Is on-line handwriting gender-sensitive? what tells us a combination of statistical and machine learning approaches

Laurence Likforman-Sulem, Gennaro Cordasco, Anna Esposito

Is on-line handwriting gender-sensitive? what tells us a combination of statistical and machine learning approaches

Laurence Likforman-Sulem[1], Gennaro Cordasco[2], and Anna Esposito[2]

[1] Telecom Paris/Institut Polytechnique de Paris, Palaiseau, France
[2] Department of Psychology, Università degli Studi della Campania "L. Vanvitelli", and International Institute for Advanced Scientific Studies (IIASS), Italy

Abstract. Handwriting is an everyday life human activity. It can be collected off-line by scanning sheets of paper. The resulting images can then be processed by a computer-based system. Thanks to digitizing tablets, handwriting can also be collected on-line. From the collected raw signals (pen position, pressure over time), the dynamics of the writing can be recovered. Since handwriting is unique for each individual, it can be considered as a biometric modality.
Biometric systems predicting gender from off-line handwriting, have been recently proposed. However we observe that, in contrast to other modalities such as speech, it is not straightforward for a human being (even expert) to predict gender. In this study we explore the limits of automatic gender prediction from on-line handwriting collected from a young adults population, homogeneous in terms of age and education. In our previous work [1], a statistical analysis of on-line dynamic features has shown differences between male and female groups. In the present study, we provide these features to a classifier, based on a machine learning approach (SVMs). Since datasets are relatively small (240 subjects), several evaluation frameworks are explored: cross validation (CV), bootstrap, and fixed train/test partitions. Accuracies obtained from fixed partitions range from 37% to 79%, while those estimated by CV and bootstrap are around 60%. This shows to our opinion the limits of the gender recognition task from on-line handwriting, for our observed young adult population.

1 Introduction

Handwriting is an everyday life human activity used from centuries for registering counts, communicating ideas, writing/copying books, sending letters or encrypted messages [2]. Many skills are involved in handwriting: gross and fine motor skills, ability to plan, eye-hand coordination [3]. Character models, learnt at school and related to an era and a geographical location, are also influencing the writer. Personal motor characteristics combined with the character models the writer has in mind, result in an unique handwriting [4].

In order to be processed by a computer-based system, handwriting must first be digitized, off-line or on-line. Off-line handwriting is collected by scanning sheets of paper. This results into images that can be processed automatically for tasks such as recognition or authentication. On-line handwriting is collected thanks to digitizing tablets. Applications range from creating/correcting documents, authenticating signatures, education (learning to write)[5] but also to detect health issues: neurological disorders, or upcoming strikes [6, 7].

Systems that use off-line handwriting for predicting gender, have been recently proposed [8–10]. These can be useful for various domains and tasks such as author profiling [11], forensics, and biometry. However we observe that, in contrast to other modalities such as speech, it is not straightforward for a human being (even expert) to predict gender from off-line handwriting [12, 13]. In our collective culture, women are assumed to write more legibly, with embellishments, using round shapes. Men are assumed to put more pressure and to use spiky shapes [14]. Indeed explanations on gender differences in handwriting, change and are related to the ideas promoted in an era, about the social condition of women. In our era, in countries such as Italy, girls and boys are taught to write together. One can hardly find an argument that would justify a difference due to gender. More often samples displayed as "male" and "female" are not so convincing (see for instance [9] [15] and our own samples in Fig. 3).

However, at school, girls are known to be more proficient in writing, than boys [16]. Writing speed, less time spent in air, are signs of writing proficiency. Features related to speed may thus be proposed for gender prediction.

Indeed a statistical analysis of features extracted from handwriting can highlight differences between male and female groups [1]. The feature means are found distinct for each group, but feature distributions may overlap so that one can hardly predict gender, from a machine learning point of view. To cope with overlapping distributions, we use a combination of several features selected by a statistical analysis and feed them as input to an SVM classifier.

In the literature, gender recognition systems provide accuracies ranging from 60% to 80% for fixed training/test partitions. A common train/test partition consists in taking 70% of the samples for training, the remaining ones for testing. For small sets, using fixed train/test partitions may over or under-estimate performance. More robust frameworks exist to evaluate performance, such as cross validation and bootstrap.

