Abstracts of Dissertations

Institute of Information and Communication Technologies

BULGARIAN ACADEMY OF SCIENCES

1 / 2014

A COMBINED APPROACH TO ON-LINE SIGNATURE RECOGNITION

Desislava Boyadzhieva

КОМБИНИРАН ПОДХОД ЗА РАЗПОЗНАВАНЕ НА ОН-ЛАЙН ПОДПИСИ

Десислава Бояджиева

Автореферати на дисертации

Институт по информационни и комуникационни технологии

БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ

ISSN: 1314-6351

Поредицата ***„Автореферати на дисертации на Института по информационни и комуникационни технологии при Българската академия на науките“*** *представя в електронен формат автореферати на дисертации за получаване на научната степен „Доктор на науките” или на образователната и научната степен „Доктор”, защитени в Института по информационни и комуникационни технологии при Българската академия на науките. Представените трудове отразяват нови научни и научно-приложни приноси в редица области на информационните и комуникационните технологии като Компютърни мрежи и архитектури, Паралелни алгоритми, Научни пресмятания, Лингвистично моделиране, Математически методи за обработка на сензорна информация, Информационни технологии в сигурността, Технологии за управление и обработка на знания, Грид-технологии и приложения, Оптимизация и вземане на решения, Обработка на сигнали и разпознаване на образи, Интелигентни системи, Информационни процеси и системи, Вградени интелигентни технологии, Йерархични системи, Комуникационни системи и услуги и др.*

The series ***Abstracts of Dissertations of the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences*** *presents in an electronic format the abstracts of Doctor of Sciences and PhD dissertations defended in the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences. The studies provide new original results in such areas of Information and Communication Technologies as Computer Networks and Architectures, Parallel Algorithms, Scientific Computations, Linguistic Modelling, Mathematical Methods for Sensor Data Processing, Information Technologies for Security, Technologies for Knowledge management and processing, Grid Technologies and Applications, Optimization and Decision Making, Signal Processing and Pattern Recognition, Information Processing and Systems, Intelligent Systems, Embedded Intelligent Technologies, Hierarchical Systems, Communication Systems and Services, etc.*

e-ISSN: 1314-6351

BULGARIAN ACADEMY OF SCIENCES

Abstract of PhD Thesis

A COMBINED APPROACH TO ON-LINE SIGNATURE RECOGNITION

Desislava Nikolova Boyadzhieva

Supervisor: Assoc. Prof. Georgi Gluhchev

Approved by Supervising Committee:

Prof. Stefan Hadjitodorov
Prof. Krasimir Atanasov
Assoc. Prof. Dimo Dimov
Assoc. Prof. Pencho Venkov
Assoc. Prof. Georgi Gluhchev

INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES
Department of Signal Processing and Pattern Recognition

The PhD thesis was discussed and allowed to be defended during an extended session of the Department of Signal Processing and Pattern Recognition at IICT-BAS, which had been held on January 17, 2014.

The defense of the PhD thesis was held on 02.06.2014, 2014 at 10:00 am in Room 507, Block 2, IICT-BAS.

The full volume of the dissertation is 129 pages. List of figures, list of tables and terminology dictionary are given at the beginning of the dissertation. The dissertation consists of five chapters (p. 7-120). It includes also a conclusion, a contribution summary, and a list of publications (p. 120-122). The list of references contains 108 titles (p. 123-129). The text of the dissertation includes 19 tables and 18 figures.

Keywords: on-line signature verification, neural networks (NNs), signature features, feature selection, Mallows *Cp*, SUsig database, genuine and forgery signatures, *k*-nearest neighbours (*k*NN), *N*-fold cross validation, Leave one out cross validation (LOOCV).

Introduction

Signature recognition is the process of confirming the identity based on the handwritten signature of the user as a form of behavioral biometrics [1]. From one hand, the signatures are a convenient, widely used and secure mean for authentication, and from the other, their input to biometric systems is fast, easy, natural and non-invasive. For these reasons, the problem of the signature verification is broadly investigated in the past years. Novel methods and algorithms are developed, mostly for on-line signatures, and lots of them are implemented in practice [1, 2, 3, 4, 16, 17]. Current research in the field of signature recognition is directed towards the development of combined approaches as well as the identifying of optimal feature subsets to discriminate between genuine and forgery signatures.

In this thesis, the subject of research is on-line signature recognition. The aim of the research work is the development of a novel combined method for on-line signature recognition and its software implementation.

The above formulated goal is achieved by fulfilling the following tasks:

1. Signatures acquisition and creation of signature database;
2. Extraction of novel and known signature features;
3. Selection of optimal signature feature subsets;
4. Investigation of various NN models for verification;
5. Carrying out experiments, evaluation of classification accuracy and comparison with known classifiers for own and public signature databases;
6. Development of a prototype of a software system for signature acquisition, signature visualization, feature extraction, feature selection and verification.

The present study is conducted in order to find out the answers of the following questions:

1. Is there a particular feature subset for each user describing his or her signature writing style?
2. Are there significant differences between the verification results for common and individual feature subset?
3. Will the verification accuracy decrease if skilled forgeries are used as negative examples for classifier training?

Methodology

The study is in the field of pattern recognition; it is related to the application of statistical methods for signature features selection and the application of NN and *k*NN classifiers for on-line signature verification. In this thesis, two signature databases are used in order to experiment with the proposed combined method for on-line signature verification. These are SUsig database of 89 users [5] and an own database of signatures of 8 users collected for the aims of the study. Both signature databases consist of genuine signatures, skilled and random forgeries.

The development of a signature recognition system consists of the following steps [6]: signature acquisition, preprocessing, feature extraction, feature selection, verification and accuracy estimation. Below we shall describe our method in the terms of these steps.

1. *Signature acquisition*: Signature data is acquired by a graphical tablet (*Wacom Intuos3 A5 PTZ-630* for own signature database and *Wacom Graphire2* for SUsig database). Raw data consist of the following information about each signature point: *x* and *y* coordinates, pressure level, timestamp, stroke indicator. For our signature database we have also information about the azimuth and the pen tilt for each point. A signature acuisition protocol is created in order to describe the steps in the signature acquisition process.

2. *Signature preprocessing*: To facilitate feature extraction, it is necessary the raw data to be preprocessed. The operations applied depend on the selected features and the acquisition protocol. The coordinates *x* and *y* of the ink coordinate space are called *himetric* units [7] and their values fall in [0, 7999] x [0, 5999]. It is necessary to transform them in the application coordinate system in [0, 1279] x [0,799]. This is performed automatically by a method from *Microsoft Tablet PC SDK*. Since the acquired signatures may be rotated, we have to align them horizontally. The next pre-processing operation is translation of the signatures to a given point of the application coordinate system because it is possible some of the coordinates to obtain negative values after rotation. So the following operations are performed to all the signature data in the databases: coordinate transformation, rotation and translation.

Table 1 Global features

A1	Signature length *L*	*A10*	distance between initial and end point	*A19*	distance between rightmost and end points
A2	Signature height *H*	*A11*	angle of the line between center and initial points	*A20*	angle of the line between leftmost and initial points
A3	height to width ratio *H/L*	*A12*	angle of the line between center and end points	*A21*	angle of the line between end and rightmost points
A4	number of points *N*	*A13*	angle of the line between initial and end points	*A22*	Number of strokes
A5	time duration	*A14*	distance between leftmost and center points	*A23*	Average tilt value
A6	Number of segments	*A15*	distance between center and rightmost points	*A24*	Average pressure value
A7	Signature density A4/A1*A2	*A16*	angle of the line between center and leftmost points	*A25*	Average pen tilt value
A8	distance between initial and center point	*A17*	angle of the line between center and rightmost points	*A26*	Average azimuth value
A9	distance between end and center point	*A18*	distance between leftmost and initial points		

3. *Feature extraction*

There are three groups of signature features: global, local and segmental [8]. Global features are extracted for the whole signature, local features are extracted for each sample point in the signature, and segmental features are extracted for each signature segment. Over 100 features

used in signature verification are listed in [9]. The extracted global signature features used are presented in Table 1. They have a clear meaning for criminology experts. To our knowledge signature features A8-A21 are not used so far in signature recognition. We extract all the features for the signatures from our database; we extract features A1-A24 for SUsig database.

4. *Feature set selection*

Since some features demonstrate higher discriminatory capability than others, feature selection should be performed. This is related to the process of selecting k features of most discrimination power out of p available ones ($k \leq p$) and it aims to identify and remove as much irrelevant and redundant information as possible. A review of the processes of feature set selection for signatures is done in [8].

We approach feature set selection step in signature verification in two ways (1) by using a *common* feature subset for all users, and (2) by using an *individual* feature subset for each user.

At the beginning we extract all the signature features for all database users and we perform *z-score* feature normalization. Let us denote by f_i the value of the ith feature before normalization, $i = 1,...,k$. The new value f_i^{zscore} of f_i is given by:

$$f_i^{zscore} = \frac{f_i - \mu_{f_i}}{\sigma_{f_i}}, i = 1,...,k$$

$$\mu_{f_i} = \frac{1}{n}\sum_{j=1}^{n} f_{ij}$$

$$\sigma_{f_i} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(f_{ij} - \mu_{f_i})^2}$$

We find all high correlated features at 0.01 confidence level, 99% confidence interval, met in more than 25% of the users. We use Pearson correlation coefficient. For particular features $\mathbf{X_1}$ and $\mathbf{X_2}$ it is given by:

$$r_{X_1,X_2} = \frac{\sum_{i=1}^{p}(X_{1i} - \mu_{X_1})(X_{2i} - \mu_{X_2})}{\sqrt{\sum_{i=1}^{p}(X_{1i} - \mu_{X_1})^2 \sum_{i=1}^{p}(X_{2i} - \mu_{X_2})^2}}$$

By applying the correlation pleiads method [12] we identify all groups (pleiads) of features having high intraclass correlation and low interclass correlation and leave only one random feature in a group. For each pleiad we retain only one feature. In this way we create the *common* feature subset.

In order to find *individual* feature subset for each user we apply the methods of Hocking, Leslie and LaMotte for selection of regression variables based on *Mallows Cp* criterion for regression [10, 11] on the already found common feature subset. This criterion is used to decide on suitable subset among contending subsets. It is a measure of the standardized total squared error defines as follows:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p),$$

RSS_p denotes the residual sum of squares for the particular regression with p variables and $\hat{\sigma}^2$ is an estimate of residual mean square σ^2 for full regression.

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{j=1}^{n}(y_j - \sum_{i=1}^{k}\beta_i x_{ij})^2 = \frac{RSS_k}{n-k}.$$

If a model is adequate, i.e. does not suffer from lack of fit, then

$$E(C_p) \approx p$$

This means that we expect *Cp* value to be about *p*. A plot of *Cp* versus *p* displays the adequate models as points close to the line *Cp* = *p*. Subsets with small values of *Cp* and values of *Cp* close to *p* are considered good. Hocking and Leslie [10] further describe a method which allows thus subset to be identified after consideration of only a small fraction of all $\binom{k}{p}$ possible subsets of size *p*. LaMotte and Hocking [11] modified this algorithm in a way that moderately large problems can be treated with minimum of computation. The algorithm specifies the subset of size *r* to be deleted. In the following, the terms *r*-subset and *p*-subset refer, respectively, to subsets being deleted and subsets being retained.

The method for selection of best subset is based on *m*-variable reductions,i.e. reductions in the regression sum of squares due to eliminating subsets of size *m* from the *k*-variable equation. Typically $1 \le m \le 4$ and $m = 1$ in the original method [10]. These *m*-variable reductions are used to determine the best *r*-subset to be removed, for $r > m$. The reduction in the regression sum of squares due to removing a set of *r* variables is given by:

$$\mathrm{Re}\,d_r = RSS_p - RSS_k$$

The set of *r* variables for which this reduction is minimum specifies the subset of size *p* ($p = k - r$) variables in the regression to be retained for which residual sum of squares is minimum. It is suggested in [10] that *Cp* statistic can also be computed by using this reduction in the following way:

$$C_p = \frac{Red_r}{\hat{\sigma}^2} - (2p - k)$$

The steps of the generalized algorithm are as follows [11]: First, the *k*-variable equation is fitted with all variables present and the reductions in the regression sum of squares, called univariate reduction, due to deleting each of the *k*-variables, are calculated. If a single variable, say *i*th is removed from the regression, its univariate reduction is given by:

$$\theta_i = \hat{\sigma}^2 t_i^2$$

$$t_i^2 = \frac{b_i^2}{\hat{\sigma}^2_{b_i}}$$

The square of *t*-statistic associated with *i*th regression coefficient is denoted by t_i^2 and the standard error of the estimated coefficient b_i is denoted by $\hat{\sigma}^2_{b_i}$. The standard errors of the estimated coeficients are the square roots of the diagonal elements of the coeficientcovariance

matrix. After this, the variables are relabeled according to the order of these univariate reductions.

$$\theta_1 \le \theta_2 \le ... \le \theta_k$$

Having the variables relabeled, the reductions in the regression sum of squares due to all of the $\binom{k}{m}$ subsets of size *m* are evaluated. These are called *m*-variable reductions. Each subset of size *m* is denoted by the subscripts of the variables in that subset in increasing order. That is, a *m*-subset is described as $(i_1, i_2, ..., i_m)$, where $i_1, i_2, ..., i_m$ where $1 < i_j < k$ and $i_1 < i_2 < ... < i_m$. Next these *m*-variable reductions are ordered in increasing order of magnitude and are used to define stages for inspecting the *r*-subsets. Only those *m*-subsets whose first index is $(r - m + 1)$ or greater are used to define a stage and hence we evaluate only them. The *r*-subsets in a stage defined by an *m*-subset contain *r* indices consisting of those indices in the defining *m*-subset and $(r - m)$ indices which are less than the first index in the *m*-subset. There are total of $\binom{k-r+m}{m}$ stages defined in this way. The stages are numbered according to the magnitude of the *m*-variable subset which define them. Thus *Stage 1* will consist of *r*-subsets defined by the *m*-subset with smallest reduction. In general, at the *q*th stage we evaluate the reductions due to all subsets defined in the *q*th stage and ask if the smallest reduction in the regression sum of squares computed for an *r*-subset obtained in all Stages before the *q*th stage $(1, ..., q)$ is less than the reduction due to the *m*-subset defining Stage *(q+1)*. If so, we terminate having identified the best subset of size *p* obtained by the corresponding *r*-subset and if not we proceed to Stage *(q+1)*.

By applying the methods of Hocking, Leslie and LaMotte we identify best feature subsets of various size for each user on the basis of his/her eight or ten genuine signatures and ten random forgeries. Among these subsets we select the best subset that have *Cp* value closest to *p*, where *p* is the number of regression coefficients. Thus, for each user we obtain the best feature subset of different size.

5. *Verification*

Neural networks are suitable to be used for signature verification since they are an excellent generalization tool (under normal conditions) and are a useful means of coping with the diversity and variations inherent in handwritten signatures [13]. Usually, a particular NN is built for each user on the basis of his/her genuine and forgery signatures. The number of input neurons is *p* where *p* is the number of the features. The single output neuron has a value *1* for genuine signature and a value *0* for forgery signature. After the training, a score threshold is determined. If the verification result (at the time of testing of a signature) is greater than the corresponding score threshold, the signature is considered genuine, otherwise – forgery. This approach is widespread because it allows fast adding and deleting of signatures for new and existing users [13]. Usually, NN training takes lots of time but in this approach it is done off-line so the users are not forced to wait.

We compare NN classifiers with *k*NN classifiers with Euclidean distance. Before applying the *k*NN classifier, we first normalize the features. Since $k_{max} = \sqrt{Ntrn}$, where $Ntrn$ is the number of signatures used for training and it is recommended for the value of *k* to be an odd number, we choose *k*= 1 and *k* = 3.

For each user, several NN and kNN models are evaluated by 10-fold cross validation and LOOCV respectively. The "optimal" models are found together with their parameters:

number of hidden neurons for NN, type of signature forgeries for training, input features and value of k.

After finding the "optimal" model of a classifier by cross validation, the corresponding classifier is being trained on all the user's signatures and is ready to be used for verification.

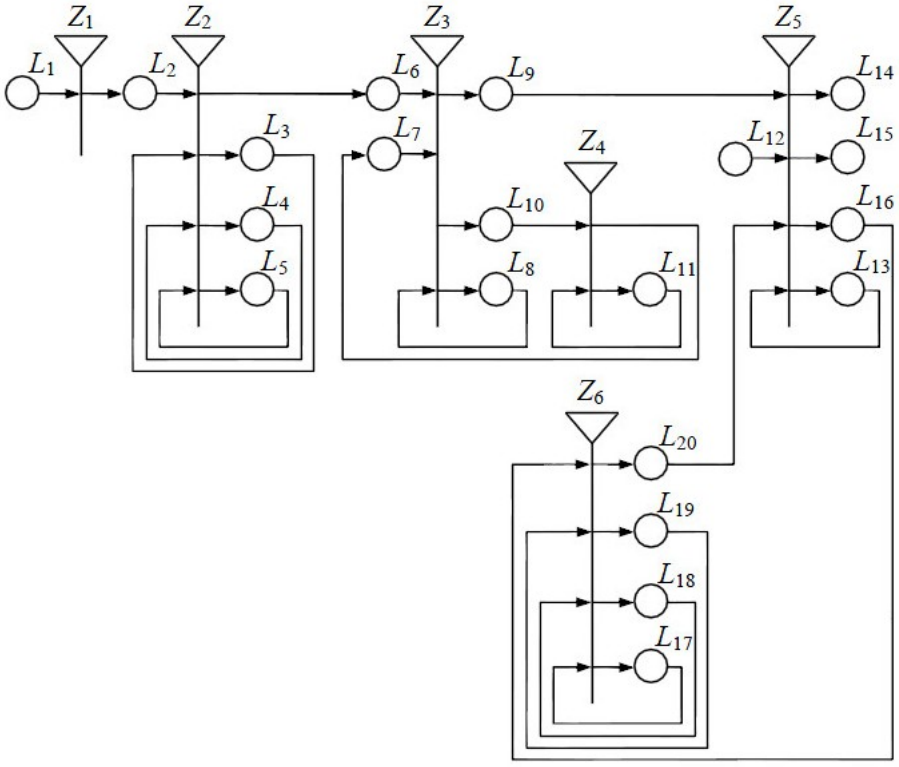
6. *Accuracy estimation*

The performances of classifiers are evaluated by the following well known metrics: FAR (false accept rate), FRR (false reject rate), TAR (true accept rate), TRR (true reject rate), and Accuracy.

