Features preferred in-identification system based on computer mouse movements

Roman Jašek and Martin Kolařík

Citation: AIP Conference Proceedings **1738**, 120024 (2016); doi: 10.1063/1.4951907

View online: http://dx.doi.org/10.1063/1.4951907

View Table of Contents: http://aip.scitation.org/toc/apc/1738/1

Published by the American Institute of Physics

Features Preferred in -Identification System Based on Computer Mouse Movements

Roman Jašek and Martin Kolařík

Tomas Bata University in Zlin, Faculty of Applied Informatics, Nám. T. G. Masaryka 5555, 760 01 Zlín, Czech Republic

Abstract. Biometric identification systems build features, which describe people, using various data. Usually it is not known which of features could be chosen, and a dedicated process called a feature selection is applied to resolve this uncertainty. The relevant features, that are selected with this process, then describe the people in the system as well as possible. It is likely that the relevancy of selected features also mean that these features describe the important things in the measured behavior and that they partly reveal how the behavior works. This paper uses this idea and evaluates results of many runs of feature selection runs in a system, that identifies people using their moving a computer mouse. It has been found that the most frequently selected features repeat between runs, and that these features describe the similar things of the movements.

Keywords: identification system, relevant feature, feature selection, behaviometric, computer mouse

INTRODUCTION

In order to identify people, behaviometric identification systems utilize features that characterize people's behavior. Usually the behavior means a selected single action or process that an identified person performs, and the features mean measured or calculated quantities that describe this action or process [1]. The used process or action is a complex result of a mixture of many sub-actions driven with many individual factors, which altogether are almost impossible to be imitated. On the other hand, this complexity makes it also almost impossible to separate the result into the factors. Any measured feature contains a part of the complete information but it is not known how well this particular part describes the observed process or action [2].

This difficulty is usually overcome with applying a two step process during the system's design phase: in the first step a bigger number of features is calculated in the feature extraction process, and in the second step a smaller number of features is then selected using the feature selection process [3]. The feature selection process selects features that are relevant and descriptive. The relevancy for the identification is linked to general relevancy of the feature. If the feature is convenient for distinguishing people, it is also likely convenient for describing the measured action or process, and consequently it suggests how the action could work.

This paper is aimed at evaluating features that are selected in an identification system based on a computer mouse movements. These movements are measured via the mouse coordinates available in the operating system, and are gradually reduced into approx. one hundred of various features.

The paper is divided into two parts: in the first part methods used to build the selected features are briefly described, and in the second part the features that are preferably selected by the feature selection are evaluated and discussed. As a result it has been found that the features related to turning and sudden changes in movements are selected more frequently than the other features.

METHODS

Obtaining features is described in three steps—detection of moves, the feature extraction and the feature selection.

Detection and Creation of Movement Units

The mouse coordinates come into the computer as a continuous stream of events interleaved with small delays. This does not correspond to the way how a person uses the mouse: typically there are longer delays separating individual

International Conference of Numerical Analysis and Applied Mathematics 2015 (ICNAAM 2015)
AIP Conf. Proc. 1738, 120024-1–120024-4; doi: 10.1063/1.4951907
Published by AIP Publishing. 978-0-7354-1392-4/$30.00