In the following, data collection and extracted features are described in Sections 2.1 and 2.2. A statistical analysis [1] was conducted in order to select the most discriminant features with respect to female and male groups. The outcomes of this analysis are recalled in Section 2.3. In Section 3, experiments are conducted with the SVM classifier and the selected features. We compare the accuracies obtained according to the evaluation framework (cross validation, bootstraping, fixed train/test partition). We conclude this study in Section 4 on the possibility of distinguishing gender for the observed population, from handwriting dynamics.

2 Data collection and extracted features

2.1 Data collection

Handwriting samples are collected by a digitizing INTUOS WACOM series 4 tablet associated with a dedicated writing pen named Intuos Inkpen. Participants write on a sheet of paper (normal paper) laid on the tablet. The tablet records each 8 ms, the following values; (x,y) positions of the pen, pen inclinations (azimuth, altitude), pressure of the pen on paper, time in milliseconds since the UNIX epoch (January 1, 1970 00:00:00 UTC), and the pen status (on paper=1 or in-air=0). It also records these values when the pen is in-air, close to the tablet. The tablet thus collects seven raw signals, one for each type of values. Fig. 2.1 shows recorded signals. The null values in the pressure signal correspond to in air movements.

The dataset includes the handwriting samples of 240 subjects. The two groups (male, female) are balanced by age, and level of education: 126 males (mean age 24.65 years old, SD=2.45) and 114 females (mean age =24.51 years old, SD=2.50), SD being the Standard Deviation. The subjects were volunteers recruited at University of Campania "Luigi Vanvitelli' in Caserta (south Italy). All subjects are right-handed and were asked to perform seven handwriting tasks: (1)drawing of two-pentagons (2) drawing of a house (3) writing of the following four Italian words in capital letters (BIODEGRADABILE, FLIPSTRIM, SMINUZZAVANO, CHIUNQUE); (4) drawing loops with left hand (5) drawing loops with right hand; (6) drawing a clock (7) writing the following Italian sentence in cursive letters (*I pazzi chiedono fiori viola, acqua da bere, tempo per sognare* meaning Crazy people are seeeking for purple flowers, drinking water and dreaming time).

Fig. 2 shows the pen position (x,y) on paper (black points) and in air (red points). The positions of the pen when the tablet is too far are not recorded. But the time spent far from the tablet can be recovered through the time-stamp raw signal. Henceforth, we will name this pen status as idle.

Figure 3 shows four samples of the copy task of a given sentence (task 7). It can be noted that guessing gender is not straightforward [3] [4]

2.2 Extracted features

From the raw signals collected by the tablet, features can be extracted in order to represent the samples in a concise way, adapted to a machine learning approach. Several features were computed at global level, for each task. These are [1]:

- time: the total time elapsed for the task, the total time spent in each pen status: in air, on paper, and idle (Total, tUp, tDown, tIdle)
- pressure: statistics about the pressure (minimum, maximum, mean, standard deviation, lower 10th and 90th percentiles),

[3] · gender ground truth: the first and third samples are from women, the other ones from men.

[4] ·the other ones from men (from top to bottom).

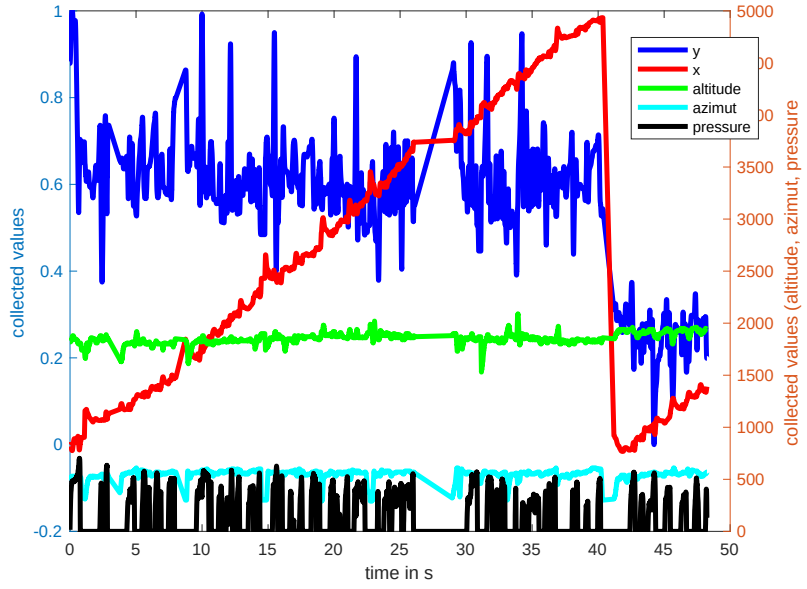

Fig. 1. Raw signals captured by the Wacom digitizing tablet when copying the task7 sentence (Fig. 2). The x and y curves have been rescaled to better fit the figure.