GN model for on-line signature verification

A reduced GN is constructed and it is presented in Fig. 1. For this model tokens keep all their history. The model is built of six transitions which correspond to the steps in online signature verification, twenty places and three types of tokens α, β and γ.

Fig. 1 GN model for on-line signature verification



Experimental results

Experiments are carried out in MATLAB environment. We use *Neural Network Toolbox*. We experiment with two types of classifiers: NN and *k*NN with varying parameters values (number of features, forgery signature types, number of hidden neurons *H* and number of neighbors' *k*). Since we have small amount of data, we evaluate classifier accuracy with *N*-fold cross validation [14]. We use 10-fold CV for NN parameters tuning and LOOCV for *k*NN parameters tuning.

Let us denote by *Var.1* the case in which feature subset is determined by using the genuine and random forgery signatures and denote by *Var.2* the case in which feature subset is determined by

using the genuine and skilled forgery signatures. Let us denote by *Case 1* the case in which only random forgeries are used for NN training, and denote by *Case 2* the case in which both random and skilled forgeries are used.

Results on own signature database

By applying the method of correlation pleiads, the initial number of features - 26, is reduced down to 17 and the remaining features are A2, A5, A6, A8, A11, A12, A13, A15, A16, A17, A20, A21, A22, A23, A24, A25, A26.

Using *Var.1* the initial number of features is reduced down to a number less than 13 for the half of the users, while using *Var.2* the nitial number of features is reduced down to 13 for 7 of 8 users.

The average value of the estimated accuracy by cross validation for all users is equal to 97.5% for NN classifier and it is equal to 92.5% for kNN classifier.

Results on SUsig database

By applying the method of correlation pleiads, the initial number of features - 24, is reduced by around 50% and the remaining features are A1, A2, A4, A6, A10, A12, A13, A16, A17, A21, A22, A23, A24.

The size of the obtained *p*-subset and the corresponding number of users are specified in Table 2. There is a significant reduction in features number for both *Var.1* and *Var.2* since its initial number - 13 is reduced down to 9 for about half of the users, reduced down to 5 or 6 features for 30% of the users.

Table 2 Individual p-subsets

Var.	Size of p-subset	Number of users
1	9	42
1	8	5
1	7	7
1	6	12
1	5	14
1	4	5
1	3	4

Var.	Size of p-subset	Number of users
2	9	48
2	8	9
2	7	4
2	6	10
2	5	15
2	4	3
2	3	0

In Table 3, all the NN and kNN models are described together with their parameters.

Table 3 Parameters of the NN models

<table>
<tr><th># of model</th><th>Features (input neurons)</th><th>Genuine signatures</th><th colspan="2">Forgery signatures</th><th>Number of hidden neurons H</th><th>Number of neighbors k</th></tr>
<tr><td>1</td><td>Common set</td><td rowspan="6">8 or 10</td><td rowspan="2">15 random</td><td rowspan="2">Case 1</td><td rowspan="4">1 to 5</td><td rowspan="4">1 or 3</td></tr>
<tr><td>2</td><td>Var. 1</td></tr>
<tr><td>3</td><td>Var. 2</td><td rowspan="4">9 random and 6 skilled</td><td rowspan="4">Case 2</td></tr>
<tr><td>4</td><td>Var. 2</td></tr>
<tr><td>5</td><td>Var. 1</td></tr>
<tr><td>6</td><td>Common set</td></tr>
</table>

All the 30 NN models are evaluated by 10-fold cross validation for each user and the best performed "optimal" NN model is selected together with its parameters: number of hidden neurons, type of signature forgeries for training and input features. All the 12 kNN models are evaluated by LOOCV and the best performed "optimal" kNN model is selected together with its parameters: value of k, type of signature forgeries for training and input features.

The average estimated accuracy of the "optimal" NN models for all users is equal to 97.95%, and the average estimated accuracy of the "optimal" kNN models for all users is equal to 96.13%. These results demonstrate the advantage of the NN classifier over the kNN classifier. The value of t-statistics is equal to 3.29 and this value is significant for a probability level equal to 0.99 and DF=176.

Dependence of the verification accuracy on the feature subset used and on the forgery signature types used for classifier training

The following research aims (1) to estimate the impact of feature set reduction on the classification accuracy, and (2) to investigate the dependenceof the classification accuracy on the forgery signature types used for classifier training.

Table 4 Average estimated accuracy for all the "optimal" models for NN

	Common feature set	***Var. 1***	***Var. 2***	**Average accuracy**	**Number of occurances**
Case 1	**97.36** *(Model #1, 13 occurances)*	**98.36** *(Model #2, 27 occurances)*	**97.85** *(Model #3, 13 occurances)*	**97.99**	**53**
Case 2	**97.71** *(Model #6, 11 occurances)*	**97.98** *(Model #5, 9 occurances)*	**97.96** *(Model #4, 16 occurances)*	**97.89**	**36**
Average accuracy	**97.52**	**98.27**	**97.91**	**-**	**-**
Number of occurances	**24**	**36**	**29**	**-**	**89**

Table 5 Average estimated accuracy for all the "optimal" models for kNN

	Common feature set	***Var. 1***	***Var. 2***	**Average accuracy**	**Number of occurances**
Case 1	**100** *(Model #1, 5 occurances)*	**95.92** *(Model #2, 27 occurances)*	**97.84** *(Model #3, 15 occurances)*	**96.97**	**47**
Case 2	**84** *(Model #6, 1 occurances)*	**95.5** *(Model #5, 20 occurances)*	**95.45** *(Model #4, 21 occurances)*	**95.20**	**42**
Average accuracy	**97.33**	**95.74**	**96.45**	**-**	**-**
Number of occurances	**6**	**47**	**36**	**-**	**89**

The average estimated accuracy for all the "optimal" models are presented in Table 4 for NN and in Table 5 for kNN.

The data presented in Table 4 reveals higher NN accuracy if using individual feature subsets (*Var. 1* and *Var. 2*) compared to the accuracy if using a common feature subset. The accuracy is slightly higher if *Var. 1* is used. The accuracy of models built on *Var. 1*, *Case 1* (random forgeries for training) suppresses the accuracy of models built on *Var. 2*, *Case 2* (random and skilled forgeries for training).

The data presented in Table 5 reveals higher *k*NN accuracy if using a common feature subset compared to the accuracy if using individual feature subsets (*Var. 1* and *Var. 2*). The accuracy of models built on *Var. 2*, *Case 1* (random forgeries for training) suppresses the accuracy of models built on *Var. 2*, *Case 2* (random and skilled forgeries for training).

The comparison between the average estimated accuracy of NN and *k*NN based on common and individual feature subsets demonstrate the advantage of NN.

The number of "optimal" models, trained on random forgeries (*Case 1*) is met in more than the half of the users. It is equal to 53 (for 60% of the users) for NN and it is equal to 47 (for 53% of the users) for *k*NN. The most frequent "optimal" models are built on *Var.1* features: for NN, it is equal to 36, and for *k*NN it is equal to 47.

Building of NN and *k*NN classifiers on the selected models for each user

NN and *k*NN classifiers are built for each user. They are evaluated on the same training (~70% of the signatures) and testing sets (~30% of the signatures). The following results are obtained: 1) for NN: average accuracy 98.46 %, 2.70 % FAR и 0 % FRR NN; 2) for *k*NN: average accuracy 89.47 %, 8.09 % FAR и 14.61 % FRR. These results demonstrate the aadvantage of NN over the kNN classifier. The value of t-statistic is equal to 5.98 and this value is significant for a probability level of 0.99.

For comparison, a classifier of *Yanikoglu* и *Kholmatov* [18] has 1.64% FRR и 1.28% FAR on the same signature database.

Software application

The proposed combined method for on-line signature recognition is implemented in a software system prototype which is built in Microsoft Visual Studio 2008 Express Edition. Tablet PC SDK 1.7 [15] is used to facilitate signature acquisition. Raw and transformed signature data is stored in a database in SQL Server Compact Edition 2008.

A signature acquisition protocol is developed do describe the process of signature acquisition with the software application.

Results interpretation

The following conclusions can be drawn from the presented experiments:

1. The number of the features is reduced by around two times by consecutive applying of the method of correlation pleiads and Mallows Cp criterion for selection of regression variables;
2. There is not a common feature subset valid for all users; there is a specific feature subset for each user which describes his signature writing style. This subset consists of 3-5 features for some users;
3. Using of random forgeries as negative cases (*Var.1*) in regression model drives to greater reduction in feature number. Initial feature set size is reduced to a higher

extend if random forgeries (*Var. 1*) are used for building the regression model for the Hocking, Leslie and LaMotte method instead of skilled forgeries;

4. If a reduced number of features is used, the verification accuracy increases for NNs and decreases for kNN;
5. The classifiers trained on only random forgeries (*Case 1*) gives better verification results than those trained on both random and skilled forgeries (*Case 2*).

The following recommendations based on the above mentioned conclusions can be formulated for using of the proposed method for on –line signature verification:

1) Signature features number to be reduced by consecutive applying of the method of correlation pleiads and Mallows Cp criterion for selection of regression variables;
2) NN classifiers to be trained on random forgeries and the most accurate of them to be selected for each user.

Conclusions

In the PhD thesis a combined method for on-line signature recognition is proposed. The method is implemented in a software application. Signature verification is considered since its widespread application in many areas. All the necessary steps in developing a signature recognition system are described: signature data pre-processing, feature extraction and selection, verification and system evaluation. The influence of the signature forgery type (random and skilled) over the feature selection and verification is investigated as well. Experiments are carried out on an own signature database which consists of genuine and forgery signatures of 8 users and on SUsig database which consists of genuine and forgery signatures of 89 users. The average accuracy is equal to 98. 46%. The results obtained by applying two different classifiers (NN and *k*-nearest neighbours) are compared. In this way, all the tasks set in the thesis, are completely fulfilled.

Research and investigation in on-line signature recognition are about to continue in the future. They will be towards 1) improvement of the proposed system – its automation and accuracy; 2) testing of the system over different databases and other signature features; 3) forgery detection. In this regard, adding a camera to give an account on the characteristics of the hand in the time of signature acquisition will be of use.

Contribution summary

The contribution of the thesis can be summarized as follows.

1. An approach to feature subset selection as a consecutive application of the correlation pleiads method and regression analyses is proposed. This approach is applied to all users of the signature database.
2. A method for classifier generation as a combination of individual user classifiers is proposed.
3. A generalized net model of the on-line signature verification process is proposed.
4. A comparative investigation of the verification accuracy depending on the forgery signatures used for training is performed.
5. A comparative investigation of the verification accuracy depending on the signature feature subset (common or individual) is performed.
6. A signature database and a signature acquisition protocol are developed.

7. A software system prototype is developed. It has the following functionality: adding of a new user, searching of an existing user, signature acquisition with a graphical tablet, signature data preprocessing, feature extraction, feature selection, genuine and forgery signatures visualization, classifier selection and training.
8. A terminology dictionary for pattern recognition in Bulgarian and English is created.

Bibliography

1. Nalwa, V.S., Ekeland, I. *Automatic on-line signature verification*. - Proceedings of the IEEE 85, 1997, pp.213-239.
2. Jain A., Stan Li. *Encyclopedia of Biometrics*, Springer, 2009.
3. Gluhchev G., M. Savov, O. Boumbarov, D. Vassileva. *A New Approach to Signature Based Authentication*, - 2nd Int. Conf. on Biometrics, Seoul, 26-29 August, 2007, pp. 594-603.
4. Savov M., G. Gluhchev. *Signature verification via Hand-Pen motion investigation*. - Proc. Int. Conf. "Recent Advances in Soft Computing", Canterbury, 2006, pp. 490-495.
5. Kholmatov, A., Yanikoglu B. *SUSIG: an on-line signature database, associated protocols and benchmark results*. - Pattern Analysis & Applications, Vol. 12, 2009, pp.227-236.
6. Plamondon, R., Lorette, G.: *Automatic signature verification and writer identification – the state of the art*. Pattern Recognition 22, pp. 107–131, 1989.
7. Ink Data: http://msdn.microsoft.com/en-us/library/ms811395.aspx
8. Richiardi, J., Ketabdar, H., Drygajlo A. In: *Local and Global Feature Selection for On-line Signature Verification*. - Eighth International Conference on Document Analysis and Recognition (ICDAR'05), 2005, pp. 625-629.
9. F. Leclerc, R. Plamondon. *Automatic signature verification: the state of the art 1989-1993*. - International Journal of Pattern Recognition and Artificial Intelligence, 8(3):643-660, 1994.
10. Hocking, R.R., Leslie, R. *Selection of the Best Subset in Regression Analysis*. – Technometrics 9., 1967, pp. 531-540.
11. LaMotte, L.R., Hocking, R.R.: *Computational Efficiency in the Selection of Regression Variables*. - Technometrics 12., 1970, pp. 83-93.
12. .С.А. Айвазян, З.И. Бежаева, О.В. Староверов, *Классификация многомерных наблюдений*, Москва, Статистика, 240 стр., 1974.
13. McCabe, A., Trevathan, J., Read, W. *Neural network-based handwritten signature verification*. - Journal of Computers, 3 (8). pp. 9-22, 2008.
14. P. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
15. Tablet PC SDK 1.7 : http://www.microsoft.com/download/en/details.aspx?displaylang=en&id=20039
16. McCabe, Alan, Trevathan, Jarrod, and Read, Wayne. *Neural network-based handwritten signature verification*. Journal of Computers, 3 (8). pp. 9-22, 2008.
17. Berrin A. Yanikoglu, Alisher Kholmatov: *Online Signature Verification Using Fourier Descriptors*. EURASIP J. Adv. Sig. Proc., 2009.
18. Kholmatov A., YanikogluB.. *Identity authentication using improved online signature verification method*. Pattern Recognition Letters, 26(15):2400–2408, 2005.

БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ

АВТОРЕФЕРАТ НА ДИСЕРТАЦИЯ

за присъждане на образователна и научна степен "доктор" по научна специалност 01.01.12 „Информатика"

КОМБИНИРАН ПОДХОД ЗА РАЗПОЗНАВАНЕ НА ОН-ЛАЙН ПОДПИСИ

Десислава Николова Бояджиева

Ръководител: доц. Георги Глухчев

Научно жури:

Проф. Стефан Хаджитодоров
Чл.-кор. Красимир Атанасов
Доц. Георги Глухчев
Доц. Димо Димов
Доц. Пенчо Венков

IICT
Институт по информационни и комуникационни технологии
Секция "Обработка на сигнали и разпознаване на образи"

Дисертацията е обсъдена и допусната до защита на разширено заседание на секция „Обработка на сигнали и разпознаване на образи “ на ИИКТ-БАН, състояло се на 17.01.2014 г.

Дисертацията съдържа 129 стр., в които 18 фигури, 19 таблици и 7 стр. литература, включваща 108 заглавия.

Защитата на дисертацията ще се състои на 02.06.2014 г. от 10:00 часа в зала 507 на блок 2 на ИИКТ-БАН на открито заседание на научно жури в състав:

Материалите за защитата са на разположение на интересуващите се в ИИКТ-БАН, ул. „Акад. Г. Бончев“, бл. 25А, стая 215.

Обща характеристика на дисертацията

Актуалност на темата

Верификацията и идентификацията по подпис, обединени от общия термин разпознаване по подпис, са задачи от областта на разпознаването на образи и анализа на документи. Целта на верификацията е потвърждаването или отхвърлянето на заявена самоличност, а на идентификацията - установяването на самоличността на даден потребител. В зависимост от начина на събиране на подписи, системите за разпознаване по подпис са статични (оф-лайн) и динамични (он-лайн). При първите подписът се привежда в цифров вид чрез сканиране или заснемане с фотокамера след като вече е бил положен върху хартия и по-нататък се работи с неговото полутоново статично изображение, докато при вторите подписите се полагат директно чрез дигитална писалка върху устройства като графичен таблет или друг вид сензорен екран, като по този начин автоматично се получава едновременно и статична, и динамична информация за подписа.

Тъй като подписът е често употребявано, удобно, сигурно и широко възприето средство за установяване на самоличност, с неинвазивно, бързо, лесно и естествено за потребителите въвеждане в биометричните системи, то през последните години се наблюдава засилен интерес към разработките в областта. Разработват се нови методи и алгоритми, част от които се внедряват в практически приложения [7, 13, 33, 40, 41, 42], като се наблюдава превес на он-лайн пред оф-лайн методите за разпознаване по подпис. С изобретяването на все повече и повече нови устройства с дигитални писалки и сензорни екрани, които предоставят възможност още с полагане на подписа да се разполага с неговия дигитален формат, се откриват нови хоризонти пред областите на приложение на он-лайн системи за разпознаване по подпис. Така практически всяко приложение за достъп до система, ресурс или устройство , за което се изисква парола за достъп, може да се замени с он-лайн верификация по подпис, а он-лайн идентификацията по подпис успешно да се прилага в случаи на хартиен носител. Изобщо, широките области на приложение в създаването и внедряването на системи за установяване на самоличност определят актуалността на задачата в световен мащаб, а оттам и актуалността на дисертационната работа.

