code, that user's resources can be accessed. The incorporation of keystroke analysis would add an extra layer of security, in that knowledge of the PIN would not guarantee access – the typing style of the impostor would also be required to be similar. Other biometric technologies are already being considered in an ATM context, an example being an iris recognition system that has been under trial in the UK by the ATM manufacturer NCR (NCR, 1999).

Another potential market for products using these techniques, is a Universal Personal Telecommunications (UPT) terminal. UPT aims to deliver a personal mobility, allowing a user to receive telecommunication services on any terminal in any network, by identifying

users by a unique personal number. Additional transparent security could be provided by this system, as apart from a terminal transmitting the UPT number to a Central Control System, the inter-keystroke times could be transmitted as well. The latencies would be measured at the terminal, and the Central Control System would calculate whether the identified user was an impostor or valid user. It would then withhold any service rights if the system decided that the user was an impostor. These techniques can also be applied to terminal mobility scenarios, such as mobile handsets, for PIN-based user authentication and other security numbers (Furnell et al. 1996).

A major concern in practice would be the robustness of the authentication technique. In an application such as ATMs, the potential users could be in various states of mind or body, i.e. they could be intoxicated, ill, or have an injured hand. This could seriously effect the performance of the classifiers, as users could deviate from their 'normal' inter-keystroke signature. Any biometric used in an ATM or UPT application has to take these factors into account. Hence, authentication techniques that rely on a user's actions, such as these, cannot perform to the same accuracy in these circumstances as physiology-based biometrics. Physiological characteristics, such as the iris, do not change whatever the state the user is in. However, two major advantages that keystroke-analysis based authentication techniques have are that they are transparent to the user and require no additional equipment, making them cheaper to implement. The one major limitation of iris and retinal recognition systems is that they rely upon the co-operation of the user and many users have reservations about staring into a device which is going to scan their eyes.

The results observed suggest that the approach is currently not accurate enough for use with large-scale systems such as ATMs and UPT. To give an example of the error rates that would be considered acceptable in these contexts, error for iris scan ATM biometrics are in the order of 0.00076% for when the FRR equals the FAR (Biometric Consulting Group 1998). A more appropriate application could include keypad-based access control for secure systems or areas within an organisation.

## **5. Conclusions**

The use of biometrics in commercial environments is rapidly increasing as the technology required decreases in cost. The low cost and transparency of the keystroke analysis biometric has made it possible for it to compete with the other more expensive alternatives. This study investigated the possibility of its use in conjunction with codes typed on a numeric keypad.

The overall results were an FAR of 9.9% with an FRR of 30%. This result was higher than observed in the previous studies summarised earlier in the paper. However, these previous results were based on a simpler problem, as a full keyboard was used in all for data collection. The code used for the main experiments in this study was only 6 digits long, dramatically shorter than in these previous experiments. Consequently, a decrease in the performance was expected. The use of an 11-digit code (i.e. with 10 measurable latencies), proved that the results of the neural network classifier could be significantly improved (with the worst error component being reduced from 60% to 0%). However, it must be acknowledged that, in a practical context, a PIN code of 11 digits would represent too long a string for most users to easily remember.

Additional experimentation is required with longer numerical codes and for an increased number of users, to test the viability of keystroke analysis for large-scale biometrics such as in ATM's and UPT services. This study demonstrated that the Multi-Layer Perceptron Classifier has potential in this application, and further development in this area could improve these results further.

## 6. References

- Biometric Consulting Group. 1998. Web Page: <http://biometric-consulting.com/bio.htm>.
- Bishop, C.M. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.
- Bleha, S., Slivinsky, C. and Hussien, B. 1990. "Computer-Access Security Systems Using Keystroke Dynamics", *Transactions on Pattern analysis and Machine Intelligence*, vol 12., no. 12.
- Furnell, S.M., Green, M., Hope, S., Morrissey, J.P and Reynolds, P.L. 1996. "Non-Intrusive Security Arrangements to support Terminal and Personal Mobility", in *Proceedings of EUROMEDIA 96* (London, UK, 19-21 December 1996): 167-171.
- Gaines, R., Lisowski, W., Press, S. and Shapiro, N. 1980. "Authentication by keystroke timing", Rand Report R-256-NSF. Rand Corporation.
- Gentner, D.R. 1983. "Keystroke timing in transcription typing". in W. E. Cooper (Ed.), *Cognitive aspects of skilled typewriting*. New York: Springer-Verlag: 95-120.
- Jobusch, D.L. and Oldehoeft, A.E. 1989. "A Survey of Password Mechanisms : Part 1", *Computers & Security*, Vol. 8, No. 7: 587-604.
- Joyce, R. and Gupta, G. 1990. "Identity Authentication Based on Keystroke Latencies", *Communications of the ACM*, Volume 33, February 1990.
- Legget, J. and Williams, G. 1988. "Verifying identity via keystroke characteristics", *International Journal of Man-Machine Studies*, 28.
- Looney, C. 1997. *Pattern Recognition using Neural Network*, Oxford University Press.
- Miller, B. 1994. "Vital signs of identity", *IEEE Spectrum*, February 1994.
- NCR. 1999. "NCR announces iris recognition trials with Nationwide Building Society". <http://www3.ncr.com/product/financial/press/sensnat.htm>
- Ord, T. 1999. *User Authentication using Keystroke Analysis with a Numerical Keypad approach*. MS.c. Thesis. University of Plymouth, Plymouth, United Kingdom.