- ductus: the number of strokes in each pen status, number of in-air strokes, number of on-paper strokes (nbUp, nbDown).
- slope: the mean inclination of the straight lines passing through the diagonals of the axis-aligned bounding boxes containing the strokes (slopeA).
- space: based on the area occupied by axis-aligned bounding boxes containing the strokes (spaceA) and the distance between consecutive strokes (spaceT). Only on-paper traits are considered.

2.3 Selected features

From these features, a small number were found significant according to gender, by statistical analysis: Anova, Manova or logistic regression. For such features, a difference in the means of each group (men/women), can be observed with a significant level measured by a p-value smaller than 0.05. Considering task7 only, a subset of features could be selected: nbDown7, nbUp7, Total7, spaceA7, slopeA7, tUp7.

In Fig. 5, are shown the boxplots of the total time spent to complete task 7 (Total7), and the number of on paper strokes (nbDown7). Total7 is the whole time spent (on-paper, in-air or idle) when completing the seventh task. According to the means: men would write the given sentence more slowly in average (37s) than women (28s). For both features, the boxplots corresponding to men and

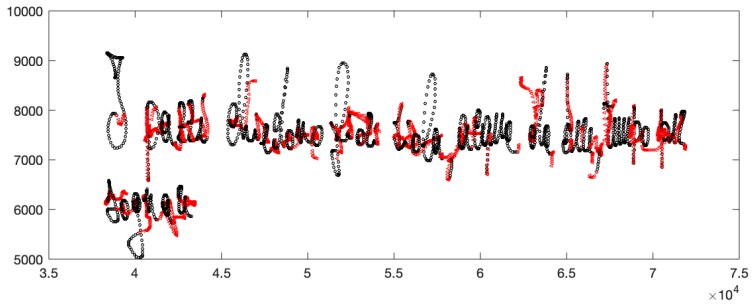

Fig. 2. Sample ink trace from the sentence copy task (task7). The red points correspond to in air movements close to the tablet area.

women largely overlap so that a classifier can hardly predict gender from one of these features alone.
However, one can generally expect prediction improvements by combining several features (see Fig. 5).

In the following, we will use a machine learning approach, and build an SVM classifier from the subset of extracted features selected by statistical analysis [1].

3 Experiments

A subset of features presented above, have been selected to build classifiers based on the SVM (Support Vector Machines) machine learning approach. SVMs are popular since they are suitable for datasets limited in size. The selected features are those which are found statistically different according to gender by an MANOVA approach [1] or a logistic regression approach.
To evaluate SVM classifiers for the gender prediction task, we use the popular accuracy metric. Accuracy is the proportion of correct predictions, computed on a test set. We propose several frameworks to compute accuracy, which differ in the way of choosing training and test data. These are:

- fixed train/test partitions. The training and test sets do not vary. A common partition is 70% training/30% test.
- Cross validation: train/test set are cycling according to K folds (K-1 folds for training, the Kth for test). The mean of the K accuracies is provided.
- Bootstrap: random train/test partitions (f.i. 70% training/ 30% test) are repeated N times (f.i. N=100). Accuracy of each repetition is collected, and the mean is provided.

Each instance of bootstrap, as well as each cycle of CV correspond to a fixed train/test partition. Thus results corresponding to the fixed train/test partition framework, may be grasped through max and min values of Tables 1 and 2 (4th column).