Цели и задачи на дисертацията

Основната цел на дисертационния труд е ***Разработване на нов комбиниран метод за разпознаване на потребители по он-лайн подписи и реализирането му като софтуерна система.*** За изпълнение на поставената цел е необходимо да се решат следните задачи:

- Организиране на база данни от подписи;
- Извличане на известни и нови признаци;
- Избор на оптимални подмножества от признаци;
- Изследване на различни модели на невронни мрежи за верификация като елементи на комбиниран класификатор;
- Провеждане на експерименти за оценка на точността на класификация и сравнение с известни класификатори с използване както на собствена база данни за подписи, така и с публично достъпна такава;
- Разработка на софтуерно приложение за разпознаване на подписи с

основни функции: събиране на собствени и подправени подписи, визуализация на подписите, извличане на признаци, избор на признаци и верификация по тях.

Особеностите на задачата налагат определени ограничения, съобразяването с които прави някои от стандартните подходи не особено подходящи. Тези особености се отнасят преди всичко към количеството на експерименталните данни, използвани за обучение и тестване на класификатора и характера на използваните признаци. За разлика от други задачи тук не е възможно събирането на голям брой образци от подписи на евентуалните потребители. Изискването за многократно подписване е неприятно. Броят на желаещите намалява значително и ако се иска неколкократно повторение на сесиите. Добавяйки и нежеланието на повечето хора да предоставят подписите си за експерименти, страхувайки се от злоупотреба, броят на събраните подписи се свежда до няколко десетки. Сравнително по-големи бази от подписи могат да бъдат събрани в университетите. Още по-голямо затруднение представлява събирането на умели фалшификати, необходими за повишаване на точността на класификатора.

Изборът на признаци също не е тривиален по две причини. Първата е свързана с малкия брой опитни данни. От теорията на разпознаването на образи е известно, че при фиксиран брой опитни данни увеличаването на броя на признаците не само не води до повишаване на точността на разпознаването, но я и намалява. Това означава, че ако при малък брой данни имаме голям брой признаци, те трябва да се намалят, т.е. да се избере подмножество, което да бъде достатъчно информативно. Намаляването на броя на признаците е свързано и с намаляване на изчислителното време. Втората причина при избора им е необходимостта от разбираем физически смисъл. Това се налага при съдебни експертизи, където резултатът от машината не се приема от съда за доказателство, ако зад него не застане експерт. (Обяснението е, че от машината не може да бъде търсена отговорност).

Независимо от тези сериозни практически ограничения, към всеки класификатор се предявяват изисквания за висока точност на разпознаване, изразяваща се в малък процент на двата вида грешки: дял на лъжливо положителните образци (FAR) и дял на лъжливо отрицателните образци (FRR). Допълнителна особеност в това отношение е различната цена на двете грешки, която трябва да бъде отчетена от класификатора. Много по-сериозна е първата грешка, която може да доведе до тежки последици за институцията, докато при втората ще бъде необходимо само полагането на допълнителен подпис.

Заедно с формулираните цел и задачи на дисертацията ще бъдат потърсени отговори и на следните въпроси.

1) Съществува ли конкретно, специфично подмножество от признаци за всеки потребител, отразяващо особеностите на неговия стил на подписване?

2) Има ли съществена разлика между резултатите, получени при верификацията с общо множество от признаци и индивидуално подмножество за всеки участник?

3) Ще се понижи ли точността на верификация с използването на умели фалшификати в качеството на отрицателни примери?

Методология на изследването

Методологията на настоящите изследвания се основава на методи и подходи от следните области:

- Статистика – корелационен анализ, избор на регресионни променливи, крос

валидация;
- Информатика - невронни мрежи, метод на *k*-най-близки съседи;
- Разпознаване на образи – схеми за верификация.

В дисертационния труд за тестване на предложеният комбиниран метод за разпознаване на он-лайн подписи са използвани образци на подписи от собствена база данни и от публично достъпна база от данни.

Апробация на резултатите

Част от резултатите са докладвани на 4 международни конференции, 2 от които в чужбина (BioID_MultiComm'09, PReMI'11) и 2 в България (UNITECH'10, 14th Int. Workshop on Generalized Nets). 1 статия е докладвана на научна конференция в България и една е приета за печат в списанието *Cybernetics and Information Technologies*.

Списък на публикациите по дисертацията

1. Boyadzhieva D., Gluhchev G., Feature Set Selection for On-line Signatures using Selection of Regression Variables, In: Proceedings of the 4th International Conference on Pattern Recognition and Machine Intelligence PReMI'11, June 27-July 01 2011, Moscow, Russia, LNCS 6744, Springer-Verlag Berlin Heidelberg, 2011, ISSN 0302-9743, pp. 440-445, http://link.springer.com/chapter/10.1007/978-3-642-21786-9_71#page-1.
2. Boyadzhieva D., Gluhchev G., An approach to feature selection for on-line signatures, In: Proceedings of the 10th Anniversary International Scientific Conference UNITECH'10 Vol. I, Nov. 19-20, 2010, Gabrovo, Bulgaria, ISSN 1313-230X, pp. I-457…I-460.
3. Бояджиева Д., Извличане на признаци на подпис от графичен таблет, В: Научни трудове на Русенски университет "Ангел Кънчев", 2010, том 49, серия 3.2, Русе, България, ISBN 1311-3321, стр. 101-105, http://conf.uni-ruse.bg/bg/docs/cp10/3.2/3.2-18.pdf.
4. Dimitrova D., Gluhchev G., Pressure Evaluation in On-Line and Off-Line Signatures, In: Proceedings of the Joint COST 2101 & 2102 International Conference on Biometric ID Management and Multimodal Communication (BioID_MultiComm'09), Sep. 16-18 2009, Madrid, Spain, LNCS 5707, Springer -Verlag Berlin Heidelberg, 2009, ISSN 0302-9743, pp. 207-211, http://link.springer.com/chapter/10.1007/978-3-642-04391-8_27#page-1.
5. Desislava Boyadzhieva and Georgi Gluhchev., *A GN model for on-line signature verification*, In: Proceedings of the 14th Int. Workshop on Generalized Nets, Burgas, Bulgaria, 29-30 November 2013, ISSN 1313-6860, pp. 65-70, http://www.ifigenia.org/w/images/f/f6/IWGN2013-14-65-70.pdf
6. Boyadzhieva D., Gluhchev G., *A Combined Method for On-Line Signature Verification* (приета за печат в списанието *Cybernetics and Information Technologies,* кн. 2, 2014 г).

Съдържание на дисертацията

Настоящата дисертация се състои от списък на фигурите, списък на таблиците, речник на термините, увод, 5 глави, заключение, приноси, публикации по темата и библиографска справка. Първа глава е озаглавена "Разпознаване на подписи", втора глава - " Избор на признаци за он-лайн подписи", трета глава - "Класификация", четвърта – "Експерименти и резултати" и пета – "Софтуерно приложение". Основното съдържание е изложено на 129 страници. Включени са 42 формули, 18 фигури, 19 таблици и 7 стр. библиографска справка, включваща 108 заглавия.

Глава 1. Разпознаване на подписи

Първа глава включва следните раздели.

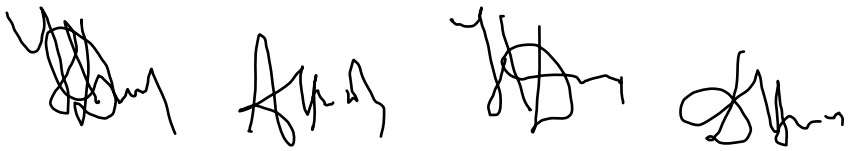
1.1 Постановка на задачата за разпознаване на подписи

1.1.1 Верификация и идентификация на самоличност
1.1.2 Биометрични данни и подписът като биометрични данни
1.1.3 Системи за разпознаване по подпис (СРП)

Разпознаването по подпис представлява процеса по потвърждаване на самоличността на потребител въз основа на неговия подпис като поведенческа биометрична характеристика [33] и е област на изследване в областта на разпознаването на образи и анализа на документи. В системите за разпознаване по подпис се фокусира върху извличането на характерните и специфични характеристики от образците на подписите за всеки потребител, които го разграничават възможно най-много от подписите на останалите потребители. Подписът като биометрична модалност по-често намира приложение в системи за достъп до определен ресурс и в системи за идентификация на базата на хартиен носител.

1.1.4 Он-лайн и оф-лайн СРП
1.1.5 Приложение на он-лайн СРП
1.1.6 Видове подправени подписи в он-лайн СРП

За разлика от някои биометрични модалности, подписът може да бъде подправен с относителна лекота. Разграничават се следните типове подправени подписи: неумели (*random*), умели (*skilled*) и прости (*simple*). При първите фалшификаторът използва собственият си подпис вместо този, който ще подправи, докато при вторите фалшификаторът е запознат с собствения подпис и динамиката на изписването му. При простите фалшификати пък, имитаторът изхождайки от името на потребителя, чийто подпис ще подправи, създава подправения подпис. На Фигура 1.1 са представени собственият и трите типа подправени подписи за потребител.



Фигура 1.1 Собствен и три типа подправени подписа (неумел, умел и прост) за потребител

Степента на сходство на подправените подписи с собствените подписи зависи от протокола за събиране на подписи, от мотивацията на фалшификаторите, от техните умения да подправят подписи, от това доколко са запознати с подписите и по какъв начин се упражняват, както и от времето, с което разполагат за трениране. Събирането на подправени подписи от хора, които нямат опит във фалшифицирането на подписи, както се случва на практика, е причина за отсъствието на качествени фалшификати. Фалшифицирането на динамиката при подписване е значително по-трудно отколкото имитирането единствено на формата на подписа, тъй като на практика фалшификаторът рядко има достъп до динамичните характеристики на подписа.

1.1.7 Етапи в създаването на он-лайн СРП

Създаването на СРП (он-лайн или оф-лайн) преминава през следните етапи: събиране на подписи, предварителна обработка, извличане на признаци и избор на най-информативните сред тях, разработване на класифициращи правила и оценяване на вероятността за грешка при класификацията [1].

1) *Събиране на подписи* – Събирането на он-лайн подписи се осъществява чрез устройства, например графични таблети, PDA, Pocket PC, Tablet PC, умни телефони и др. Най-често с тях се измерват следните характеристики: а) *x*, *y* координатите на върха на писалката; б) приложеният от писалката натиск; в) времеви отпечатък; г) ъгъл на азимут на писалката (ротация на писалката по часовниковата стрелка около оста x, със стойности между 0° и 359°); д) ъгъл на наклон на писалката по отношение на повърхността на графичния таблет (със стойности между 0° и 90°).
2) *Предварителна обработка* – това е обработката, която се извършва, за да се подготвят подписите за следващия етап. Тя обикновено включва филтриране, за премахване на точките, непринадлежащи на подписа, транслация на подписа в дадена точка, ротация, целяща подравняването на подписа хоризонтално, а понякога и нормиране, за да се стандартизира подписа по отношение на размера.
3) *Извличане на признаци* – Те се пресмятат като функции от извлечените характеристики. Признаците би трябвало да позволят разграничаването на истинските от подправените подписи и да са подходящи за автоматично сравнение. Намирането на добро дискриминационно множество от признаци предопределя до голяма степен успеха на всяка система за разпознаване.
4) *Класификация* - Основната задача в областта на разпознаването на образи е построяването но класификатор. При нея се извършва причисляване на входните образци към един от множество от *N* класа.
5) *Оценка на вероятността за грешка* - Качеството на биометричните системи се оценява в термините на оценки на грешката при вземане на решение (*decision error rates*). В качеството на оценки се използват FAR (*false accept rate*, дял на лъжливо положителните образци) и FRR (*false reject rate*, дял на лъжливо отрицателните образци).

1.1.8 Извличане на признаци и подходи за разпознаване по подписи

Признаците на подписи представляват функции от извлечените характеристики, а изборът на оптимално признаково множество е важна стъпка в създаването на СРП. От първостепенно значение е извлечените признаци да имат невисоки изисквания към изчислителни ресурси и да водят до добро разпознаване с ниска стойност на FRR и FAR. Различните СРП се различават по използваните признаци и метода на разпознаване. Повече от 100 признаци са използвани в разработките в областта [2]. В зависимост от критерия, по който се групират, признаците за подписи се делят на [5]:

А) Според начина на събиране на подписите: статични и динамични;

Б) Според обхвата: глобални и локални;

В) Според начина на представяне: функционални и параметрични (скаларни).

1.1.9 Трудности при създаването на СРП

Подписът има някои характерни особености, които създават известни затруднения при създаването на надеждна СРП. На първо място следва да се отбележи, че представителите на подписите от един и същи потребител са подобни, почти еднакви, но не и идентични (съществува голяма вътре класова вариация), както и че

последователно положени подписи от един и същ потребител се различават по мащаб и по ориентация. Върху качеството на подписите оказват влияние и различните условия, при които те се полагат. Например набързо положения подпис в изправена поза би се различавал от подписа, положен докато участникът седи. Значение би оказала и употребата на необичайна писалка, с която потребителят не е свикнал, както и моментното психологическо състояние на участника. Трудностите при разпознаването произтичат от късата дължина на някои подписи, която би могла да разкрие малка информация, а оттам да доведе до неточното им разпознаване. Също така, потребители с подобни или еднакви имена би могло да имат сходни подписи по геометрична форма [5].

Трудности произтичат и от процеса по набавянето на собствени и подправени подписи. Обикновено СРП разполага с ограничен брой подписи. Създаването на тестова база от данни от подписи, която да е представителна за реални приложения е тежка задача, поради трудностите, които обичайно се срещат при набирането на доброволци, които да се съгласят да представят известен брой подписи и то в различни сесии. Повечето от хората не биха се съгласили подписите им да се съхраняват в бази от данни и да се предоставят за фалшифициране.

1.2 Критичен обзор

В тази част е извършен преглед на разработките в областта на разпознаването на подписи от последните години. Обърнато е внимание на обзорните статии по темата и са разгледани различни подходи за решаване на задачата, свързани с настоящата дисертация. Тези методи са дискутирани и са представени конкретни системи за разпознаване по подпис, проектирани върху тях. За всяка от системите са посочени използваните видове признаци и метода за разпознаване, вида на използваните фалшифицирани подписи, базата от данни, върху която са проведени експериментите и получената оценка на системата.

През последните 30 години са разработени голям брой методи и модели за верификация на он-лайн подписи, които се основават най-общо на следните методи: статистически подходи, динамично времево изкривяване (*Dynamic Time Warping*), скрит Марковски модел (Hidden *Markov Model*), невронни мрежи, комбинирани подходи [8] и класификатор по *k*-най-близки съседи.

На база на направения литературен обзор може да се направи извода, че текущите изследвания и разработки в областта на разпознаването на подписи са насочени основно към създаването на комбинирани подходи и по-добър избор на признаци, които да разграничават надеждно собствените от фалшифицираните подписи.

1.3 Цели на дисертационната работа

1.4 Резултати в Глава 1

1. Представена е биометричната модалност „подпис“, начините за получаване и възможните приложения.
2. Направена е характеристика на използваните признаци, етапите на изграждане на СРП и съпътстващите ги трудности.
3. Представени са основните методи за разпознаване: статистически, динамично времево изкривяване, скрити Марковски вериги, невронни мрежи, комбинирани подходи. Посочени са съответните литературни източници и данни за ефективността на методите.
4. Формулирана е целта на дисертацията и съпътстващите изпълнението й задачи.

Глава 2. Избор на признаци за он-лайн подписи

.

Втора глава се състои от следните раздели.

2.1 Предварителна обработка

Една от основните задачи в разпознаването на образи е изборът на признаци, по които да се извършва разпознаването. Те се избират от изследователя в зависимост от неговия опит, а също и от спецификата на разпознаваните обекти. Предварително не е известно кои са признаците, които биха довели до задоволително разпознаване и затова обикновено изследователят разглежда възможно най-голям брой признаци, стремейки се по този начин да не пропусне някой важен за разпознаването признак. Същевременно, ефективността на системата за разпознаване и изискванията към изчислителните ресурси зависят в голяма степен от броя на признаците [34]. Затова колкото е по-малък той, толкова е по-голяма скоростта на верификация [6].

Така възниква и задачата за избор на най-информативните признаци, решаването на която се състои в минимизиране на признаковото пространство, така че да се премахнат признаците, които са носители на излишна или повтаряща се информация.

Бихме могли да формулираме следните критерии към признаците, които да участват в признаковото множество за он-лайн подписи:

1. Да бъдат извличани и пресмятани бързо и лесно;
2. Ако по някакъв начин станат достъпни от неоторизирани лица, да не позволяват лесното пресъздаване на подписа чрез тях;
3. Броят им да бъде достатъчно малък, за да не се забавят по-нататъшните изчисления и достатъчно голям, за да осигури надеждна верификация;
4. Желателно е да имат разбираем физически смисъл, което ще позволи установяването на фалшиви подписи от експерт в случай на съдебна експертиза.

Наред с избора на модел и вида на класификатора избраните признаци играят много важна роля за постигането на успех на всяка СРП.

2.1.1 Предварителна обработка на данните от подпис

Както е известно, според вида на използвания хардуер (графичен таблет, PDA и др.) за събиране на подписи, се измерват различни характеристики за всяка точка от траекторията. Данните за тези характеристики са в "суров" вид. С прилагането на някои операции към тях (т.нар. предварителна обработка), стойностите на данните се трансформират, за да може да се улесни извличането на признаците от тях.