TABLE 1. Quantities and their calculation scheme

${}_ix$			${}_iy$			${}_i\mathrm{d}t^*$	=	${}_{i+1}t - {}_it$
${}_i\mathrm{d}x^*$	=	${}_{i+1}x - {}_ix$	${}_i\mathrm{d}y^*$	=	${}_{i+1}y - {}_iy$	${}_i\mathrm{d}s^*$	=	$\sqrt{{}_i\mathrm{d}x^2 + {}_i\mathrm{d}y^2}$
${}_iv_x^*$	=	$\frac{{}_i\mathrm{d}x}{{}_i\mathrm{d}t}$	${}_iv_y^*$	=	$\frac{{}_i\mathrm{d}y}{{}_i\mathrm{d}t}$	${}_iv$	=	$\sqrt{{}_iv_x^2 + {}_iv_y^2}$
${}_i\mathrm{d}v$	=	$\frac{{}_{i+1}v - {}_iv}{{}_i\mathrm{d}t}$	${}_i\mathrm{d}^2v$	=	$\frac{{}_{i+1}\mathrm{d}v - {}_i\mathrm{d}v}{{}_i\mathrm{d}t}$			
${}_i\phi_{vxy}^*$	=	$\mathrm{atan}({}_iv_y, {}_iv_x)$	${}_i\phi_v$	=	${}_{i+1}\phi_{vxy} - {}_i\phi_{vxy}$	${}_iv_n$	=	$\sin{}_i\phi_v$
${}_i\omega$	=	$\frac{{}_i\phi_v}{{}_i\mathrm{d}t}$	${}_i\mathrm{d}\omega$	=	$\frac{{}_{i+1}\omega - {}_i\omega}{{}_i\mathrm{d}t}$			
${}_ia_x^*$	=	$\frac{{}_iv_x}{{}_i\mathrm{d}t}$	${}_ia_y^*$	=	$\frac{{}_iv_y}{{}_i\mathrm{d}t}$	${}_ia$	=	$\sqrt{{}_ia_x^2 + {}_ia_y^2}$
${}_i\phi_{axy}^*$	=	$\mathrm{atan}({}_ia_y, {}_ia_x)$	${}_i\phi_a$	=	${}_{i+1}\phi_{axy} - {}_i\phi_{axy}$	${}_ia_n$	=	$\sin{}_i\phi_a$
${}_ic_s$	=	$\frac{{}_{i+1}\phi_v - {}_i\phi_v}{{}_i\mathrm{d}s}$	${}_i\mathrm{d}c_s$	=	$\frac{{}_{i+1}\mathrm{d}c_s - {}_i\mathrm{d}c_s}{{}_i\mathrm{d}s}$	${}_ic_a$	=	$\frac{{}_{i+1}\phi_a - {}_i\phi_a}{{}_i\mathrm{d}s}$
$\mathrm{d}x_d^*$	=	${}_{\mathrm{last}}x - {}_0x$	$\mathrm{d}y_d^*$	=	${}_{\mathrm{last}}y - {}_0y$	s_d	=	$\sqrt{\mathrm{d}x_d^2 + \mathrm{d}y_d^2}$
s_I^*	=	$\sum_i {}_i\mathrm{d}s$	t_I	=	$\sum_i {}_i\mathrm{d}t$	c_{sI}	=	$\sum_i {}_i\mathrm{d}c_s$
v_d	=	$\frac{s_d}{t_I}$	r	=	$\frac{1 - s_d}{s_I}$	j	=	$\frac{s_{UI}^*}{s_I}$

movement actions that have single purpose (e.g. to move the mouse to a button). Therefore the first step of the reducing the data is a detection of these movement actions. According to [4], these actions are called strokes.

The identification system used in this paper detects the strokes with measuring a delay between consecutive input stream events: if the delay is longer than a chosen threshold, the preceding piece of the input stream is considered to be a stroke. For the purpose of this paper the threshold used was 1.0 s.

The movement is smooth but the stroke is not. The main reason for this is that the tracking of the mouse position has limited resolution, both space and time. In order to get rid of these imperfections the detected strokes are smoothed and re-sampled. The smoothing is realized by the smoothing spline [5], which is then sampled to obtain three times more points than the source stroke had.

Creation of Features—Feature Extraction

The obtained strokes are then processed to get quantities describing each individual stroke in three steps: in the first step quantities relating to each stroke's event are computed, in the second step the vectors of quantities' values are reduced to statistical markers, and in the third step these statistical markers are modeled with a random variable.

The stroke is composed of events containing only x and y coordinates and a time instant t. These three quantities are recalculated into all other quantities as it is summarized in table 1. The quantities with an asterisk in the exponent are auxiliary, they only show how calculations advance. Two kinds of quantities are used: the first kind having index is a vector quantity, and the second kind without the index means the quantity is a scalar. Index i that indexes events in the stroke is put to the left side of the quantity symbol to avoid confusing with quantities' proper indexes.

In the second step, for each stroke, the following statistical markers are computed for each vector quantity: the maximum M, the minimum m, the average A, the deviation D and the spread (S = M-m).

In the third step, the values of each marker across all strokes are considered to be samples of an unknown probability density function, which is modeled with a random variable [6]. The random variable is then called the feature.

The overall result of the feature extraction is that many strokes (containing many events) of a particular person are reduced to 89 random variables having one or two parameters.

Feature Selection

The number of features (89) is big and the features share substantial amount of information in an unknown way. In order to eliminate this shared information, and also to select the most relevant features, the feature selection is utilized. The used feature selection runs SFFS [3] algorithm and compares a candidate combinations using d_s metric [7]. The feature selection is driven with randomly selected strokes, so its results vary between different runs.

TABLE 2. Numbers of how many times features were selected

	x	y	v	dv	d^2v	ϕ_v	v_n	ω	dω	a	ϕ_a
m = minimum	—	—	47	19	23	2	23	3	29	38	0
A = average	40	24	0	14	38	6	47	7	18	0	1
M = maximum	30	18	42	12	55	11	31	2	50	4	71
D = deviation	—	—	5	—	16	21	4	14	3	2	40
S = spread	5	18	4	—	3	11	5	10	12	5	1
$\sum$	**75**	60	**98**	45	**135**	51	**110**	36	**112**	49	**113**
$\eta = \frac{1}{n}\sum$	25	20	20	15	**27**	10	22	7.2	22	10	23

	a_n	c_s	dc_s	c_a	s_d	t_I	c_{sI}	v_d	r	j
m = minimum	72	0	0	0	—	—	—	—	—	—
A = average	23	14	14	0	48	34	57	15	58	68
M = maximum	23	18	64	1	—	—	—	—	—	—
D = deviation	23	9	—	13	—	—	—	—	—	—
S = spread	2	13	36	6	—	—	—	—	—	—
$\sum$	**143**	54	**114**	20	48	34	57	15	58	68
$\eta = \frac{1}{n}\sum$	**29**	11	**29**	4.0	**48**	**34**	**57**	15	**58**	**68**

