UNITECH '10 GABROVO

INTERNATIONAL SCIENTIFIC CONFERENCE
19 – 20 November 2010, GABROVO

AN APPROACH TO FEATURE SELECTION FOR ON-LINE SIGNATURES

Desislava Boyadzieva
IICT-BAS

Georgi Gluhchev
IICT-BAS

Abstract

The selection of reliable and proper feature set is an important step in creating a reliable signature verification system. This paper presents an approach to feature set selection for signatures acquired by a graphical tablet. Twenty features are first extracted from 160 on-line signatures acquired from sixteen users by a GUI designed for that purpose. After that the redundant features are identified trough an analysis of the histogram built for significant correlation coefficients of the feature correlation matrices of the users.

Keywords: Feature selection, On-line signatures, Correlation analysis, Graphical tablet, Signature verification, Histogram.

INTRODUCTION

Signatures are recognized and accepted modality for authentication. Signature verifycation is the process of confirming the identity of a user based on his/her handwritten signature as a form of behavioral biometrics [1]. Generally, the creation of a signature verification system passes through the following steps: signature acquisition, feature extraction, feature selection and creation of a classification criterion for verification and error evaluation. There are two methods for signature acquisition: off-line and on-line [2]. The off- line method uses a captured image of a written signature after the writing process is over while the on-line method uses devices such graphical tablets to capture signature during signing and thus a lot of writer specific features like pressure, speed, pen tilt, azimuth, etc. are available.

Generally, signature features can be [4] local, where one feature is extracted for each sample point in the signature, global features, where one feature is extracted for a whole signature, and segmental features, where the signature is subdivided into segments and one feature is extracted for each segment. Over 100 features used in signature verification are listed in [2].

Some features demonstrate more discriminatory capability than others and some feature selection should be performed after feature extraction. Feature selection is defined as the process of selecting k most discriminatory features 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 on the work on feature set selection for signatures is performed in [4].

In this paper we identify the redundant features trough an analysis of the histogram built for significant correlation coefficients of the feature correlation matrices for the sixteen users of our database.

The rest of the paper is structured as follows: Section 2 considers the proposed approach to feature selection and Section 3 draws some conclusions.

AN APPROACH TO FEATURE SELECTION FOR ON-LINE SIGNATURES

SIGNATURE ACUISITION

We create a GUI using C# programming language and .NET Tablet PC SDK 1.7 library to facilitate signature acquisition and feature extraction (Figure 1). All signatures are acquired by using a digitizing tablet Wacom Intuos3 A5 PTZ-630 having a resolution of 5080 lines per inch and an acquisition area of the pad of 152.4 x 210.6 mm and sampling rate of 200 points per second. Each of the sixteen users sign ten times by using an inking pen on a sheet of paper placed on the acquisition area of the graphical tablet. Thus, the GUI facilitate users since allow them to sign in a natural way and to have visual control over the process of signing. The signature database for our experiment contains 160 signatures collected from sixteen users.

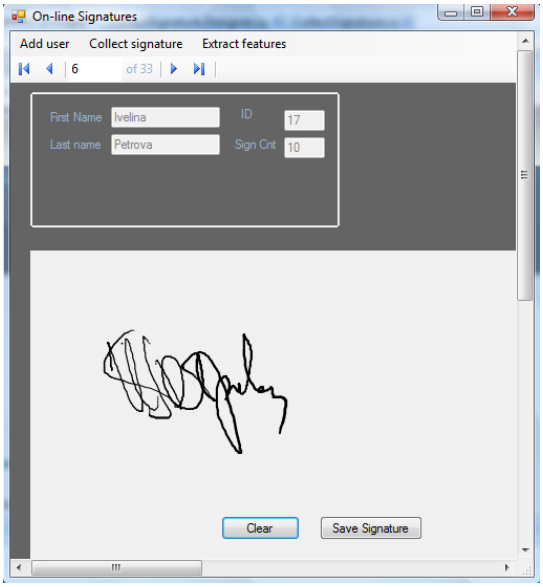


Figure 1 *GUI for signature acquisition*

FEATURE EXTRACTION

All the features used in our work are listed in Table 1.They are evaluated on the basis of the raw signature data obtained using the software program.

The following formulas are used for computing part of the features:

The distance d between two points in a signature (resp. (x_1, x_2) and (y_1, y_2)) is given by:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}. \qquad (1)$$

The angle φ of a line between two points in a signature (resp. (x_1, x_2) and (y_1, y_2)) is given by:

$$\varphi = arctg\left(\frac{(y_2 - x_2)}{(y_1 - x_1)}\right). \qquad (2)$$

The center point $(\bar{x}, \bar{y})$ of a signature containing *n* points is given by:

$$(\bar{x}, \bar{y}) = (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i). \qquad (3)$$

Considering the *j*th user $j, j = 1,...,k$ of the database let us denote the number of his/her signatures with n ($n = 10$) and the number of their features with p ($p = 20$). Thus, each feature X_i^j can be considered as an *n*-dimensional vector of the form:

$$X_i^j = [X_{i1}^j, ..., X_{in}^j], i = 1,..., p. \qquad (4)$$

A1	width	*A11*	angle of the line between center and end points
A2	height	*A12*	angle of the line between initial and end points
A3	height to width ratio	*A13*	distance between leftmost and center points
A4	number of points	*A14*	distance between center and rightmost points
A5	time duration	*A15*	angle of the line between center and leftmost points
A6	signature density =A4/A1*A2	*A16*	angle of the line between center and rightmost points
A7	distance between initial and center point	*A17*	distance between leftmost and initial points
A8	distance between end and center point	*A18*	distance between rightmost and end points
A9	distance between initial and end point	*A19*	angle of the line between leftmost and initial points
A10	angle of the line between center and initial points	*A20*	angle of the line between end and rightmost points