Координатите x и y от координатната система на таблета (*ink coordinate space*) се наричат *himetric units* [28] и приемат високи стойности в [0,7999]x[0,5999]. Първата извършвана операция е *(1) трансформиране* на тези координати към x и y координатите (нар. пикселови координати) от координатната система на приложението със стойности в [0,1279]x[0,799]. Това се извършва автоматично от метод, достъпен от пакета за разработка на приложения Microsoft Tablet PC. При събирането на подписи с графичен таблет има възможност подписите от даден участник да не са подравнени по отношение на оста x, а да са завъртени на определен ъгъл. Това от своя страна налага нуждата от извършването на *(2) ротация*. Другата извършвана операция е *(3) транслация* на подписите от всички участници в избрана точка от координатната система на приложението.

Ротация на подпис на определен ъгъл

За извършването на ротацията на подпис се прилага подхода на собствените вектори. Първоначално се намират главните оси на подписа, съответстващи на собствените вектори v_1 и v_2 на ковариационната матрица Σ_{xy}. Пресмятат се и съответните собствени стойности λ_1 и λ_2. Максималната собствена стойност $\lambda_{\max} = \max(\lambda_1, \lambda_2)$ показва оста на най-голямата промяна, а ъгъла θ, който подписа сключва с оста x може да се намери от компонентите на съответният й собствен вектор $v_{\max}$.

$$\theta = arctg(\frac{v_{\max 2}}{v_{\max 1}}) \tag{2.1}$$

За даден подпис S с n на брой точки, представен във вида:

$$S = \{S_i(x_i, y_i), i = 1,...,n\} \tag{2.2}$$

завъртането на ъгъл θ (докато главната ос съвпадне с тази на оста x) се задава със следното матрично уравнение:

$$\begin{pmatrix} x'_i \\ y'_i \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix}, i = 1,...,n \tag{2.3}$$

Траекторията от точки $S' = \{S'_i = (x'_i, y'_i), i = 1,...,n\}$ представя подписа след ротация.

Транслация на подпис в избрана начална точка

За даден подпис S с n на брой точки, представен във вида (2.2) и избрана начална точка с координати (x_o, y_o) първоначално се намират съответните компоненти на транслация T_x и T_y:

$$\begin{pmatrix} T_x \\ T_y \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \tag{2.4}$$

след което се извършва самата транслация, като за всяка двойка координати (x_i, y_i) от стойностите на x_i и y_i се извадят съответно стойностите T_x и T_y:

$$\begin{pmatrix} x_t^{Tr} \\ y_t^{Tr} \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} T_x \\ T_y \end{pmatrix}, i = 1,...,n \tag{2.5}$$

В резултат на трансформацията, траекторията от точки $S^{Tr} = \{S_i^{Tr}(x_i^{Tr}, y_i^{Tr}), i = 1,...,n\}$ представлява подписа след транслация.

2.2 Извличане на признаци на он-лайн подпис от графичен таблет

С помощта на пакета за разработка на приложения *Tablet PC SDK 1.7* се получава динамична информация за събираните с графичния таблет подписи. Тази информация е структурирана в пакети (по един за всяка точка от подписа) и съдържа множество от

данни, което се поддържа от съответния графичен таблет. Някои от тези данни са следните: *X* и *Y* координати, натиск, наклон на писалката, времеви отпечатък (*timestamp*) и азимут. Координатите определят местоположението на върха на писеца върху координатната система на таблета, натискът приема стойности в интервала [0,1023] и представлява приложения от писалката натиск в съответната точка. Наклонът на писалката е ъгъла, който тя сключва с повърхността на таблета и приема стойности в интервала [0,90], а азимутът е ротацията по часовниковата стрелка на писалката в равнината XY и приема стойности в интервала [0,360].

Таблица 2.1 Глобални признаци

№	Признак	№	Признак
A1	Дължина на подписа *L*	A14	Разстояние от най-лявата точка до центъра
A2	Височина *H*	A15	Разстояние от центъра до най-дясната точка
A3	Сплеснатост *H/L*	A16	Ъгъл на отсечката, определена от центъра и най-лявата точка
A4	Брой точки *N*	A17	Ъгъл на отсечката, определена от центъра и най-дясната точка
A5	Общо време	A18	Разстояние от най-лявата точка до началната
A6	Брой сегменти	A19	Разстояние от най-дясната точка до крайната
A7	Плътност A4/A1*A2	A20	Ъгъл на отсечката, определена от най-лявата точка и началната
A8	Разстояние от началната точка до центъра	A21	Ъгъл на отсечката, определена от крайната и най-дясната точка
A9	Разстояние от крайната точка до центъра	A22	Брой щрихи
A10	Разстояние от началната до крайната точка	A23	Наклон на подписа
A11	Ъгъл на отсечката, определена от центъра и началната точка	A24	Средна стойност на натиск
A12	Ъгъл на отсечката, определена от центъра и крайната точка	A25	Средна стойност на наклон на писалката
A13	Ъгъл на отсечката, определена от началната и крайната точка	A26	Средна стойност на азимут

Стойностите на признаците за даден подпис се изчисляват въз основа на неговата динамичната информация. В Таблица 2.1 са представени използваните глобални признаци, като във формулите им са използвани следните означения.

Нека с *Si* сме означили *i*-тата точка от даден подпис, с *n* – броя на точките в подписа, (x_i, y_i) са координатите на точката *Si*. Натискът в точката *Si* означаваме с p_i ,с s_i – наклонът в *i*-тата точка и с t_i – изминалото време между изчертаването на *i*-тата и (*i+1*)-вата точка. С $(\bar{x}, \bar{y})$ сме означили координати на центъра, с (x_{lev}, y_{lev}) и с (x_{desen}, y_{desen}) сме означили координати на най-лявата и най-дясната точка на подписа.

Центърът $(\bar{x}, \bar{y})$ на подпис, съставен от *n* на брой точки се задава с формулата:

$$(\bar{x}, \bar{y}) = (\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i). \quad (2.6)$$

Разстоянието d между две точки от подписа (x_1, y_1) и (x_2, y_2) се задава с:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{2.7}$$

Ъгълът φ на отсечката от точка (x_1, y_1) до точка (x_2, y_2) от подписа се задава с:

$$\varphi = arctg\left(\frac{(y_2 - y_1)}{(x_2 - x_1)}\right). \tag{2.8}$$

Една точка *Si* е локален екстремум, ако е изпълнено едно от условията (2.9):

$$\begin{aligned} y_{i-\varepsilon} < y_i > y_{i+\varepsilon} \\ y_{i-\varepsilon} > y_i < y_{i+\varepsilon} \\ x_{i-\varepsilon} < x_i > x_{i+\varepsilon} \\ x_{i-\varepsilon} > x_i < x_{i+\varepsilon} \end{aligned} \tag{2.9}$$

Големината на ε се определя експериментално, като може да бъде равна на 2,3..., но рядко се използва $\varepsilon = 5$. Всяка двойка последователни екстремуми разделя подписа на сегменти.

2.3 Минимизиране на броя на признаците

В този раздел е представен метод за избор на регресионни променливи, базиран на критерият *Cp* на *Mallows* [10]. Подборът на подмножество от независими променливи (фактори), които да се включат в регресионния модел, така че той да дава най-надеждна прогноза, е важна задача в регресионния анализ. Този проблем се нарича избор на най-добро подмножество или избор на най-добро регресионно уравнение [4].

2.3.1 Нормиране на признаците

Тъй като всеки от признаците получава стойности в различни интервали се налага те да се нормират. Чрез извършването на нормировка стойностите на данните попадат в обща скала на отчитане и се премахват единиците за измерване. Сред известните техники за нормиране са стандартизацията (нормировка *z-score*) и нормировката *min-max* [14]. В първия случай, стойностите на признаците се трансформират в такива със средна стойност, равна на нула и стандартно отклонение, равно на единица. Новополучената стойност f_i^{zscore} за признак f_i се намира по формулата:

$$f_i^{zscore} = \frac{f_i - \mu_{f_i}}{\sigma_{f_i}}, i = 1,...,k \tag{2.10}$$

където с μ_{f_i} и σ_{f_i} са означени съответно средната стойност и стандартното отклонение за *i*-тия признак, пресметнати за всички подписи от съответния потребител, изчислени по формулите:

$$\begin{aligned} \mu_{f_i} &= \frac{1}{n}\sum_{j=1}^{n} f_{ij} \\ \sigma_{f_i} &= \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(f_{ij} - \mu_{f_i})^2} \end{aligned} \tag{2.11}$$

Нормировката *min-max* в интервала $[c, d]$ се извършва за стойностите на всеки

признак по формулата:

$$f_i^{\min\max} = \frac{f_i - \min_{f_i}}{\max_{f_i} - \min_{f_i}}(d-c) + c, i = 1,...,k \tag{2.12}$$

където с min_{f_i} и max_{f_i} са означени съответно минимума и максимума за *i*-тия признак, пресметнати по всички подписи от съответния потребител.

$$\min_{f_i} = \min_j(f_{ij}) \tag{2.13}$$

$$\max_{f_i} = \max_j(f_{ij})$$

В задачите за разпознаване на образи нормировката най-често се извършва в интервала [0,1].

2.3.2 Премахване на един от група признаци с висока корелация

Корелацията е мярка за измерване както на посоката, така и на силата на връзката между две променливи [16]. Стойността на корелация се определя от стойността на корелационния коефициент. Най-често използван е корелационния коефициент r на Pearson, който за два p-мерни вектора $\mathbf{X_1}$ и $\mathbf{X_2}$ се определя по формулата:

$$r_{X_1,X_2} = \frac{\sum_{i=1}^{p}(X_{1i} - \mu_{X_1})(X_{2i} - \mu_{X_2})}{\sqrt{\sum_{i=1}^{p}(X_{1i} - \mu_{X_1})^2 \sum_{i=1}^{p}(X_{2i} - \mu_{X_2})^2}} \tag{2.14}$$

където с μ_{X_j} е означено средното за вектора $\mathbf{X_j}$, $j = 1,2$

$$\mu_{X_j} = \sum_{i=1}^{p} X_{ji}, j = 1,2 \tag{2.15}$$

Големината на корелационния коефициент е показател за силата на линейната връзка между векторите $\mathbf{X_1}$ и $\mathbf{X_2}$. Колкото той е по-близко до 1 или -1, толкова е по-силна връзката между двата вектора. Стойността на коефициента на корелация би могла в голяма степен да се повлияе от един или повече признаци, силно отдалечени от групата данни (*outliers*) [16].

Корелационната матрица на признаците с размер $k \times k$ може да се представи по следния начин за даден потребител $m = 1,...,M$

$$C_{k\times k}^m = \{C_{ij}^m\},\ i,j = 1,...,k \tag{2.16}$$

Всеки не диагонален елемент C_{ij}^m, $i \neq j$ на корелационната матрица представлява стойността на коефициента на корелация между признака i и признака j за съответния потребител m.

Методът на корелационните плеяди [30] се състои в намирането на групи от признаци с висока стойност на корелация между признаците в групата и ниска стойност на корелация между признаците от отделните групи. Тези групи признаци се наричат корелационни плеяди. С прилагането на този метод се намалява броя на признаците като от всяка група се оставя единствен признак. Няма теоретически обосновано становище кой признак от плеядата да бъде запазен. Най-простият и практически оправдан подход е да се избере този признак, получаването на който е най-бързо и

лесно. Значително по-убедително би било, ако за всяка комбинация от избрани признаци се оцени точността на класификацията. Това обаче изисква значителни изчислителни усилия, а резултатът от практическа гладна точка едва ли ще бъде съществен точно поради силната корелационна връзка в плеядата.

2.3.3 Избор на най-информативни признаци

Регресионният анализ е често използван метод за установяване на функционална зависимост между независимите променливи и зависимата променлива. Намира широко приложение в различни приложни задачи [4]. Изборът на регресионни променливи в регресионния анализ се състои в определянето на подмножество от променливи, които да се включат в модела, така че той да дава надеждна прогноза за принадлежността на дадена точка към коректния клас.

Формулировката на задачата за избор на регресионни променливи е следната [14]. Нека са дадени $n \geq k+1$ наблюдения над k-мерния вектор от независими променливи $\mathbf{x}_\mathbf{j}^t = (x_{1j},...,x_{kj})$ и зависимата променлива y_j, $j = 1,...,n$, стойността на която се определя от:

$$y_j = \sum_{i=1}^{k} \beta_i x_{ij} + e_j \tag{2.17}$$

Моделът на регресията (10) има следния матричен вид:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{e} \tag{2.18}$$

където $\mathbf{Y}$ е n-мерен вектор на зависимите променливи, $\mathbf{X}$ е матрица на независимите променливи с размерност $n \times k$, $\boldsymbol{\beta}$ е k-мерен вектор на неизвестните регресионни коефициенти и $\mathbf{e}$ е n-мерен вектор на остатъците, за които се предполага, че са независими и нормално разпределени $N(0,\sigma^2)$ с математическо очакване, равно на нула и дисперсия, равна на σ^2. Регресионните параметри (коефициенти) β_i, $i = 1,...,k$ са неизвестни константи, които следва да се оценят по метода на най-малките квадрати от изходните данни. Техните оценки се означават с b_i. Тогава, векторът $\mathbf{b}$ от оценките на вектора на коефициентите по метода на най-малките квадрати се задава чрез:

$$\mathbf{b} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y} \tag{2.19}$$

Най-популярният подход за избор на регресионен модел е този за "избор на най-добро подмножество" (*best subset*). При прилагането му се разглеждат всичките $2^k - 1$ на брой подмножества от регресионни променливи. Това на практика е неприложимо, особено ако броят на променливите е голям. В литературата съществуват методи, при които търсенето на подмножествата от променливи не включва пресмятания върху цялото признаково подмножество, а само върху част от него [15]. Към тях се отнася и разглеждания по-долу метод на Hocking и Leslie. Повечето от критериите за избор на регресионни променливи са монотонни функции на остатъчната сума от квадрати (RSS, *residual sum of squares*) и поради тази причина задачата за избор на регресионни променливи се редуцира до намирането на тези подмножества от променливи, за които остатъчната сума от квадрати е малка. Остатъчната сума от квадрати RSS_p за n наблюдения над p на брой променливи се дефинира като:

$$RSS_p = \sum_{j=1}^{n}(y_j - \sum_{i=1}^{p}\beta_i x_{ij})^2 \tag{2.20}$$

Mallows [17] предлага критерий за сравняване на няколко регресионни модела с цел откриването на най-добрите от тях. За целта се пресмята стойността на C_p за всеки един от моделите. За модел с p на брой променливи и n наблюдения, C_p се дефинира като:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p) \tag{2.21}$$

Оценката $\hat{\sigma}^2$ на σ^2 се задава по формулата:

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{j=1}^{n}(y_j - \sum_{i=1}^{k}\beta_i x_{ij})^2 = \frac{RSS_k}{n-k}. \tag{2.22}$$

където с σ^2 е означена дисперсията на n-мерния вектор на остатъците $\mathbf{e}$.

Така, ако даден модел с p на брой променливи е добър, то от (2.21) и (2.22) се намират следните стойности за математическото очакване на стойностите на RSS_p и C_p

$$\begin{aligned} E(RSS_p) &= (n-p)\hat{\sigma}^2 \\ E(C_p) &\approx p \end{aligned} \tag{2.24}$$

За добри подмножества от променливи се считат тези с малки стойности на C_p или със стойности на C_p, близки до p [17].

Hocking и Leslie [15] създават метод, който позволява откриването на най-добри подмножества от регресионни променливи след разглеждането на малка част от всичките $\binom{k}{p}$ възможни подмножества с размер p. LaMotte и Hocking [9] модифицират метода така, че сравнително големи по обем задачи могат да се решават с минимум изчисления. С т.нар. r-подмножество се означава подмножеството от променливи, които се изключват от регресионния модел, а с p-подмножество – подмножеството от променливи, които остават в модела. По същество методът се базира на т.нар. редукции на m-подмножества, т.е. на редукции в остатъчните суми от квадрати поради премахването на подмножества с размер $1 \le m \le 4$ от модела с k на брой променливи. Редукциите на m-подмножества служат за определянето на най-добро r-подмножество, което да се премахне (за $r > m$). Редукцията на регресионната сума от квадрати $\mathrm{Re}\,d_r$, получена от премахването на подмножество от r променливи се задава чрез:

$$\mathrm{Re}\,d_r = RSS_p - RSS_k \tag{2.25}$$

Множеството от r променливи, за което стойността на тази редукция е най-малка определя подмножеството от p ($p = k - r$) регресионни променливи, за което остатъчната сума от квадрати е минимална. В [15] е предложена следната формула за намиране на C_p чрез тази редукция:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (2p - k) \tag{2.26}$$

По-долу са изброени стъпките на обобщения алгоритъм [9].

1. Намират се регресионните коефициенти за модела с всичките k на брой променливи;

2. За всяка от променливите x_i се пресмята редукцията θ_i от премахването ѝ:

$$\theta_i = \hat{\sigma}^2 t_i^2 \tag{2.27}$$

където t_i е t-статистиката на i-тия регресионен коефициент.

Процедурата продължава с пренареждане на променливите по нарастване на съответните им редукции:

$$\theta_1 \leq \theta_2 \leq \cdots \leq \theta_k$$

Методът се основава на следното фундаментално свойство на квадратичните форми: "Ако редукцията на регресионната сума от квадрати, получена от премахването на подмножество от променливи с най-голям индекс j не надвишава стойността на редукцията θ_{j+1} от премахването на $(j+1)$-вата променлива, то не съществува подмножество, включващо променливи с индекси, по-големи от j, за което редукцията да е по-малка".