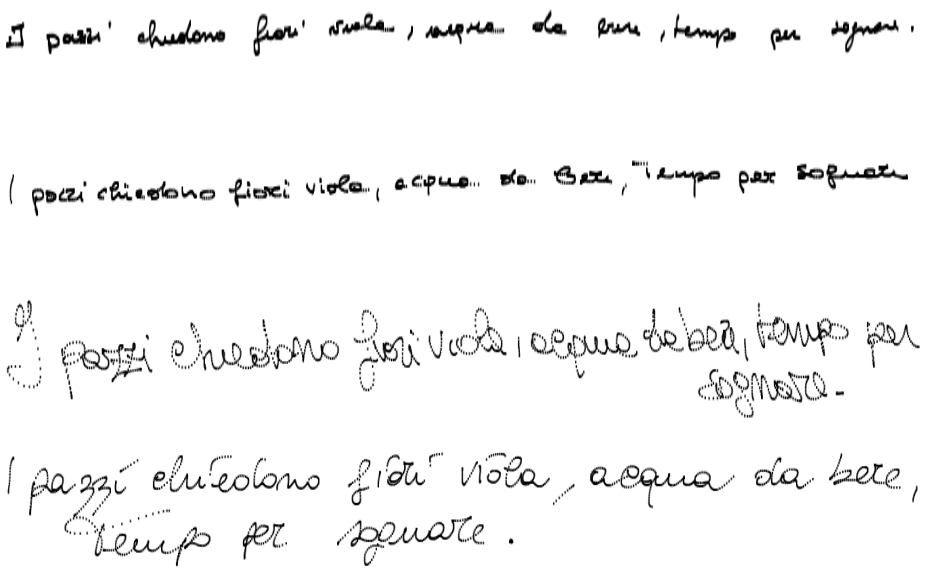

Fig. 3. Four task7 samples from subjects 16, 1, 20 and 141. Guessing gender is not straightforward.

The accuracies provided by these frameworks may largely differ, especially when dealing with small-size datasets (several hundred of subjects). Publicly available datasets often provide fixed train and test partitions. This is practical in order to compare approaches. However the actual accuracy may not correspond to this particular partition, especially for small datasets. Second, the test partition is often released, in contrast to keep it apart by the dataset designers who evaluate themselves the test accuracy. As a consequence, hyper-parameters may be tuned including test data, as well as feature selection. Accuracy may thus be over-estimated.

Figure 6 shows the interest of using cross validation, in contrast to fixed partitions. The mean CV accuracy is equal to 61.6% while the accuracy of one fold partition is much lower, equal to 37.5%, and that of another partition is much higher than the mean, and equal to 75%. Indeed the actual accuracy is around 60% which differs from a 75% accuracy provided by one particular partition.

The accuracies shown in CV (with $K = 10$) and Bootstrap (100 repetitions with 160 training samples and 72 test samples) results (see Tables 1 and 2) are rather low: 64.8% for bootstrap, 65.4% for cross-validation. However these accuracies may be still over-estimated since the selection of features was conducted from all samples (240 subjects), thus including the samples of the testing folds.

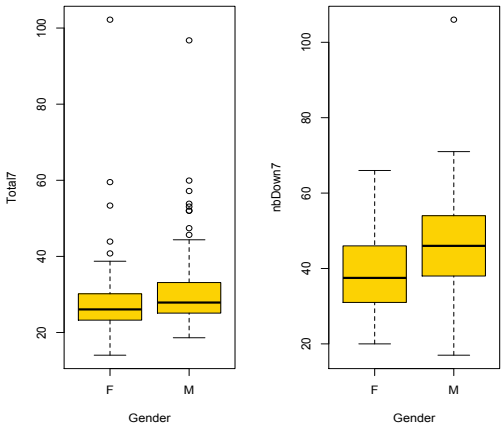

Fig. 4. Boxplots of Total7 and nbDown7 features, according to gender. Total7 feature is equal to the total time in seconds spent for completing the copy of the cursive sentence (task7). b) nbDown7 feature is equal to the number of strokes performed when writing the same sentence. Whiskers denote quartiles.

A fairer approach consists in clearly separating training data from test data. Thus an alternate feature selection approach for the CV framework consists in:

1. start with all samples divided in K folds and a set of extracted features
2. select the more efficient features from the K-1 training folds by a statistical approach
3. from the selected features, and the samples of the K-1 folds, build a gender prediction model
4. use the model to predict gender for the samples of the Kth fold.

This approach can directly be extended to the bootstrapping framework, by selecting features on the current training partition.