Table 1*Signature features*

FEATURE CORRELATION MATRIX

Correlation is a useful quantity for measuring both the direction and the strength of the relationship between two variables [3]. The value of the correlation is determined by the correlation coefficient. The most widely-used type of correlation coefficient is Pearson *r* and for variables X_1 and X_2 (being *p*-dimensional vectors) it is defined as follows:

$$r_{X_1,X_2} = \frac{\sum_{i=1}^{p}(X_{1i} - \overline{X_1})(X_{2i} - \overline{X_2})}{\sqrt{\sum_{i=1}^{p}(X_{1i} - \overline{X_1})^2 \sum_{i=1}^{p}(X_{2i} - \overline{X_2})^2}}. \qquad (5)$$

The mean $\overline{X_j}, j = 1,2$ is defined as follows:

$$\overline{X_j} = \sum_{i=1}^{p} X_{1i}, j = 1,2. \quad (6)$$

The magnitude of correlation coefficient measures the strength of the linear relationship between X_1 and X_2. The closer r_{X_1,X_2} is to 1 or -1, the stronger is the relationship between X_1 and X_2. It is important to examine the scatter plot of X_1 versus X_2 before interpreting the numerical value of r_{X_1,X_2} since: (a) the correlation coefficient can be substantially influenced by one or few outliers in the data, and (b) the correlation coefficient measures only linear relationships [3].

Information about the reliability of each correlation is given by the corresponding significance level which depends on the size of the sample [3]. In our experiment this level is 0.632 since the number of collected signatures is ten ($n = 10$).

For each user $j, j = 1,...,k$ of the database, features correlation matrix can be represented in the following way:

$$C^j = \{C_{kl}^j\}, k, l = 1,..., p. \quad (1)$$

Each off-diagonal element $C_{kl}^j, k \neq l$ of this matrix is called pairwise correlation coefficient between feature k and feature l for the particular user j.

FEATURE SELECTION

First, the correlation matrix on all features is computed for every user of the database and after that all the significant pairwise correlation coefficients at a level of 5% are extracted. Afterwards these significant correlation coefficients are counted on all users and the corresponding histogram is created. On the basis of that histogram it is decided which features to be excluded from the initial feature set. We create a histogram according to the procedure described in section 2.4. Part of the histogram for highly correlated features which number is greater than 4 (i.e. occur in more than 25% of users) is displayed on Table 2.

On the basis of that histogram and the pairwise scatter plots, we made the following conclusions: (a) pairs *A4;A5* and *A3;A6* are high correlated and none of these features is correlated with other feature so we can exclude for example *A5* from the first pair and *A6* from the second pair; (b) features *A1, A9, A13, A14, A18* are related each other so we decide to exclude *A1, A9, A13, A14*. Finally, six features respectively labeled with *A1*, *A9*, *A14*, *A5*, *A6*, and *A13* are excluded from the initial feature set.

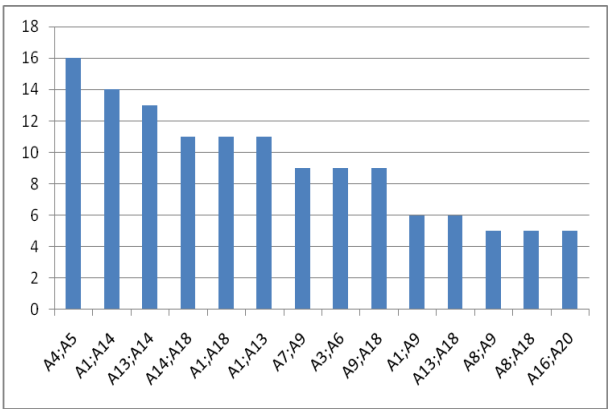


Figure 2 *Histogram*

In this paper we present a GUI which simplify the signature acquisition process and automate signature feature extraction. Also, we describe our approach to feature set selection and conduct an experiment in order to test it. The obtained results are good and they demonstrate that independently of the small size of the used database, it is reasonable to remove six out of the twenty features.

The focus of our future work is to apply the described procedure on a larger signature database and to investigate the problem further in order to find the most discriminating features to be used in signature verification.

ACKNOWLEDGEMENT

This work is supported by the European Social Fund and Bulgarian Ministry of Education, Youth and Science under Operative Program "Human Resources Development", Grant BG051PO001-3.3.04/40.

REFERENCE

[1] V.S. Nalwa, "Automatic on-line signature verification", Proceedings of the IEEE, vol. 85, no. 2, pp. 213-239, Feb. 1997.

[2] 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.
[3] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, Regression analysis by example, 4th ed, Wiley, 2007.
[4] Jonas Richiardi, Hamed Ketabdar, Andrzej Drygajlo, "Local and Global Feature Selection for On-line Signature Verification," icdar, pp.625-629, Eighth International Conference on Document Analysis and Recognition (ICDAR'05), 2005.