Fuzzy approach in biometric authentication by keystroke dynamics

JAN CAPEK, MILOSLAV HUB
Institut of System Engineering and Informatics
Faculty of Economic and Administration
University of Pardubice
Studentska 84, 530 09 Pardubice
THE CZECH REPUBLIC

http://fes.upce.cz

Abstract: A person’s identity verification became very important in information society. This article presents some results of our research of biometric authentication by keystroke dynamics. The comparison of the stochastic approach and using fuzzy numbers is presented as well.

Key-Words: authentication, biometric, keystroke dynamics, password, fuzzy numbers

1 Introduction
Authentication as a data security instrument in our information society is very important for keeping our data as safe as possible. The aim of authentication is to decide whether some subject is really claimed subject. There are 3 types of authentication: authentication by knowledge, authentication by ownership of something, and authentication by attribute. Each one has both advantages and disadvantages. They can be combined to increase the security of our information as well. It is well known that password do not prove high security level. From previous works for example [7, 14].

The one possible way to increase security level of access to information systems is the combination of authentication by knowledge and authentication by attribute ie parallel usage of passwords and keystroke dynamics. Everyone has different keyboard typing [12]. It is similar to one’s own signature.

In keystroke dynamics it is possible to recognize various kinds of identifyably characteristics which are measurable: duration times (times between key press and key release of the same key), latency times (times between key release of the first key and key press of the second key), key typing speed, position of the finger on the key, pressure on the key and so on.

A template records the specific rhythm and touch, which is used to verify the user's identity at next logon. Research on keystroke dynamics—the study of individual typing patterns—shows that an individual's typing style is unique. That means that even if someone knows your password, he won't be able to authenticate as you.

The most identifiable characteristics for this biometric authentication system appear to be keystroke duration and latency times because no special hardware is necessary.

But using of keystroke dynamics together with passwords entails trouble. Passwords are usually very short in practice. In spite of all recommendations, and they provide very little amount of identifiable characteristics [7]. For this reason, it is necessary to find a good algorithm for user and impostor recognition. This means to find an algorithm which minimizes potential errors.

2 Related works
Beginning of keystroke dynamics is dated in 19th century when telegraph operators where able to recognize each other only on their keying dynamics [15]. Using of keystroke dynamics in authentication was suggested by Spillan in 1975 [12]. However, the first scientific study in this field was carried out till in 1980 by Gaines and his research group [5]. In this research was studied keyboard typing of 300 – 400 word’s long text by 7 secretaries. Conclusion of this work is proved statistical dependence of times of big-rams typing by the same user. This experiment had a lot of imperfections, especially very small number of testing persons. In thr 90’s a lot of researches continued in Gaines work [10, 11, 15]. They were oriented to long texts too.

Radical changes were in Garcia’s patent of 1986 [6]. His idea was that the best data for this
authentication is an individual username. The
username is several times written and by means is
created template of latency times. When a person
wants to access computer resource, he is required to
type his name. The latency vector of the keystroke
times is compared with template stored in the
computer. If this claimant’s latency vector and
template are similar, the user is granted access to
the system. As a criterion of similarity he used so
called Mahalanobis distance function.

The next important approach was patented by
Young and Hammon [16]. They are using not only
keystroke latency, but keystroke latency and
duration at the same time. As criterion they selected
Euclid distance function \(d_E\):

\[
d_E(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T \times (\mathbf{x}_1 - \mathbf{x}_2)}
\]

\(\mathbf{x}_1, \mathbf{x}_2 \ldots\) N-dimension vectors

After the success of this work, a lot of work
followed. The most important and the most cited
are Joice [8] and Bleha [2] researches. Joice had
suggested using user’s name, surname and
password together as data for this biometric
authentication. As a criterion of similarity between
template and access vectors, he recommended sum
of absolute differences between template and
access times \(d_A\):

\[
d_A(\mathbf{x}_1, \mathbf{x}_2) = \sum_i |x_{1,i} - x_{2,i}|
\]