3. Пресмятат се редукциите на m-подмножества за всяко от всичките $\binom{k}{m}$ на брой подмножества с размер m. Всяко от тези подмножества с размер m се представя във вида $(i_1, i_2, ..., i_m)$, където $i_1, i_2, ..., i_m$ са променливите, от които е съставено, подредени в нарастващ ред, $1 < i_j < k$ и $i_1 < i_2 < ... < i_m$. Така пресметнатите редукции на m-подмножества се подреждат в нарастващ ред. Тези m-подмножества, чиито първи индекс е по-голям или равен на $(r - m + 1)$ служат за дефиниране на етапи за изследване на r - подмножества. Подмножествата с размер r, определени на даден етап съдържат m на брой индекси от съответното m-подмножество и $(r - m)$ индекси, по-малки от първия индекс в съответното m- подмножество. Броят на дефинираните по този начин етапи е $\binom{k - r + m}{m}$.

Етапите се номерират според големината на стойността на редукцията на m-подмножествата, които ги дефинират. Така, първият етап се състои от r-подмножествата, определени от m-подмножеството с най-малка стойност на редукция.

4. Изобщо, на всеки етап q се прави проверка дали най-малката от редукциите на r-подмножествата, дефинирани от всички етапи до момента $(1, ..., q)$ е по-малка от редукцията на m-подмножеството, дефиниращо следващия $(q+1)$-ви етап, и ако това е изпълнено, се счита, че е открито търсеното r-подмножество, което да

напусне изходния модел и съответстващото му най-добро p-подмножество, а в случай, че не е изпълнено, се преминава към следващия ($q+1$)-ви етап.

2.4 Резултати в Глава 2

1. Формулирани са изискванията към използваните за верификация признаци.
2. Описани са трансформациите на получените от графичния таблет изображения на подписи, необходими за привеждането им в стандартен вид и извличането на признаци. Представени са 26 глобални признака, необходими за математическото описание на подписа. Признаците с номера A8-A21 не се срещат в литературата до момента.
3. Описан е подход, основаващ се на последователното прилагане на метода на корелационните плеяди и C_p статистиката на *Mallows*, за избор на най-информативно от гледна точка на класификацията индивидуално за всеки потребител подмножество от признаци.

Глава 3. Класификация

Тази глава обхваща следните раздели.

3.1 Невронните мрежи като класификатори

3.1.1 Видове невронни мрежи

3.1.2 Обучение на невронна мрежа

3.1.3 Предимства и недостатъци на невронните мрежи

3.1.4 Невронна мрежа за класификация

3.1.5 Приложение на невронните мрежи в задачата за разпознаването по подпис

Невронните мрежи са подходящо средство за решаване на сложни задачи като разпознаването на подпис поради следните причини [13]. На първо място, в случай че са добре обучени, те обикновено дават верни отговори за входни образци, неизползвани при обучението. Също така многослойните невронни мрежи могат да моделират всяка функция на множество от променливи, а подписите могат лесно да се представят като такава функция. Поради възможността им за обобщаване, невронните мрежи могат да се справят с разнообразието и изменчивостта, характерни за образците на подписите. Тъй като в повечето случаи обучението на невронна мрежа отнема дълго време, в приложенията за разпознаване по подпис то обикновено се извършва еднократно и в оф-лайн режим, като по този начин на потребителите се спестява неудобството да чакат до приключването му. През последните години в литературата са предложени няколко системи за разпознаване на подпис, създадени чрез прилагането на невронни мрежи [13]. Идеалното обучение на една невронна мрежа за разпознаване по подпис е практически невъзможно, поради нестабилния характер на данните (подписът на даден потребител зависи от възрастта му, от неговото физическото и психическо състояние, от употребата на медикаменти и опиати, от вида на хартията, върху която се подписва, от вида на писалката и още много други).

Съществуват два режима за реализация на система за разпознаване на подписи с невронни мрежи, а именно верификация и идентификация. За режим верификация са необходими както собствени, така и чужди подписи от всеки потребител. За всеки участник се създава и съхранява единствена невронна мрежа, която има само един изходен неврон, чийто краен отговор е ниво на достоверност, приемащо стойности между нула (чужд подпис) и единица (собствен подпис), която показва степента на сходство на тестовия подпис с подписите, използвани при обучението за съответния участник. Тази стойност се сравнява с предварително избрана стойност на прага за участника и ако я надвишава, тестовият подпис се счита за собствен, а в противен случай за чужд. Този подход е широко приложим на практика и позволява бързото добавяне и премахване на подписи на нови и стари потребители. Така, при добавяне на нов потребител е необходимо единствено да се обучи малка по размер невронна мрежа.

За реализацията на режим идентификация са необходими единствено собствени подписи от всеки потребител. В този случай се създава и съхранява една обща невронна мрежа, която има толкова изходни неврони, колкото е броя на участниците. Изходът от идентификацията са толкова на брой стойности между нула и единица, колкото е и броят на всички потребители и като резултат се избира участника, съответстващ на

максималната изходна стойност на мрежата. Когато се използва обща невронна мрежа, е необходимо тя да се обучи наново при добавяне на участник, което несъмнено изисква повече време.

3.1.6 Проектиране на невронни мрежи

Задачата за проектиране на невронна мрежа включва следните подзадачи : *събиране на данни, създаване на мрежата, конфигуриране на мрежата, инициализиране на теглата и отклоненията й, обучение, оценка на мрежата и същинското й приложение* [11].

1) *Събиране на данни* - Качеството на събраните данни и множеството от наличните входни данни (признаци), които постъпват на входа й са от голямо значение.
2) *Създаване на мрежата* - Броят на елементите от входния слой зависи от размерността на входните данни, а броят на елементите от изходния слой - от броя на разпознаваните класове. Добра практика е [25] да се изберат некорелирани входни променливи и да се извърши тяхната нормиране.
3) *Конфигуриране на мрежата* - При проектирането на невронната мрежа следва да се вземе решение относно нейната топология (брой на слоевете и броя на невроните във всеки слой, техния тип и свързаност) и нейните параметри. Те са силно зависими от решавания проблем и наличните данни и затова стойностите им се намират експериментално.
4) *Инициализиране на теглата и отклоненията на мрежата* - Тъй като успехът на проектираната НМ зависи от произволните стойности на началните тегла, е необходимо да се обучат няколко НМ с различни инициализации на теглата и да се избере тази, която демонстрира най-добро качество.
5) Обучение - След като топологията на мрежата е дефинирана, е необходимо да се избере вида на обучаващия алгоритъм и да се премине към нейното обучение, така че да се оптимизира качеството й.
6) Оценка на мрежата и приложение – За оценяване на мрежата може да служи оценката на грешката *PCTError*, отразяваща процентното съотношение на грешно класифицираните образци към общия брой образци.

3.2 Метод на *k*-най-близки съседи

Методът на *k*-най-близки съседи е непараметричен метод за класификация. Отношение към точността на класификацията по метода на *k*-най-близки съседи имат: (1) стойността на *k*, (2) избора на мярка за разстояние или сходство, както и (3) метода за комбинирането на етикетите на класовете. В случай на класификация в два класа (дихотомия) се препоръчва нечетна стойност на *k*. Максимална стойност за *k* е $k_{max} = \sqrt{n}$ [13]. За определяне на стойността на *k* е удачно да се оцени точността на метода чрез LOOCV за множество от стойности на *k* и да се избере тази от тях, която води до най-висока точност [6].

3.3 Оценяване и сравняване на качеството на класификатори

След проектирането на даден класификатор (в частност невронни мрежи) е необходимо да се оцени неговото качество. Оценката на точността на класификатори, обучени с учител, е важна не само за предсказване на точността при бъдещото им използване, но също така и при избор на класификатор и комбинирането на класификатори. Традиционната практика е множеството от проектираните модели да се оценят и измежду тях да се избере този с най-малка грешка.

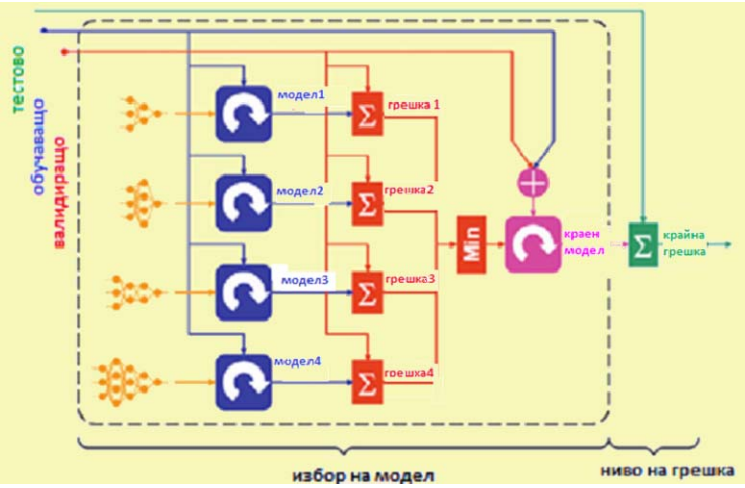
3.3.1 Оценка на класификатори

В случая на разпознаване в два класа, съответно *Positive* (положителни образци)

и *Negative* (положителни образци) се намира стойност на праг, който да ги разграничава най-добре. Така, всеки един пример попада в даден клас в зависимост от това дали резултата от разпознаването надхвърля стойността на прага (клас *Positive*) или не (клас *Negative*) [29]. Това разпределяне би могло да бъде вярно (*True*) или грешно (*False*). Така, за даден образец от клас *Positive*, ако изходът от класификацията е положителен, то образецът се явява истински положителен (*True Positive, TP*), а ако е отрицателен, то образецът е лъжливо отрицателен (*False Negative, FN*). Аналогично, даден образец от клас *Negative* би могъл да бъде истински отрицателен (*True Negative, TN*) или лъжливо положителен (*False Positive*, *FP*). Въз основа на *TP, FN, TN* и *FP* са дефинирани следните мерки за оценка на качеството на класификатори: *FAR, TAR, FRR, TRR, AC, Precision, PCTError, Sensitivity, Specifity*. За нагледно изобразяване на качеството на биометричните системи обикновено се изчертават т.нар. *Receiver Operating Characteristic* (ROC) криви, които представляват визуални средства за оценяване на точността на класификатор и за сравняване на различни класификационни модели.

3.3.2 Метод Holdout

Методът *Holdout* се прилага при наличието на достатъчно примери за обучение. Множеството от образци се разделя в две извадки: за обучение и за тестване. Образците от обучаващата извадка се използват за определяне на параметрите на модела, а тези от тестовата извадка се използват за получаването на оценка на качеството на класификатора върху непознати и неизползвани за обучението му данни и за сравняването на различни мрежови архитектури.



Фигура 3.2 Метод Holdout

Прилагането на метода *Holdout* в случая на невронни мрежи (Фигура 3.2) се състои в следното. Образците се разделят в [11]: обучаваща, валидираща и тестова извадки. Данните от обучаващата извадка се представят по време на обучението и се използват за настройване на теглата на невронните мрежи. Тези от валидиращата извадка се използват за измерване на качеството на всяка от проектираните невронни мрежи върху непознати данни и при взимането на решение за прекратяване на нейното обучение, когато качеството престане да се подобрява. Данните от тестовата извадка се използват за симулиране на невронната мрежа и измерване на нейното качество след приключване на обучението ѝ. Така кандидат невронните мрежи се нареждат във възходящ ред по качеството им, изчислено върху валидиращата извадка, а крайната грешка се смята върху данните от тестовата извадка.

3.3.3 Метод на N-кратна крос валидация (*N-fold cross validation*)

Методът на *N*-кратна крос валидация е приложим в случаите, когато на разположение са малък брой примери за обучение. Същността на метода се състои в следното. Данните се разделят произволно на *N* части (*folds*) с приблизително равен брой образци във всяка част. Последователно се съставят всичките *N* на брой възможни извадки, всяка от които съдържа обучаваща извадка, състояща се от (*N-1*) части и тестова извадка, съдържаща останалата една част. По този начин всяка част се използва за тестване точно един път. С усредняване на стойностите на грешките, получени за всяко разделяне на данните, се намира оценка на стойността на крайната грешка на класификатора.

Фигура 3.4 Метод на N-кратна крос-валидация

Характерно за този метод е това, че за обучението на класификатор се използват всичките налични примери за обучение и по този начин не се губят никакви данни. Методът на *N*-кратна крос валидация е изобразен схематично на Фигура 3.4.

Ако стойността на *N* е равна на размера на обучаващата извадка, то методът се нарича *Leave One Out Cross Validation (LOOCV)*. В този случай, тестовата извадка се състои от единствен образец, а за обучението се използват възможно най-много данни.

С прилагане на метода на *N*-кратна крос валидация може да се провери способността за обобщаване на невронната мрежа [11, 12] в случаите, когато на разположение са малък брой примери за обучение.

Невронната мрежа се обучава за всяко от тези *N* на брой множества с *K* на брой инициализации на началните тегла на връзките. По този начин за всяко разделяне $i = 1,\dots,N$ се намират *K* на брой оценки на грешките $e_{i,j}\ j = 1,\dots,K$, а стойността на крайната грешка за всяко разделяне E_i се изчислява като средната стойност от тези *K* на брой оценки:

$$E_i = \frac{1}{K}\sum_{j=1}^{K} e_{i,j} \qquad i = 1,\dots,N; j = 1,\dots,K \tag{3.13}$$

Оценките E_i на съответните класификатори са неизместени, тъй като данните от тестовата извадка не се съдържат в обучаващата извадка. Оценката на конкретния модел $\hat{E}$ се намира като средната стойност от грешките, получени за всяко разбиване на данните:

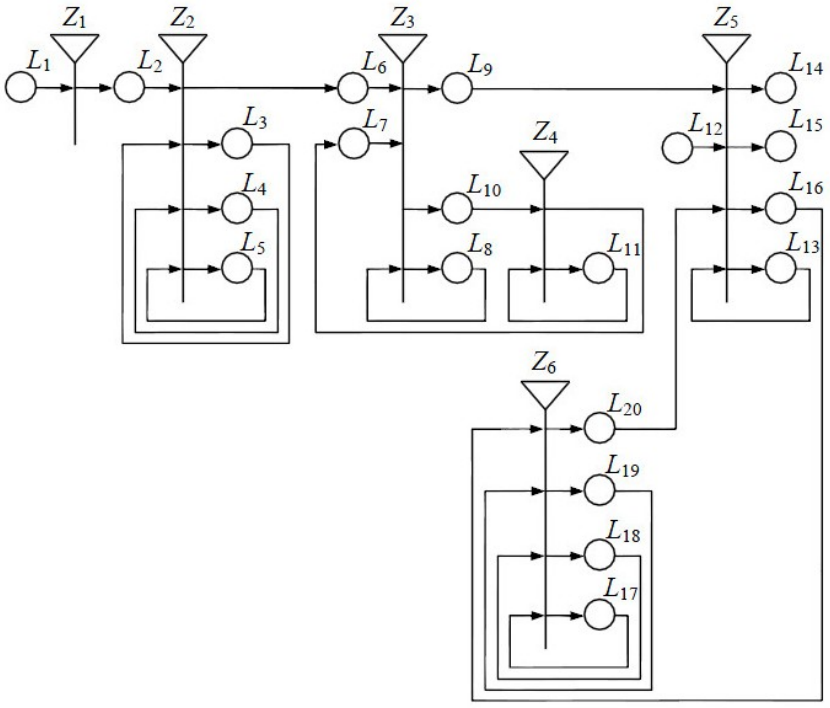
$$\hat{E} = \frac{1}{N}\sum_{i=1}^{N} E_i \qquad (3.14)$$

Оценката $\hat{E}$ разчита на допускането, че точността на класификацията би се променила плавно с промяна в размера на обучаващата извадка. Следователно, ако точността на модела е задоволителна, изследователят би могъл или да обучи единствен класификатор върху цялото множество от данни, или да комбинира класификаторите, построени по време на обучението [27].

Методът на *N*- кратна крос валидация е удачен избор при търсенето на подходяща стойност на *k* в класификатора *k*NN, а също и при търсенето на оптимална топология на НМ [6]. След намирането на оптимални стойности на параметрите на класификатор, той се обучава върху цялото множество от данни и бива използван на практика.

3.4 Обобщен мрежови модел за он-лайн верификация на подписи

С апарата на обобщените мрежи (ОМ) е направен опит за моделиране на процеса на верификация по подпис с цел ясно и нагледно представяне на предложения в дисертацията метод, който включва следните стъпки за верификация на заявена самоличност по подпис: полагане на подпис, предварителна обработка на подпис, извличане на признаци, избор на признаци, обучение и верификация.



Фигура 3.6 Обобщено мрежови модел на процеса

Представената редуцирана обобщена мрежа е без времеви компоненти, без зададени приоритети на преходите, позициите и ядрата и представлява модификация на предложените в [35, 36] модели на ОМ за различни задачи от разпознаването на образи. Обобщено мрежовият модел е представен на Фигура 3.6. Той е изграден от 6 прехода (всеки един от които представя стъпка от процеса), 20 позиции и 3 вида ядра α, β и γ.

3.5 Резултати в Глава 3

1. Представена е кратка класификация на различни видове невронни мрежи

според архитектурата и алгоритмите за обучение. Коментирани са предимствата и недостатъците им.
2. Анализирани са въпросите за изграждане и обучение на невронни мрежи.
3. Описани са видовете грешки от класификация, методите за оценка на точността и построяването на комбиниран класификатор.
4. Изграден е модел на процеса на он-лайн верификация с апарата на обобщените мрежи.

Глава 4. Експерименти и резултати

В настоящата глава са представени получените резултати от провеждането на експериментите за създаване и тестване на комбинираната система за разпознаване по подпис. Тази система може да служи за верификация въз основа на признаците, описани в Раздел 2.2. За решаване на поставената задача са проектирани класификатори невронни мрежи от тип многослоен перцептрон и класификатори по *k*-най-близки съседи, които са тествани върху образците на подписи от две бази данни. Комбинираният подход за разпознаване по подпис се прилага като поредица от следните стъпки: (1) предварителен избор на признаци и (2) верификация.

4.1 Избор на признаци за разпознаване

За всеки от потребителите се извличат стойностите на признаците, изчислени по характеристиките на всичките му осем или десет собствени и десет подправени подписа. В настоящата дисертация е експериментирано с умели и неумели фалшификати. Неумелите фалшификати са избрани произволно от собствените подписи на останалите участници. Признаците са глобални и са описани в раздел 2.2. Тъй като всеки от тях получава стойности в различни интервали при различна скала на измерване, се налага необходимостта от нормиране на признаците в интервала [0,1]. Извършва се нормировка *min-max* (раздел 2.3.1), която запазва взаимовръзките между оригиналните стойности на данните. След извличането на признаци, се пристъпва към намирането на (1) ***общо*** за всички потребители подмножество от признаци с прилагане на метода на корелационните плеяди (точка 2.3.2); (2) ***индивидуално*** подмножество от признаци на базата на собствени подписи и ***неумели*** фалшификати (наричано по-долу Вариант "1"); (3) ***индивидуално*** подмножество от признаци на базата на собствени подписи и ***умели*** фалшификати (наричано по-долу Вариант "2"). Така намерените подмножества от признаци служат за построяването на класификационните модели.

След намирането на общо за всички потребители подмножество от признаци, се прилага метода за избор на регресионни променливи, описан в Раздел 2.3 за намирането на индивидуалните подмножества от признаци (2) и (3). Така се откриват най-добрите подмножества от признаци с различен размер за всеки участник въз основа на неговите собствени и подправени подписи. Сред тях се избира това подмножество от признаци, за което стойността на *Cp* е най-близка до тази на *p*, където *p* е броят на регресионните коефициенти. Получените по този начин подмножества от признаци са с различен размер за различните потребители. Нека означим броя на признаците с *k*, а този на подписите на даден потребител с *n*. За всеки потребител се създава текстов файл с двадесет реда, равен на броя подписи, определени за обучение. Всеки ред съдържа стойности на признаците за съответния подпис, разделени с точка и запетая, следвани от 1 (ако конкретният подпис е собствен, тоест принадлежи на първия клас) и -1 (ако подписът е подправен, тоест принадлежи на втория клас). Необходимо е да се подберат тези признаци (независими променливи), които възможно най-добре апроксимират наблюдаваните значения на зависимата променлива *y* ($y = 1$ за точките от първия клас, собствен подпис и $y = -1$ за тези от втория клас, фалшифициран подпис). Този файл се обработва и за всеки потребител / файл се откриват най-добрите подмножества с размер p ($p = k - r$) за стойности на r ($3 < r < 13$), а сред тях се намира подмножеството със стойност на *Cp* най-близка до *p,* но по-малка от нея. Това подмножество от признаци се приема за "най-добро".

4.2 Построяване на модели за класификация

Експериментите са извършени с класификатори невронни мрежи и класификатор по *k*-най-близки съседи. Построени са класификатори с вариращи параметри (брой

признаци, вид фалшифицирани подписи за обучение, стойност на *H* при НМ и стойност на *k* при *k*NN). Поради малкия обем данни за обучение, точността на различните модели за класификация е оценена с 10-кратна крос валидация (НМ) и LOOCV (*k*NN).

4.2.1 Класификатор невронна мрежа

За всеки потребител се създава отделна НМ. Входът *p* на всяка от тестваните невронни мрежи представлява матрица [*FeatureSize Ntrn*] с брой редове, равен на броя на признаците *FeatureSize* и брой стълбове, равен на броя на образци за обучение *Ntrn*. Изходът зависи от режима на разпознаване. В случай на верификация той е матрица с размерност [*1 Ntrn*], със стойност в [0,1] във всеки от редовете, равна на 1 за собствен подпис и 0 – за подправен.

В процеса на обучение участват и собствени и подправени подписи. Експериментирано е с използването само на неумели фалшификати (наричано по-долу Случай ”1”) и едновременно на умели и неумели фалшификати (наричано по-долу Случай ”2”). В търсене на оптимална архитектура на НМ, за всеки потребител се експериментира с различни топологии на един и същ вид невронна мрежа според променливия размер на броя на невроните в скрития слой. Топологията на НМ се избира чрез крос валидация.

Така, за всеки потребител се намират параметри на модел на НМ (брой скрити неврони, вид фалшифицирани подписи за обучение, брой признаци на входа и др.). Изходът от верификацията е стойност от интервала [0,1]. Окончателната невронна мрежа се запаметява заедно с архитектурата и стойностите на теглата и може да се използва по-нататък за верификация.

4.2.2 Класификатор по *k*-най-близки съседи

За създаването на класификатор по *k*-най-близки съседи са следвани препоръките, изложени в Раздел 3.2. Първоначално се прилага нормиране на признаците, така че те да имат стандартно отклонение, равно на единица и средна стойност, равна на нула. Използвано е Евклидово разстояние за пресмятане на разстоянията между образците. Тъй като $k_{max} = \sqrt{Ntrn}$, $Ntrn$ – брой примери за обучение и стойността на *k* е препоръчително да е нечетно число, то са тествани са две стойности на *k*: 1 и 3.

4.3 Получени резултати върху собствена база от данни за подписи

Резултатите от избора на най-информативните признаци по метода на Hocking и Leslie (Раздел 2.3) за всеки от осемте потребителя са обобщени в таблица 4.2. От нея е видно, че в случая на вариант “2” броя на признаците за всички потребители без един се редуцира до 13, като само за потребител #5 те са 5 признака, докато при вариант “1” броя на признаците за половината участници се редуцира до брой, по-малък от 13.

Таблица 4.2 Редуцирани признаци за всеки потребител

Участник	**#1**		**#2**		**#3**		**#4**		**#5**		**#6**		**#7**		**#8**	
Вар.	“1”	“2”	“1”	“2”	“1”	“2”	“1”	“2”	“1”	“2”	“1”	“2”	“1”	“2”	“1”	“2”
A2	+	+	+			+		+	+			+	+	+	+	
A5		+	+	+	+	+	+	+	+	+	+		+	+		+
A6	+		+	+		+		+	+	+	+	+			+	+
A8		+	+	+		+	+	+	+	+		+	+		+	

A11	+	+				+					+			+	+	+
A12	+			+		+		+	+			+	+	+	+	+
A13	+	+		+	+	+		+			+	+	+	+		+
A15	+	+	+	+		+		+			+	+	+	+		
A16			+	+	+	+		+	+		+	+	+	+	+	
A17		+	+	+	+	+	+	+	+		+	+		+		+
A20	+	+	+	+	+		+		+		+			+		+
A21	+	+	+	+	+	+	+	+	+		+	+	+	+	+	+
A22	+	+						+	+	+	+	+	+	+	+	+
A23	+	+	+		+		+				+		+		+	+
A24	+	+	+	+	+	+	+	+			+	+	+		+	+
A25	+		+	+					+			+	+	+		+
A26	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Ср	10,	9,0	9,09	9,5	8,1	9,05	7,1	9,2	10,	4,8	9,08	9,33	9,12	9,24	10,3	9,77
p	13	13	13	13	9	13	8	13	12	5	13	13	13	13	11	13

Проектираните невронни мрежи са 30 на брой (Таблица 4.3), а проектираните класификатори по *k*-най-близки съседи са 12 на брой за всеки участник.

Таблица 4.3 Параметри на моделите

# на модел	**Брой признаци (входни**	**Брой собствени подписи**	**Брой подправени подписи**		**Брой скрити неврони**	**Брой съседи *k***
1	Общи признаци	10	15 неумели	Случай “1”	от 1 до 5	1 или 3
2	Вариант “1”					
3	Вариант “2					
4	Общи признаци		9 неумели и 6 умели	Случай “2”		
5	Вариант “1”					
6	Вариант “2”					

В случай на равни максимални стойности на точността за два или повече модела за даден участник, е избран този от тях, построен с по-малък брой признаци и / или брой скрити неврони (НМ) и / или брой съседи (kNN). Средната стойност на приближената оценка на точността за класификатор НМ за всички участници е равна на 97.5%, а тази на точността за класификатор *k*NN е равна на 92.5%.

След намирането на оптимални стойности на параметрите на класификатор по метода на крос валидация, той се обучава върху цялото множество от данни и бива използван на практика.

4.4 Получени резултати върху база данни за подписи SUsig

Резултатите от избора на най-информативни признаци по метода на *Hocking* и *Leslie* (Раздел 2.3) за всеки от потребителите са обобщени в таблица 4.7.

Таблица 4.7 Редуцирани признаци за всеки потребител от SUsig

Участник	**Вар.**	**A1**	**A2**	**A4**	**A6**	**A10**	**A12**	**A13**	**A16**	**A17**	**A21**	**A22**	**A23**	**A24**	**Ср**	**p**
#1	"1"	+		+	+	+		+	+			+	+	+	5,35	9
	"2"	+			+		+	+	+		+	+	+	+	5,25	9
#2	"1"	+			+	+	+	+		+	+	+		+	7,77	9
	"2"		+	+	+		+		+	+	+	+		+	5,06	9

	"1"		+	+	+	+	+	+	+	+				+	7,65	9
#3	"2"	+	+		+			+	+		+			+	6,39	7
	"1"					+			+	+		+		+	4,8	5
#4	"2"				+			+	+			+		+	2,94	5
	"1"	+	+	+		+	+				+			+	4,94	7
#5	"2"	+	+	+		+		+			+	+	+	+	5,22	9
	"1"	+	+	+	+	+	+			+				+	8	9
#6	"2"	+	+	+	+		+	+		+	+			+	6,42	9
	"1"	+	+	+	+		+			+			+	+	5,24	8
#7	"2"	+	+	+	+	+			+	+		+		+	5,01	9
	"1"		+	+	+			+			+			+	2,22	6
#8	"2"	+	+	+	+	+	+				+	+		+	5,23	9
	"1"			+	+		+		+			+	+	+	4,47	7
#9	"2"		+	+	+		+		+	+		+	+	+	5,28	9
	"1"						+		+	+				+	2,03	4
#10	"2"	+		+	+	+	+				+	+	+	+	6,02	9
	"1"	+	+				+		+	+				+	3,1	6
#11	"2"	+	+	+	+			+	+		+	+		+	5,75	9
	"1"	+		+				+	+	+	+	+	+	+	5,57	9
#12	"2"			+						+				+	0,71	3
	"1"		+	+		+	+					+		+	5,39	6
#13	"2"	+	+	+	+	+		+			+	+		+	5,05	9
	"1"			+				+	+			+	+	+	3,24	6
#14	"2"	+			+		+	+		+	+		+	+	5,93	8
	"1"	+	+	+	+	+	+			+		+		+	5,39	9
#15	"2"		+	+	+	+	+	+			+	+		+	5,13	9
	"1"	+			+	+	+	+	+		+	+		+	5,01	9
#16	"2"	+				+		+	+	+	+	+	+	+	5,22	9
	"1"				+					+				+	2,96	3
#17	"2"	+	+		+		+		+	+	+		+	+	5,12	9
	"1"	+	+	+	+	+			+		+		+	+	5,29	9
#18	"2"			+		+			+	+		+		+	5,74	6
	"1"	+	+	+	+		+		+		+		+	+	7,11	9
#19	"2"	+		+	+		+	+		+		+	+	+	5,38	9
	"1"		+	+	+	+	+			+		+	+	+	5,48	9
#20	"2"				+			+			+		+	+	4,65	5
	"1"						+	+		+		+		+	4,94	5
#21	"2"		+	+	+			+	+	+	+	+		+	5,68	9
	"1"	+	+		+				+				+	+	3,75	6
#22	"2"	+		+	+	+		+	+		+	+		+	5,14	9
	"1"	+	+	+	+	+	+		+	+				+	5,25	9
#23	"2"	+		+			+	+	+		+	+	+	+	5,12	9
	"1"	+	+	+	+			+	+			+	+	+	6,71	9
#24	"2"	+	+		+			+				+		+	2,34	6
	"1"		+	+		+	+	+	+		+	+		+	5,04	9
#25	"2"	+	+	+	+			+	+		+		+	+	5,05	9
	"1"		+		+	+			+	+	+	+		+	6,84	8
#26	"2"		+	+		+				+			+	+	3,72	6
	"1"			+	+	+	+		+					+	3,47	6
#27	"2"		+	+	+		+	+	+		+	+		+	5,28	9
	"1"	+		+	+	+	+		+		+		+	+	8,41	9
#28	"2"	+	+	+								+		+	1,55	5
	"1"	+	+	+	+	+		+				+	+	+	12,15	9
#29	"2"	+	+	+		+	+	+					+	+	7,52	8
	"1"	+		+	+	+	+		+		+	+		+	5,27	9
#30	"2"				+				+			+		+	1,18	4
	"1"	+	+		+		+		+	+	+	+		+	8,81	9
#31	"2"		+	+	+		+	+	+		+			+	5,7	8

	"1"	+	+		+	+	+			+		+	+	+	8,38	9
#32	"2"	+	+	+			+		+	+	+	+		+	5,21	9
	"1"		+		+			+			+		+	+	4,49	6
#33	"2"	+			+	+			+	+	+	+	+	+	5,71	9
	"1"	+		+	+	+			+	+		+	+	+	5,39	9
#34	"2"	+		+	+	+				+	+	+	+	+	5,24	9
	"1"	+	+		+								+	+	4,16	5
#35	"2"	+			+						+	+		+	2,62	5
	"1"	+	+	+	+	+		+				+	+	+	5,2	9
#36	"2"	+		+		+	+	+	+		+	+		+	5,1	9
	"1"		+	+		+		+	+			+	+	+	7,43	8
#37	"2"	+	+			+		+		+	+	+		+	7,58	8
	"1"	+							+		+		+	+	4,38	5
#38	"2"	+	+		+		+		+		+	+	+	+	5,38	9
	"1"		+	+	+	+	+	+				+	+	+	5,13	9
#39	"2"	+	+	+		+					+			+	5,21	6
	"1"			+			+				+		+	+	4,67	5
#40	"2"		+	+	+					+				+	2,17	5
	"1"	+	+			+	+		+	+	+		+	+	7,48	9
#41	"2"	+		+	+	+			+	+		+	+	+	5,11	9
	"1"						+	+			+	+		+	2,5	5
#42	"2"	+		+		+			+					+	2,73	5
	"1"	+	+	+	+	+	+	+				+		+	5,04	9
#43	"2"	+		+			+				+		+	+	4,5	6
	"1"		+		+				+	+		+	+	+	3,68	7
#44	"2"	+				+	+		+	+		+		+	6,06	7
	"1"		+	+		+		+	+	+	+	+		+	5,35	9
#45	"2"		+	+	+	+		+		+	+	+		+	5,09	9
	"1"	+		+	+			+					+	+	4,72	6
#46	"2"	+	+	+	+			+		+	+	+		+	5,16	9
	"1"	+	+				+			+				+	1,84	5
#47	"2"	+	+	+	+	+				+		+	+	+	5,14	9
	"1"		+	+	+	+	+				+	+	+	+	5,18	9
#48	"2"	+	+		+	+	+	+			+	+		+	5,36	9
	"1"		+	+		+						+		+	3,11	5
#49	"2"	+	+	+	+	+			+		+		+	+	5,67	9
	"1"		+	+	+	+	+		+		+	+		+	7,57	9
#50	"2"			+					+		+	+		+	2,65	5
	"1"		+	+	+	+	+		+	+	+			+	5,22	9
#51	"2"	+		+	+	+	+	+			+		+	+	6,35	9
	"1"	+	+	+	+	+		+	+		+			+	5,16	9
#52	"2"	+	+		+		+		+		+	+	+	+	5,08	9
	"1"	+	+	+		+			+	+	+		+	+	5,38	9
#53	"2"			+	+	+		+	+					+	3,93	6
	"1"		+			+		+	+		+		+	+	5,92	7
#54	"2"	+		+	+	+	+		+			+	+	+	8,69	9
	"1"	+					+						+	+	2,3	4
#55	"2"		+	+	+	+	+		+		+	+		+	5,21	9
	"1"	+		+		+	+	+	+		+		+	+	8,13	9
#56	"2"	+	+		+		+	+			+	+		+	7,58	8
	"1"		+	+			+	+	+	+		+	+	+	8,19	9
#57	"2"	+		+			+		+	+	+	+	+	+	5,73	9
	"1"		+	+		+	+	+		+	+		+	+	5,98	9
#58	"2"	+	+		+			+	+	+	+		+	+	8,96	9
	"1"	+		+	+	+	+			+		+	+	+	6,46	9
#59	"2"		+	+		+		+		+	+	+	+	+	5,44	9
#60	"1"		+	+			+	+				+		+	3,93	6