The set of features selected may vary from one training set to another. In the experiment conducted with the CV approach, the features selected by generalized logistic model (logistic regression) approach were the same for all training partitions: spaceA7, nbUp7, nbDown7, and spaceT7. But in general, important features could vary from one partition to another. We observe with CV (see Table 1) that selecting features from all samples (240) performs better than selecting features from the training set only: 65.4 % accuracy versus 61.2 %. Thus using testing data for feature selection may provide an over-estimation of the classifier accuracy.

Our results are consistent with those obtained in the literature for predicting gender from on-line handwriting [8] where accuracies range from 60% to 80%. Results show that if an accuracy of 75% or even 79% can be obtained for a

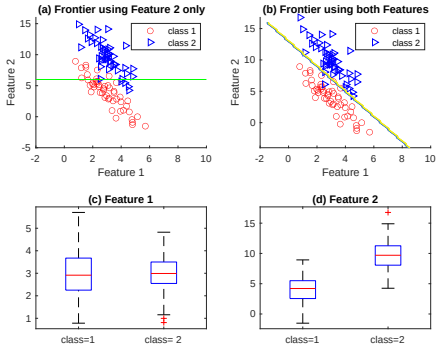

Fig. 5. (c) Feature 1 alone is inefficient for making class predictions in this two-class problem since whisker boxes largely overlap). (a) Feature 2 alone is more efficient but the resulting classifier is weak (many misclassifications). (b) An efficient classifier can be obtained by combining both features 1 and 2 (few misclassifications). Inspired from [17]

Approach	Features	Accuracy (in %)	Min/Max
CV & feat select. on whole set	Total7, tUp7 nbDown7, nbUp7	65.4 [11.3]	41.6/79.2
CV & feat select. on training folds	spaceA7, spaceT7 nbDown7, nbUp7	61.6 [12.5]	37.5/75

Table 1. Accuracies (in %) obtained by Cross Validation (K=10 folds). Standard deviations, min and max fold accuracies are provided. Feature selection is performed from whole set (240 samples) or from the training folds (216 samples) .

number of train/test partitions (see Max column in Tables 1 and 2), a reliable estimation of the accuracy is around 60-65 %, using CV and bootstrap. To our opinion, these low accuracies show that gender has a weak influence on on-line handwriting for the observed population (an accuracy of 50% would be obtained just with random guess).

4 Conclusion

Handwriting results from the combination of motor programs and social and environmental conditions. In this study, global features linked to hand movements and writing speed were extracted. First, features were selected according to a statistical analysis. These features were found distinct according to gender, but they could not be used in isolation because of large overlaps in the feature female/male distributions. Thus, selected features were combined through a machine learning based classifier (SVM). Low accuracies (around 65%) were obtained, estimated by cross validation and bootstraping, while higher ones were

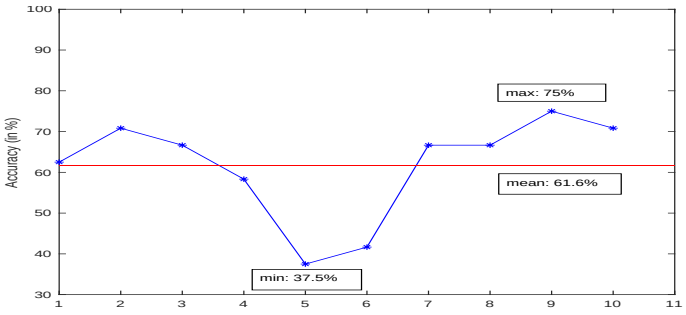

Fig. 6. Cross validation accuracies are varying according to the folds. The so-called CV accuracy is the mean accuracy equal to 61.6 %. The train/test partition corresponding to test the 9th fold and train on the remaining ones, reaches a maximum 75% accuracy.

Approach	Features	Accuracy (in %)	Min-Max
Bootstrap & feat select. on whole set	Total7, tUp7, nbDown7, nbUp7	64.8 [4.5]	52.7/76.4
Bootstrap & feat select. on 168 training samples	spaceA7, tUp7 nbDown7, nbUp7	66 [4.3]	55.6/76.4

Table 2. Accuracies (in %) obtained by Bootstrapping (100 repetitions) with 168 training samples/72 test samples resampled at each repetition). Standard deviations, min and max accuracies are provided. Feature selection is performed on all samples or on the training set built at each repetition.

obtained with fixed train/test partitions (79%). Such differences are observed due to the small dataset-size (several hundreds of subjects). To our opinion, the actual low accuracies (around 65%) corroborate the fact that the dynamics of handwriting may not be gender-based among our european young adult population.

The handwriting tasks performed by young adults are impersonal, and as mentioned before, there is no reason to find in the samples a difference due to gender. This could be different in another cultural context. In such case, experts should be able to illustrate and justify gender differences, if any.

References

1. G. Cordasco, M. Buonanno, M. Faundez-Zanuy, M. Riviello, L. Likforman-Sulem, and A. Esposito, "Gender Identification through Handwriting: an Online Approach," in *11th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2020*, pp. 197–202, 2020.
2. B. Megyesi, B. Esslinger, A. Fornés, N. Kopal, B. Láng, G. Lasry, K. de Leeuw, E. Pettersson, A. Wacker, and M. Waldispühl, "Decryption of historical manuscripts: the decrypt project," *Cryptologia*, vol. 44, no. 6, pp. 545–559, 2020.

3. T. team, "The handwriting book." https://www.yourtherapysource.com/blog1/2016/01/20/gross-motor-skills-and-handwriting-3/.
4. C. Sirat, J. Irigoin, and E. Poulle, "L'écriture: le cerveau, l'oeil et la main," in *Colloque International du CNRS, IRHT Paris*, pp. 1–6, 1990.
5. D. Simonnet, N. Girard, E. Anquetil, M. Renault, and S. Thomas, "Evaluation of children cursive handwritten words for e-education," *Pattern Recognition Letters*, vol. 121, pp. 133–139, 2019.
6. R. Plamondon, C. O'Reilly, and C. Ouellet-Plamondon, "Strokes against stroke - strokes for strides," *Pattern Recognit.*, vol. 47, no. 3, pp. 929–944, 2014.
7. C. Taleb, L. Likforman-Sulem, C. Mokbel, and M. Khachab, "Detection of parkinson's disease from handwriting using deep learning: a comparative study," *Evolutionary Intelligence*, 2020.
8. M. Liwicki, A. Schlapbach, and H. Bunke, "Automatic gender detection using online and off-line information," *Pattern Anal. Appl.*, vol. 14, no. 1, pp. 87–92, 2011.
9. N. Al-Qawasmeh and C. Suen, "Gender detection from handwritten documents using concept of transfer-learning," in *ICPRAI 2020 Intern. conf. on Pattern Recognition and Artificial Intelligence*, vol. 12068 of *Lecture Notes in Computer Science*, pp. 3–13, Springer, 2020.
10. F. Alaei and A. Alaei, "Gender detection based on spatial pyramid matching," in *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland* (J. Lladós, D. Lopresti, and S. Uchida, eds.), vol. 12824 of *Lecture Notes in Computer Science*, pp. 305–317, Springer, 2021.
11. A. Sotelo, H. Gómez-Adorno, O. Esquivel-Flores, and G. Bel-Enguix, *Gender Identification in Social Media Using Transfer Learning.* MCPR 2020 Lecture Notes in Computer Science, vol 12088, Springer, 2010.
12. M.-J. Berrichon-Seyden. personal communication, July 2021.
13. Y. Akbari, K. Nouri, J. Sadri, C. Djeddi, and I. Siddiqi, "Wavelet-based gender detection on off-line handwritten documents using probabilistic finite state automata," *Image Vis. Comput.*, vol. 59, pp. 17–30, 2017.
14. P. Maken and A. Gupta, "A method for automatic classification of gender, based on text-independent handwriting," *Multimedia Tools and Appl*, vol. 80, pp. 24573–24602, 2021.
15. E. Sokic, A. Salihbegovic, and M. Ahic-Djokic, "Analysis of off-line handwritten text samples of different gender using shape descriptors," in *2012 IX International Symposium on Telecommunications (BIHTEL)*, pp. 1–6, 2012.
16. S. Rosenblum, "Development, reliability, and validity of the handwriting proficiency screening questionnaire (hpsq)," *American Journal of Occupational Therapy*, 2008.
17. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. 7-8, pp. 1157–1182, 2003.