\(\mathbf{x}_1, \mathbf{x}_2 \ldots\) N-dimension vectors

Bleha had chosen different metrics of
similarity. He suggested applying of two different
metrics. The first one was normalized distance function \(d_N\), the second one was normalized Bayes
classificator \(d_B\):

\[
d_N(\mathbf{x}_1, \mathbf{x}_2) = \frac{(\mathbf{x}_1 - \mathbf{x}_2)^T \times (\mathbf{x}_1 - \mathbf{x}_2)}{\|\mathbf{x}_1\| \cdot \|\mathbf{x}_2\|}
\]

\(\mathbf{x}_1, \mathbf{x}_2 \ldots\) N-dimension vector

\(\|\mathbf{x}_1\|, \|\mathbf{x}_2\| \ldots\) Norms of vectors are defined in
following eq. (4)

\[
\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \times \mathbf{x}}
\]

\(\mathbf{x}\) \ldots N-dimension vector

As at present, the American firm Bionet Systems offers BioPassword. BioPassword is
patented authentication software based on
keystroke dynamics with undisclosed algorithm
[17]. However, Bragg [3] after her own experience
gave following result of BioPassword. She says:
“In my test, logon worked as documented. I could
keep that post-it note with my password on the
monitor, but as long as I logged off, no one could
log on as me. Similarly, I couldn’t logon if I
purposefully changed my typing style.”

3 Our fuzzy approach

The first idea of this approach, as we know, was
published by Capek in 2004 [4]. This idea is
seemingly simply – to create a fuzzy inference
system for every template. Suppose the subject
which is going through authentication is posing as
a subject \(S_j\) and he is presenting \(n\) identifiable
features \(\{i_1, \ldots, i_n\}\). Then the fuzzy inference system for
j-th template is defined as:

\[
\text{if } i_1 \text{ is } I_1 \text{ and } i_2 \text{ is } I_2 \text{ and } \ldots \text{ and } i_n \text{ is } I_n \text{ than access}
\]

A task is to find suitable membership functions
of \(I_1, I_2 \ldots I_n\) for every template. It means to find
an algorithm capable of creating this membership
functions only from template data.

Certainly, the center of a membership function
is localized at a mean of the concrete feature \(\bar{x}_{i,j}\).
However, some features are more typical for j-th
user than others. Suitable characteristic of
“typicality” of some features is his standard
deviation \(s_{i,j}\). Thus, possible membership function
\(\mu(x_{i,j})\) of i-th feature of j-th template can look as
follows:

\[
\mu(x_{i,j}) \uparrow \quad \text{The width depends on } s_{i,j}
\]

\[
0 \quad \bar{x}_{i,j} \quad 1 \quad \text{Value of feature}
\]

\[Fig. 1 A possible membership function for I_{i,j}\]
\[
\mu(i, j) = \begin{cases} 
0 & x \leq \bar{x}_{i, j} - 2 \cdot \sqrt{6} \cdot s_{i, j} \\
\frac{x - \bar{x}_{i, j}}{2 \cdot \sqrt{6} \cdot s_{i, j}} + 1 & \bar{x}_{i, j} - 2 \cdot \sqrt{6} \cdot s_{i, j} < x \leq \bar{x}_{i, j} \\
\frac{\bar{x}_{i, j} - x}{2 \cdot \sqrt{6} \cdot s_{i, j}} + 1 & \bar{x}_{i, j} < x < \bar{x}_{i, j} + 2 \cdot \sqrt{6} \cdot s_{i, j} \\
0 & x \geq \bar{x}_{i, j} + 2 \cdot \sqrt{6} \cdot s_{i, j} 
\end{cases}
\]

\[x_{i, j, k} \text{ ..........k-th value of i-th feature during creating of j-th template}
\]

\[n_j \text{ ..........Number of measures for creating of j-th template}
\]

\[s_{i, j} = \frac{1}{n_j - 1} \cdot \sum_{k=1}^{n_j} (x_{i, j, k} - \bar{x}_{i, j})^2
\]

\[n_j \text{ ..........Number of measures for creating of j-th template}
\]

\[x_{i, j, k} \text{ ..........k-th value of i-th feature during creating of j-th template}
\]

\[\bar{x}_{i, j} \text{ ..........Mean of i-th feature of j-th template}
\]