	"2"	+	+		+	+		+			+			+	6,03	7
	"1"		+	+	+	+		+				+	+	+	7,52	8
#61	"2"			+		+			+	+				+	3,79	5
	"1"	+	+			+	+	+		+		+	+	+	8,31	9
#62	"2"		+		+	+	+		+	+	+		+	+	5,45	9
	"1"		+	+	+					+			+	+	3,56	6
#63	"2"			+								+		+	0,69	3
	"1"				+			+			+		+	+	4,72	5
#64	"2"	+		+	+		+	+	+		+	+		+	5,03	9
	"1"	+			+	+	+	+	+			+	+	+	5,1	9
#65	"2"		+	+			+		+			+	+	+	6,89	7
	"1"	+	+	+	+	+	+			+	+			+	8,34	9
#66	"2"		+	+	+			+	+	+	+		+	+	5,22	9
	"1"		+		+		+		+	+			+	+	5,67	7
#67	"2"	+			+		+	+				+		+	3,61	6
	"1"		+		+				+				+	+	4,38	5
#68	"2"				+		+					+	+	+	4,95	5
	"1"	+		+		+		+	+		+		+	+	5,74	8
#69	"2"		+	+	+	+		+	+	+	+			+	5,27	9
	"1"	+	+	+	+	+	+		+			+		+	8,81	9
#70	"2"		+		+	+	+	+	+		+		+	+	5,1	9
	"1"							+					+	+	2,76	3
#71	"2"			+		+		+					+	+	3,5	5
	"1"		+								+			+	2,6	3
#72	"2"		+		+			+	+					+	4,66	5
	"1"	+	+									+		+	1,28	4
#73	"2"		+	+		+	+		+	+		+		+	5,89	8
	"1"	+			+	+	+	+		+	+		+	+	5,09	9
#74	"2"	+		+		+				+				+	3,68	5
	"1"	+	+			+	+	+	+					+	4,29	7
#75	"2"	+			+				+	+				+	4,15	5
	"1"	+	+	+	+			+	+			+	+	+	8,02	9
#76	"2"			+		+		+				+		+	4,36	5
	"1"			+		+		+	+					+	2,67	5
#77	"2"		+	+		+	+		+	+	+		+	+	5,02	9
	"1"	+	+				+		+	+	+	+	+	+	6,63	9
#78	"2"					+			+				+	+	1,1	4
	"1"	+		+		+	+	+			+	+	+	+	8,06	9
#79	"2"	+	+	+	+		+			+		+	+	+	8,41	9
	"1"			+						+			+	+	3,34	4
#80	"2"			+	+	+			+	+				+	5,78	6
	"1"	+				+				+	+		+	+	4,26	6
#81	"2"			+						+	+	+	+	+	4,28	6
	"1"	+				+		+					+	+	3,66	5
#82	"2"				+			+			+			+	2,89	4
	"1"	+	+		+	+		+	+	+	+			+	5,33	9
#83	"2"	+		+	+		+	+		+	+		+	+	5,26	9
	"1"			+					+	+	+		+	+	4,44	6
#84	"2"		+				+		+	+	+	+	+	+	6,81	8
	"1"						+			+	+		+	+	3,89	5
#85	"2"	+		+					+	+				+	4,23	5
	"1"						+					+		+	1,01	3
#86	"2"		+	+	+	+		+		+	+	+		+	5,68	9
	"1"		+	+	+		+		+			+		+	6,64	7
#87	"2"	+	+		+	+		+	+			+	+	+	5,96	9
	"1"		+						+	+				+	3,36	4
#88	"2"			+	+	+			+			+		+	5,26	6
#89	"1"	+			+					+			+	+	3,65	5

	"2"	+		+			+	+	+	+		+	+	+	5,61	9

В Таблица 4.8 и за двата варианта са посочени брой на редуцираните признаци и съответния брой участници.

.

Таблица 4.8 Брой срещания на редуцирани признаци

Вар.	Брой на редуцирани признаци	Брой участници
“1”	9	42
“1”	8	5
“1”	7	7
“1”	6	12
“1”	5	14
“1”	4	5
“1”	3	4

Вар.	Брой на редуцирани признаци	Брой участници
“2”	9	48
“2”	8	9
“2”	7	4
“2”	6	10
“2”	5	15
“2”	4	3
“2”	3	0

Вижда се, че е постигнато значително редуциране на броя на признаците, понеже първоначалния им брой (равен на 13) се редуцира до 9 за около половината участници и при двата варианта, а за 30% от участниците (26 души) и при двата варианта се намалява до 5 или 6 признака. При малък брой участници признаците достигат брой, по-малък от 5 (за Вариант “1” – 9 души, а за Вариант “2” – само 3). Тези резултати показват, че първоначалният брой признаци се редуцира значително и е по-малък при използване на неумели фалшификати за отрицателни примери (Вариант “1”).

Таблица 4.9 съдържа обобщена информация за всеки от построените модели (с номера от 1 до 6) за базата подписи SUsig. Те се различават по използваните отрицателни примери за обучение и по броя на признаците.

За всеки участник се извършва оценка на точността за всяка НМ по метода на 10-кратната крос валидация и оценка на точността на всеки класификатори *k*NN по LOOCV, а за окончателен се избира този модел и съответната стойност на *H* или *k*, който демонстрира най-висока точност. В Таблица 4.10 са посочени параметрите на така избраните модели за всеки участник.

Таблица 4.9Параметри на моделите за базата SUsig

# на модел	Брой признаци (входни неврони)	Брой собствени подписи	Брой подправени подписи		Брой скрити неврони H	Брой съседи *k*
1	Общи признаци	8 или 10	15 неумели	Случай “1”	от 1 до 5	1 или 3
2	Вариант “1”					
3	Вариант “2”					
4	Вариант “2”		9 неумели и 6 умели	Случай “2”		
5	Вариант “1”					
6	Общи признаци					

Таблица 4.10 Окончателни модели за всеки участник от SUsig

Участ ник	№ на избран модел за НМ	Брой скрити неврони *H*	Точност по 10-кратна CV (%)	№ на избран модел за kNN	Брой съседи *k*	Точност по

						LOOCV (%)
#1	1	5	100	2	1	100.00
#2	3	3	98.33	4	1	95.65
#3	5	5	93.7	5	3	86.96
#4	4	2	95.93	3	1	92.00
#5	2	5	98.89	2	1	96.00
#6	4	4	98.15	5	1	100.00
#7	3	4	97.41	5	1	95.65
#8	1	3	97.41	4	3	86.96
#9	5	2	98.89	4	1	96.00
#10	4	4	99.07	2	3	96.00
#11	6	2	98.52	3	1	92.00
#12	6	1	99.63	5	1	92.00
#13	1	1	93.33	5	1	96.00
#14	2	3	99.26	2	3	100.00
#15	2	4	99.26	2	3	100.00
#16	2	2	99.07	2	1	100.00
#17	3	1	97.78	3	1	100.00
#18	2	1	96.48	2	1	92.00
#19	5	4	99.63	5	1	100.00
#20	6	3	97.78	5	1	95.65
#21	1	5	99.26	1	1	100.00
#22	5	5	97.41	2	1	95.65
#23	1	3	90	4	1	80.00
#24	5	3	98.7	2	1	96.00
#25	2	2	97.78	4	1	100.00
#26	4	2	97.22	4	1	100.00
#27	2	4	97.78	5	1	96.00
#28	2	2	98.89	4	1	96.00
#29	2	4	99.63	4	1	96.00
#30	3	3	97.04	4	1	92.00
#31	6	3	99.44	4	1	100.00
#32	2	2	98.15	5	1	100.00
#33	4	4	98.15	2	1	91.30
#34	2	2	97.96	2	1	91.30
#35	6	4	97.04	4	1	91.30
#36	2	4	99.44	2	1	95.65
#37	2	5	98.33	5	1	92.00
#38	1	2	99.63	2	1	100.00
#39	4	3	98.52	1	1	100.00
#40	2	2	94.07	5	1	88.00
#41	1	5	98.89	3	1	100.00
#42	2	4	98.52	2	1	96.00
#43	4	2	99.26	3	1	96.00
#44	4	2	97.59	4	1	100.00
#45	3	5	96.3	5	1	92.00
#46	1	5	96.67	2	3	92.00

#47	3	2	99.26	2	1	100.00
#48	4	4	97.41	4	1	86.96
#49	2	2	100	3	1	95.65
#50	1	3	97.04	3	1	96.00
#51	1	3	97.96	2	1	96.00
#52	2	3	99.44	1	1	100.00
#53	3	4	97.59	2	3	96.00
#54	3	4	96.11	2	1	92.00
#55	5	3	99.07	5	1	100.00
#56	5	4	98.52	5	1	92.00
#57	6	5	99.44	3	1	100.00
#58	6	4	91.67	2	1	76.00
#59	2	4	99.63	5	1	100.00
#60	6	2	96.3	5	3	88.00
#61	2	4	98.33	5	1	100.00
#62	6	2	97.59	4	1	96.00
#63	6	2	99.44	5	1	100.00
#64	3	5	99.07	2	1	100.00
#65	2	5	98.52	3	1	100.00
#66	1	2	99.26	3	1	100.00
#67	3	3	99.26	2	1	100.00
#68	2	2	98.7	3	1	96.00
#69	3	4	98.33	3	1	100.00
#70	5	2	100	4	3	95.65
#71	3	5	95.56	4	1	96.00
#72	1	2	99.44	2	1	100.00
#73	2	4	100	2	1	100.00
#74	4	4	99.63	4	1	100.00
#75	2	5	97.22	2	1	96.00
#76	4	5	92.78	6	3	84.00
#77	5	3	95.93	5	1	95.65
#78	2	5	98.52	3	1	100.00
#79	2	4	95.19	2	1	92.00
#80	4	2	97.96	4	1	96.00
#81	6	5	97.96	5	1	100.00
#82	4	4	99.63	4	1	100.00
#83	3	4	100	3	1	100.00
#84	2	1	98.89	3	1	100.00
#85	2	2	97.78	2	1	100.00
#86	1	5	96.85	4	1	100.00
#87	4	1	96.48	1	1	100.00
#88	4	3	100	1	1	100.00
#89	4	3	99.63	4	1	100.00

Сравнение на точността на класификаторите HM и *k*NN

Целта на сравнението е да се провери как избраният подход за построяване на класификатор (HM) се съотнася към други широко използвани класификатори. При

малки извадки, когато оптимални в статистически смисъл подходи, не са приложими, най-често се използват евристични подходи като k – най-близки съседи. Следва да се отбележи, че при голям брой наблюдения и голямо k този подход е квази оптимален, т.е. води до получаване на класификатор, близък до оптималния Бейсов класификатор. В Таблици 4.10 – 4.12 са дадени резултатите от проведеното изследване.

Средната стойност на точността на НМ, оценена за всички участници от SUsig по избраните за тях модели чрез 10-кратна крос-валидация, е равна на 97.95%. Средната стойност на точността на kNN, оценена за всички участници от SUsig по избраните за тях модели чрез LOOCV, е равна на 96.13%. Тези резултати показват предимство за невронната мрежа. За отбелязване е, че най-добри резултати за kNN се получават при $k = 1$ за 79 потребители или 89% от случаите. За изчислените средни стойности и средноквадратични отклонения е изчислена стойността на t-статистиката $t = 3.29$. При 176 степени на свобода тази стойност е значима с вероятност 0.99.

Зависимост на точността на верификация от подмножеството признаци за разпознаване и вида фалшиви подписи за обучение

Изследването има за цел (1) да се оцени ефекта от редуцирането на броя на признаците върху точността на класификацията (2) да се изследва зависимостта на точността на класификация от вида фалшиви подписи, използвани за обучение на класификаторите. За целта изчисляваме средните стойности на оценената точност по най-добрите модели за НМ (Таблица 4.11) и kNN (Таблица 4.12), средната точност по Общи признаци, по признаци от Вариант “1” и по признаци от Вариант “2”.

Данните от Таблица 4.11 показват по-голяма точност на НМ при използване на индивидуалните признаци (Вариант “1” и Вариант “2”) спрямо точността по общи признаци като се наблюдава превес на точността по признаци от Вариант “1”. Точността по Вариант “1”, Случай “1” (неумели фалшификати за обучение) надвишава точността по Вариант “2”, Случай “2” (умели и неумели фалшификати за обучение).

Таблица 4.11 Таблица със средни стойности на оценената точност по най-добрите модели на НМ

	Общи признаци	**Вариант 1**	**Вариант 2**	**Средна точност**	**Брой случаи**
Случай 1	**97.36** *(Модел 1, 13 случая)*	**98.36** *(Модел 2, 27 случая)*	**97.85** *(Модел 3, 13 случая)*	**97.99**	**53**
Случай 2	**97.71** *(Модел 6, 11 случая)*	**97.98** *(Модел 5, 9 случая)*	**97.96** *(Модел 4, 16 случая)*	**97.89**	**36**
Средна точност	**97.52**	**98.27**	**97.91**	**-**	**-**
Брой случаи	**24**	**36**	**29**	**-**	**89**

При прилагане на класификатор kNN Таблица (4.12) точността по общите

признаци надвишава тази по индивидуалните. Точността по Вариант "2", Случай "1" (неумели фалшификати за обучение) надвишава точността по Вариант "2", Случай "2" (умели и неумели фалшификати за обучение).

Таблица 4.12 Таблица със средни стойности на оценената точност по най-добрите модели на kNN

	Общи признаци	**Вариант 1**	**Вариант 2**	**Средна точност**	**Брой случаи**
Случай 1	**100** *(Модел 1, 5 случая)*	**95.92** *(Модел 2, 27 случая)*	**97.84** *(Модел 3, 15 случая)*	**96.97**	**47**
Случай 2	**84** *(Модел 6, 1 случай)*	**95.5** *(Модел 5, 20 случая)*	**95.45** *(Модел 4, 21 случая)*	**95.20**	**42**
Средна точност	**97.33**	**95.74**	**96.45**	**-**	**-**
Брой случаи	**6**	**47**	**36**	**-**	**89**

Сравнението на средната точност между НМ и *k*NN по общи и индивидуални признаци показва превъзходство на НМ.

Броят на избраните модели, обучени с неумели фалшификати (Случай "1"), се среща при повече от половината участници. Той е равен на 53 (при 60 % от участниците) за НМ и равен на 47 (при 53 % от участниците) за *k*NN. Броят на избраните модели, построени по признаци от Вариант "1" е най-голям за НМ (за 36 участника), както и за *k*NN (за 47 участника).

Построяване на класификатор НМ и *k*NN по избраните модели за всеки потребител

В Таблица 4.13 са показани резултатите от оценката на горните класификатори върху едни и същи извадки от 830 собствени и 890 фалшиви подписи. Около 70% от подписите се използват за проектиране на класификатора, а останалите 30% за тестване. При невронните мрежи за обучение са използвани по 3 или 4 собствени подписа и 7 фалшиви, а валидацията е извършена върху 2 или 3 собствени и 3 фалшиви подписа за всеки потребител. Тестването е извършено върху 3 собствени и 5 фалшиви подписа. При 9-кратна инициализация е избрана тази невронна мрежа, за която е получена най-добра оценка. За *k*NN са използвани 5 или 7 собствени подписа и 10 фалшиви, а за тестване – 3 собствени и 5 фалшиви. Разликата в броя използвани собствени подписи за обучение се дължи на различния брой събрани подписи от всеки участник. В Таблица 4.13 за съответните модели за всеки участник са посочени стойностите на праг на НМ, номера на инициализация на теглата, стойности на FAR, FRR и точността по НМ и *k*NN.

Таблица 4.13 Оценка на качеството на класификация с НМ и kNN за всички участници

Участник	**Праг (НМ)**	**№ на инициа лизация**	**FAR (НМ)**	**FRR (НМ)**	**Точност по НМ**	**FAR (*k*NN)**	**FRR (kNN)**	**Точност по kNN**
#1	0.82	7	0	0	100	0	0	100
#2	0.82	6	0	0	100	20	0	87.5
#3	0.77	4	0	0	100	0	0	100

#4	0.83	5	0	0	100	0	0	100
#5	0.92	3	0	0	100	0	0	100
#6	0.92	2	0	0	100	0	0	100
#7	0.82	2	0	0	100	0	0	100
#8	0.81	4	0	0	100	20	33.33	75
#9	0.92	1	0	0	100	0	0	100
#10	0.92	7	0	0	100	0	0	100
#11	0.91	3	0	0	100	0	0	100
#12	0.87	4	0	0	100	0	0	100
#13	0.92	2	0	0	100	0	66.67	75
#14	0.82	8	0	0	100	0	0	100
#15	0.81	9	0	0	100	40	0	75
#16	0.82	7	0	0	100	0	33.33	87.5
#17	0.77	4	0	0	100	0	33.33	87.5
#18	0.92	6	0	0	100	0	33.33	87.5
#19	0.82	9	0	0	100	0	0	100
#20	0.8	9	0	0	100	0	0	100
#21	0.82	9	0	0	100	20	0	87.5
#22	0.86	2	40	0	75	0	33.33	87.5
#23	0.9	1	0	0	100	40	0	75
#24	0.92	2	0	0	100	0	33.33	87.5
#25	0.92	2	0	0	100	20	66.67	62.5
#26	0.92	3	0	0	100	40	0	75
#27	0.91	7	0	0	100	20	33.33	75
#28	0.92	1	0	0	100	0	33.33	87.5
#29	0.91	3	0	0	100	20	0	87.5
#30	0.82	1	40	0	75	60	66.67	37.5
#31	0.92	7	0	0	100	0	0	100
#32	0.8	6	0	0	100	0	0	100
#33	0.8	1	0	0	100	40	33.33	62.5
#34	0.82	8	0	0	100	20	0	87.5
#35	0.82	6	0	0	100	0	33.33	87.5
#36	0.82	5	0	0	100	0	0	100
#37	0.85	5	40	0	75	0	66.67	75
#38	0.82	4	0	0	100	0	0	100
#39	0.92	4	0	0	100	0	0	100
#40	0.8	2	0	0	100	0	66.67	75
#41	0.92	8	0	0	100	0	0	100
#42	0.91	8	0	0	100	0	0	100
#43	0.92	9	0	0	100	0	33.33	87.5
#44	0.91	9	0	0	100	20	0	87.5
#45	0.87	8	0	0	100	0	33.33	87.5
#46	0.91	8	0	0	100	0	0	100
#47	0.92	4	0	0	100	0	0	100
#48	0.82	1	0	0	100	40	33.33	62.5
#49	0.81	1	0	0	100	0	0	100

#50	0.89	9	0	0	100	0	0	100
#51	0.91	1	0	0	100	20	0	87.5
#52	0.82	3	0	0	100	0	0	100
#53	0.91	2	0	0	100	0	33.33	87.5
#54	0.83	2	0	0	100	0	33.33	87.5
#55	0.82	3	0	0	100	0	33.33	87.5
#56	0.88	5	0	0	100	0	33.33	87.5
#57	0.92	3	0	0	100	0	0	100
#58	0.89	5	40	0	75	40	66.67	50
#59	0.92	4	0	0	100	0	66.67	75
#60	0.87	2	0	0	100	40	0	75
#61	0.91	7	0	0	100	0	33.33	87.5
#62	0.92	1	0	0	100	20	33.33	75
#63	0.9	5	0	0	100	0	0	100
#64	0.81	4	0	0	100	0	0	100
#65	0.85	5	0	0	100	0	0	100
#66	0.92	3	0	0	100	60	0	62.5
#67	0.91	7	0	0	100	0	0	100
#68	0.91	8	0	0	100	0	33.33	87.5
#69	0.82	9	0	0	100	0	0	100
#70	0.82	2	0	0	100	0	0	100
#71	0.75	5	0	0	100	0	0	100
#72	0.92	1	0	0	100	0	0	100
#73	0.92	3	0	0	100	20	33.33	75
#74	0.92	1	0	0	100	0	33.33	87.5
#75	0.87	4	0	0	100	0	0	100
#76	0.9	6	0	0	100	0	0	100
#77	0.82	2	0	0	100	0	0	100
#78	0.88	9	0	0	100	0	0	100
#79	0.86	9	0	0	100	20	0	87.5
#80	0.65	3	40	0	75	20	33.33	75
#81	0.92	4	0	0	100	0	33.33	87.5
#82	0.82	8	0	0	100	0	33.33	87.5
#83	0.81	5	0	0	100	0	0	100
#84	0.9	8	0	0	100	0	0	100
#85	0.9	6	0	0	100	0	0	100
#86	0.92	7	0	0	100	20	0	87.5
#87	0.78	2	0	0	100	0	0	100
#88	0.92	9	0	0	100	20	0	87.5
#89	0.82	3	40	0	87.5	20	0	87.5

Получените стойности за средната точност, лъжливо положителните и лъжливо отрицателните отговори за невронната мрежа са съответно 98.46 %, 2.70 % и 0 %, а за kNN – 89.47 %, 8.09 % и 14.61 %. Те показват превъзходство на невронната мрежа спрямо kNN класификатора. Изчислената стойност на t-статистиката t = 5.98 показва значима разлика с вероятност 0.99.

Сравнение на резултатите с тези от други автори

Върху същата база данни за подписи SUsig е проектиран класификатор от авторите *Yanikoglu* и *Kholmatov* [3], който е оценен с 1.64% *FRR* и 1.28% *FAR*.

В [13] са тествани различни архитектури на невронни мрежи за верификация на подпис. Най-добрият получен експериментален резултат е 2% FAR и 1.3% FRR за многослоен перцептрон с един скрит слой. Във всичките експерименти са използвани по 5 подписа от всеки потребител за обучение и 41 глобални признаци на входа.

В [37] е представен класификатор *k*NN за оф-лайн подписи. Стойността на точността е оценена чрез крос валидация и е раван на 97.47%. В [38] и [39] са представени класификатори *k*NN за оф-лайн подписи, за които е постигната точност съответно от 70% и 78%.

Така представените в дисертацията класификатори имат оценки сходни с тези, които се срещат в литературата с използване на същата база данни за подписи, както и със същите по вид класификатори.

4.5 Резултати в Глава 4

От проведените експерименти за избор на най-информативно подмножество от признаци от собствената база от подписи и базата SUsig, както и за точността на верификацията, могат да се направят следните изводи:

1. Последователното използване на метода на корелационните плеяди и регресионния подход, основан на *Cp* критерия на Mallows, води до съществено намаляване на броя на необходимите признаци (приблизително два пъти).
2. За всеки от участниците в експериментите съществува специфично подмножество от признаци, което описва най-добре съвкупността от неговите подписи или с други думи, не съществува подмножество от признаци, което еднакво добре да разграничава всеки от участниците. За някои от участниците индивидуалното подмножество се състои от 3-5 признака.
3. По-значителна редукция в броя на признаците се получава при използване на неумели фалшификати (Вариант “1”) като отрицателни примери.
4. Редукцията на броя на признаците при невронните мрежи води до повишаване на точността на верификация, докато при *k*NN се наблюдава намаляване на точността.
5. Обучените с неумели фалшификати класификатори (Случай ”1”) водят до по-добри резултати в сравнение с обучените с умели и неумели фалшификати (Случай “2”).

Тези изводи водят до следните препоръки за прилагане на изложения в дисертационната работа комбиниран метод за верификация по подпис:

1) да се редуцира броя на признаците с последователно прилагане на метода на корелационните плеяди и *Cp* критерия на Mallows;

2) Да се обучат класификатори по НМ с използване на неумели фалшификати и да се избере най-точния от тях за всеки конкретен участник.

Глава 5. Софтуерно приложение

Пета глава включва следните раздели.

5.1 Бази от данни за подписи

Експериментите с класификатори за верификация по подпис са извършени върху базата данни за подписи, събрана за целта, както и върху публично достъпната база данни за подписи SUsig.

5.1.1 Събиране на подписи

Подписите се събират мобилно с помощта на лаптоп и графичен таблет *Wacom Intuos3 A5 PTZ-630* [61], който дава възможност за визуална обратна връзка и позволява писането както със стандартна писалка за таблет (*grip pen*), така и с писалка с мастило (*inking pen*), с която може да се пише по стандартен начин върху лист, поставен върху таблета като същевременно се получава и динамичната информация за всяка точка от траекторията на подписа, която се записва в създадената база данни за подписи.

Участниците в експериментите са 8 колеги от секция "ОСРО"- БАН с различна възраст и пол. Всеки от тях демонстрира индивидуален стил на подписване, така подписите на някои от участниците представляват неразбираеми щрихи, а на други четим и стегнат почерк.

Целта на протокола за събиране на подписи е описанието на процеса и детайлите по практическото събиране на образци на подписи, което се извършва с графичен таблет с помощта на разработеното за целта софтуерно приложение.

И така, броят на събраните собствени подписи е 80 (по 10 подписа от потребител), а този на умелите фалшификати - 72 (по 9 подписа за потребител).

5.1.2 Бази данни за подписи SUsig

Базата данни за подписи SUsig [24] е предоставена от университета Sabanci University – Турция. Тя се състои от две части, съответно Visual и Blind. Първата съдържа подписи, събрани с помощта на таблет за подписи с вграден екран Interlink Elec. ePad. За всеки потребител се събрани 20 собствени и 10 подправени подписа, като истинските са събрани в две сесии. Броят на участниците е 84. За събирането на подписите от втората част (Blind) е използван графичният таблет Wacom Graphire2. С него са събрани по 8 или 10 собствени и 10 подправени подписа в една сесия за всеки участник. Броят на участниците е 89. За някои от потребителите броят на събраните собствени подписи е 8, а не 10 и поради това общият брой събрани подписи е 830. За всеки от подписите, в отделен текстов файл, се съдържа информация за координатите x и y, времеви отпечатък, ниво на натиск и индикатор за начало на щрих. Въз основа на предоставените характеристики на подписите, се пресмятат времетраенето и дължината (като брой точки) на всеки един от щрихите.

5.1.3 Диаграма на базата данни за подписи SUsig

5.2 Среда за разработка

5.3 Екрани

Заключение

В настоящата дисертационна работа е предложен комбиниран метод за разпознаване на он-лайн подписи, който е реализиран в софтуерно приложение.

Разгледан е случая на верификация на подпис, поради все по-широкото му използване в обществената практика. Обхванат е целия процес по създаването на система за разпознаване по подписи с всички необходими за целта стъпки, включващи най-общо събиране на подписи, избор на признаци и създаване на модели за верификация. Изследвано е значението на типа подправени (умели и неумели) подписи както за избора на признаци, така и за самата верификация. Експериментите са проведени върху създадена за целта база от данни за подписи от осем участника, както и върху базата от данни за подписи SUsig, съдържаща подписи от 89 участника. Постигнатите резултати по време на експериментите демонстрират задоволителна точност на разпознаване от 98,46 %. Извършено е сравнение на получените резултати с класификатор НМ и класификатор по *k*-най-близки съседи. По такъв начин поставените в дисертационната работа цели задачи са изпълнени напълно.

Разработките и изследванията в областта на разпознаването на он-лайн подписи ще продължат и в бъдеще и ще бъдат насочени: 1) към усъвършенстване и подобряване на работата на системата като степен на автоматизация и точност; 2) тестването на системата върху различни бази данни и с използването на други признаци; 3) поради все по-широкото навлизане на такива системи в практиката и възможността за измами все по-често ще възниква въпросът за установяването на фалшификати. В тази връзка от голяма полза ще бъде добавянето на телевизионна камера, която ще отчита допълнително характеристиките и движението на ръката на подписващия се. Очертава се възможност и за съвместна изследователска работа с експерти от НИКК при МВР и участие в национални и международни проекти.

Авторска справка

Научно-приложните приноси са:

- Предложен е подход за минимизиране на признаково пространство чрез последователно прилагане на метода на корелационните плеяди и регресионния анализ. С прилагането му е получено минимално подмножество от признаци за всеки потребител, включен в база данни;
- Предложен е метод за създаване на класификатор като комбинация от индивидуални за всеки потребител класификатори;
- Предложен е обобщен мрежов модел на процеса за он-лайн верификация на подписи;
- Извършено е сравнение на точността на верификацията с използването на умели и неумели фалшификати в процеса на обучението;
- Извършено е сравнение на точността на верификацията по общи и индивидуални признаци.

Приложни приноси:

- Разработена е база данни и протокол за събиране на он-лайн подписи;
- Разработен е прототип на софтуерна система, даваща възможност за: добавяне на потребител, търсене на даден потребител, събиране на подписи с графичен таблет, предварителна обработка на данните за подписи, извличане на признаци, избор на признаци, визуализация на собствени и фалшиви подписи, избор на модел на класификатор и обучението му;
- Създаден е речник на част от специфичните термини в разпознаването на образи на български и английски език.

Благодарности

Изказвам своята сърдечна благодарност на научния си ръководител доц. д-р Георги Глухчев за компетентните напътствия, професионалните съвети и търпението, оказани при разработването на дисертационния труд.

Благодаря и на колегите от колектива на секция ”Обработка на изображения и разпознаване на образи” с ръководител доц. д-р Димо Димов за предоставянето на образци от подписите им, необходими за провеждане на част от експериментите.

Благодаря на съпруга и близките ми за подкрепата и разбирането, които проявиха по време на моята работа.

Работата по дисертационния труд е осъществена с подкрепата на Проект: BG051PO001-3.3.04/40 по Оперативна програма “Развитие на човешки ресурси” на ЕСФ и МОМН - Република България.

Литература

[1] Plamondon, R., Lorette, G.: *Automatic signature verification and writer identification – the state of the art*. Pattern Recognition 22, pp.107–131, 1989.
[2] J. Fierrez-Aguilar, L. Nanni, J. Lopez-Penalba, J. Ortega-Garcia, and D. Maltoni. *An online signature verification system based on fusion of local and global information*. In Proc. of IAPR Intl. Conf. on Audio- and Video-Based Biometric Person Authentication, AVBPA, pp. 523–532. Springer LNCS-3546, 2005.
[3] A. Kholmatov and B. Yanikoglu. *Identity authentication using improved online signature verification method*. Pattern Recognition Letters, 26(15):2400–2408, 2005.
[4] Chatterjee, S., Hadi, A.: *Regression Analysis by Example*. 4th Ed. New York,2006.
[5] D. Impedovo et al. *Automatic signature verification – The state of the art*, IEEE Transaction on Systems, Man and Cybernetics part C, pp. 609-635, Vol. 38, Issue 5, 2008
[6] P. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[7] Jain A., Stan Li. *Encyclopedia of Biometrics*, Springer, 2009.
[8] L. Likforman-Sulem, S. Garcia-Salicetti, J. Dittmann, J. Ortega-Garcia, N. Pavesic, G. Gluhchev, S. Ribaric, and B. Sankur, *Report on the hand and other modalities stated of the art*, Biometrics for Secure Authentication, 2005 .
[9] LaMotte, L.R., Hocking, R.R.: *Computational Efficiency in the Selection of Regression Variables. Technometrics.* 12, pp. 83-93, 1970.
[10] Boyadzhieva D., Gluhchev G., *Feature Set Selection for On-line Signatures using Selection of Regression Variables*, In: Proceedings of the 4th International Conference on Pattern Recognition and Machine Intelligence PReMI’11, pp. 440-445, 2011.
[11] Neural Network Toolbox™ 7
[12] G. P. Zhang, *Neural networks for classification: a survey,* Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 30, pp. 451-462, 2000.
[13] McCabe, Alan, Trevathan, Jarrod, and Read, Wayne. *Neural network-based handwritten signature verification*. Journal of Computers, 3 (8). pp. 9-22, 2008.
[14] Hocking, R.R.: *The Analysis and Selection of Variables in Linear Regression*. Biometrics. 32, 1976.
[15] Hocking, R.R., Leslie, R.n.: *Selection of the Best Subset in Regression Analysis*., Technometrics. 9,pp. 531-540, 1967.
[16] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, *Regression analysis by example,* 4th ed, Wiley, 2007.
[17] Mallows, C.L. *Some Comments on CP*. Technometrics 15 (4): pp. 661–675, 1973.
[18] http://www.microsoft.com/visualstudio/en-us/products/2010-editions/express
[19] http://www.microsoft.com/sqlserver/en/us/editions/compact.aspx
[20] http://www.microsoft.com/download/en/details.aspx?displaylang=en&id=20039
[21] http://www.wacom-asia.com/intuos3/intuos3_630.html

[22] http://www.mathworks.com/products/matlab/
[23] http://www.mathworks.com/matlabcentral/answers/22944-neural-network-design
[24] Alisher Kholmatov, Berrin A. Yanikoglu: *SUsig : an on-line signature database, associated protocols and benchmark results*. Pattern Anal. Appl. 12(3): 227-236, http://biometrics.sabanciuniv.edu/SUsig, 2009 .
[25] LeCun, L. Bottou, G. Orr and K. Muller: *Efficient BackProp*, in Orr, G. and Muller K. (Eds), Neural Networks: Tricks of the trade, Springer, 1998.
[26] Dougherty G., *Pattern Recognition and Classification*, Springer-Verlag New York Inc. ISBN: 9781461453222.
[27] Kuncheva L.I. *Combining classifiers: Soft computing solutions*, in: S.K. Pal and A. Pal (Eds.) Pattern Recognition: From Classical to Modern Approaches, World Scientific Publishing Co., Singapore, pp.427-452, 2001.
[28] Ink Data: http://msdn.microsoft.com/en-us/library/ms811395.aspx
[29] Fawcett, T. *ROC graphs: Notes and practical considerations for researchers*, Tech Report HPL-2003-4, HP Laboratories. http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf,2003.
[30] С.А. Айвазян, З.И. Бежаева, О.В. Староверов, Классификация многомерных наблюдений, Москва, Статистика, 240 стр., 1974.
[31] Gluhchev G., M. Savov, O. Boumbarov, D. Vassileva. *A New Approach to Signature Based Authentication*, 2nd Int. Conf. on Biometrics, Seoul, pp. 594-603, 2007.
[32] Savov M., G. Gluhchev. *Signature verification via Hand-Pen motion investigation*, Proc. Int. Conf. "Recent Advances in Soft Computing", Canterbury, 2006, pp. 490-495.
[33] Nalwa, V.: *Automatic on-line signature verification.* Proc. IEEE 85, 215–239 , 1997.
[34] J. Richiardi, H. Ketabdar, A. Drygajlo, *Local and Global Feature Selection for On- line Signature Verification*, IEEE 8th International Conference on Document Analysis and Recognition (ICDAR'05), vol.2, pp. 625 – 629, 2005.
[35] Atanassov, K., G. Gluhchev, S. Hadjitodorov, A. Shannon, V. Vassilev*, Generalized Nets and Pattern Recognition*. KvB Visual Concepts Pty Ltd, Monograph No. 6, Sydney, 2003.
[36] *A generalized net model of the process of handwriting identification*. In: [35].
[37] Schmidt T., Riffo V., Mery D.: *Dynamic Signature Recognition Based on Fisher Discriminant*. CIARP, volume 7042 of Lecture Notes in Computer Science, Springer, pp.433-442, 2011.
[38] Abdulla Ali, *Offline Signature Verification using Radon Transform and SVM/kNN Classifiers*, ISSN 0136-5835, Вестник ТГТУ, Том 15. № 1, 2009.
[39] Meenakshi K., Sargur S., Aihua X., *Offline Signature Verification And Identification Using Distance Statistics*, In: International Journal of Pattern Recognition and Artifficial Intelligence, Vol. 18, No. 7, pp.1339-1360, 2004.
[40] Gluhchev G., M. Savov, O. Boumbarov, D. Vassileva. *A New Approach to Signature Based Authentication*, 2nd Int. Conf. on Biometrics, pp. 594-603, 2007.
[41] Savov M., G. Gluhchev. *Signature verification via Hand-Pen motion investigation*, Proc. Int. Conf. Recent Advances in Soft Computing, pp. 490-495, 2006.
[42] Berrin A. Yanikoglu, Alisher Kholmatov: *Online Signature Verification Using Fourier Descriptors*. EURASIP J. Adv. Sig. Proc., 2009.

Abstracts of Dissertations

Number 1, 2014

INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES

BULGARIAN ACADEMY OF SCIENCES

БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ

ИНСТИТУТ ПО ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Брой 1, 2014

Автореферати на дисертации