TABLE 3. Comparison of used statistical markers

	m = minimum	A = average	M = maximum	D = deviation	S = spread
$\sum$	256	246	432	150	131
$\xi = \frac{1}{n}\sum$	19.7	16.4	**28.8**	13.6	9.36

For the purpose of this paper, the feature selection was run 40 times for each of 16 people, what in total produced 640 results of the feature selection available for experiments.

RESULTS AND DISCUSSION

The results of all 640 runs are summed in table 2. Values represent how many times the particular feature (e.g. deviation of v) appeared in all 640 runs. The last line ($\eta = \frac{1}{n}\sum$) shows average number of appearances aggregated over the particular quantity. The value η better represents the contribution of the feature, because it equalizes different counts of markers that the individual features have. Dashes in the table represent unused features (that were impossible to be modeled with a random variable), and the scalar features which have no markers.

Values of η printed in bold are above the average, which is $\bar{\eta} = 26.5$. Three groups of these features can be recognized: static values s_d (direct distance) and t_I (total stroke time), d^2v (jerk), and values a_n, dc_s, c_{sI}, r and j, which relate to turning and smoothness of the path. These features are the most frequent. It could be concluded that:

- these features are well suitable to describe the mouse movements,
- these features makes the biggest difference among involved people.

The table 3 displays the frequency of appearance of each of used statistical markers. Extreme values m and M are the most frequent, with the M winning obviously. It could be concluded that people primarily differ in the limiting values of the moves, e.g. that each person has remarkably different maximal jerk (d^2v) or curvature difference (dc_s).

CONCLUSION

Identification systems that are based on people behavior work with a complex results of measuring a pre-selected activity, like a gait or a daily routine. The measured data contains enough information to identify people, but it is difficult to decide which measurement or feature is relevant and which is not. The reason for this is that the structure of the data is unknown and complex.

The feature selection process can discover relevant features, so it can help: at first, the selection of features improves the identification system, because it makes the system simpler, running faster, and working better as the features distinguish between people as well as possible. At second, it can be concluded, that features that well distinguish between people, could also well describe the relevant part of the observed behavior, and, consequently, that these features could point to the mechanism or principle how the behavior is realized and controlled by the people.

According to these ideas, this paper focuses on evaluation of features that are selected inside an identification system based on analyzing computer-mouse movements. The process of reducing source input data (coordinates x, y and a time t) up to features that best describe the individual person is outlined first, and then results of many feature selections are summarized and evaluated.

In order to obtain enough statistics, the feature selection was run 640 times (40 times for 16 people) in the system. In each run the set of selected features was remembered. It has been revealed that:

- the selected features are related to turning and smoothness of movements,
- the features are typically extremes of measured quantities.

Deductions of these observations are:

- the features that carry the most information about people are related to turning and smoothness of the movement,
- the biggest difference in people can be seen in the extremes of the quantities (like in the maximal jerk).

The overall result of this paper is as follows:

- the evaluation of behavioral features selected for the identification in the system may reveal information about how the behavior works,
- the analysis done on real data supported this idea.

The knowledge about which features are preferably selected in identification systems can further help with improving identification systems at all, and it can also help with better understanding of fundamentals of identifying people in general.

ACKNOWLEDGEMENT

This work was supported by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089 and by the Bio-Inspired Methods: research, development and knowledge transfer project, reg. no. CZ.1.07/2.3.00/20.0073 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic.

REFERENCES

1. Jain, A. K., Ross A., 2008. *Handbook of Biometrics*. Springer, ISBN 978-0-387-71040-2.
2. Schuckers, M. E., 2010. *Computational Methods in Biometric Authentication*. Springer, ISBN 978-1-84996-202-5.
3. Guyon I., Gunn S., Nikravesh M., Zadeh L. A., 2006. *Feature Extraction*. Springer, ISBN 978-3-540-35487-1.
4. Gamboa H., Fred A., 2003. An identity authentication system based on human computer interaction behaviour. *3rd workshop on pattern recognition in information systems PRIS 2003*. Vol. 3, pp. **49**–**55**.
5. Pollock, D. S. G., 1993. *Smoothing with Cubic Splines.* London, Queen Mary University of London.
6. Johnson, N. L., Kotz S., Balakrishnan N., 1994. *Continuous Univariate Distributions*. John Wiley, ISBN 978-0-471-58495-7.
7. Kolařík, M., Jašek R., 2015. *Fast method of comparing performance of identification systems*. Paper submitted to Expert Systems With Applications Journal, Elsevier.