### 4 Experiments

For purpose of testing of our approach there has been created a special software which is able to measure latency and duration times during typing appointed text works with precision microseconds. This software consists of two parts – client part (programmed in Java) which measures relevant times and server part (programmed in PHP) which saves measured times into database. Three different passwords have been selected: “biometrika”, “informatika” and “koleje” (“biometrics”, “informatics” and “hostel” in English) for testing. These words have been selected for their different lengths. Students with age range 19-25 were asked to write these words fifteen times. For precision it is necessary to add that chosen students are common users of computer and they have been familiarized of this experiment. The next week they were asked to write this words one time and the next week again. Measured data has been continuously saved into database for the future processing. The following number of students has complied to the criterion of fifteen measures for template and consecutive two measures for an access:

- 80 students for the word “koleje”
- 42 students for the word “biometrika”
- 45 students for the word “informatika”

Individual accesses were tested both against student’s own template and against other student’s templates. It has been obtained by the following number of valid accesses \(N_V\) and number of impostor accesses \(N_I\):

\[N_V = N_a \cdot N
\]

\[N_I = N_a \cdot N \cdot (N - 1)
\]

\[N_a \text{ .......... Number of access by one person}
\]

\[N \text{ .......... Number of templates}
\]

The following table describes numbers of valid accesses and impostors for the selected words:

<table>
<thead>
<tr>
<th>Word</th>
<th>Valid accesses</th>
<th>Impostor accesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>koleje</td>
<td>160</td>
<td>12640</td>
</tr>
<tr>
<td>biometrika</td>
<td>84</td>
<td>3444</td>
</tr>
<tr>
<td>informatika</td>
<td>90</td>
<td>3960</td>
</tr>
</tbody>
</table>

The goal of the authentication process is to decide whether a subject is claimed subject, in other words “if you are really you”. This decision may not always be valid. The possible situations are shown in the next table.

<table>
<thead>
<tr>
<th>Valid access</th>
<th>Impostor access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptance</td>
<td>Rejection</td>
</tr>
<tr>
<td>Suitable situation</td>
<td>False rejection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Valid access</th>
<th>Impostor access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptance</td>
<td>Suitable situation</td>
</tr>
</tbody>
</table>

This is a reason why as criterion for comparison of our model and contemporary models we have chosen false acceptance ratio FAR and false rejection ratio FRR, which are defined by following formulas:
\[
FAR = \frac{NFA}{NIA} \quad \text{(11)}
\]
\[
FRR = \frac{NFR}{NVA} \quad \text{(12)}
\]

\begin{align*}
NFA & \quad \text{Number of false acceptances} \\
NIA & \quad \text{Number of impostor accesses} \\
NFR & \quad \text{Number of false rejections} \\
NVA & \quad \text{Number of valid accesses}
\end{align*}

It is necessary to note that FAR is dependent on FRR and it is not possible to decrease FAR without increasing FRR. Reversing statement is valid too. Various membership functions for i-th template and j-th feature were tested. Testing was done both for duration and latency times and for duration together with latency times.

In the following graphs the negative exponentially-weighted smoothing is used [13].

Fig. 2 Word "biometrika", duration times

Fig. 3 Word "biometrika", latency times

Fig. 4 Word "biometrika", both duration and latency times

Fig. 5 Word "informatika", both duration and latency times

Fig. 6 Word "koleje", both duration and latency times
For the last word “koleje” we show pictures (7) and (8) the individual fuzzy numbers constructions for both the accepted (valid) access and the rejected (impostor) access.

If we compare both figures (7) and (8), we can see that for letter o, both numbers are inside core of the fuzzy numbers.

This situation is for one letter only, so it is not enough for accepting both persons. Theoretically it is possible to create a table for how many alphabetic characters from password must be inside core of the fuzzy number for accepting user.
5 Conclusion and future work
It is evident from previous graphs fuzzy approach gives better results than conventional algorithms, based on the stochastic approach. This attitude offers a smaller danger of false acceptance errors and false rejection errors. A special hardware is not necessary for implementation of this algorithm and hence the costs will not increase.

References: