PAPER

A New Automated Method for Evaluating Mental Workload Using Handwriting Features

Zhiming WU[†], **Hongyan XU**[†,††], ***Student Members***, *and* **Tao LIN**[†a)], ***Nonmember***

SUMMARY Researchers have already attributed a certain amount of variability and "drift" in an individual's handwriting pattern to mental workload, but this phenomenon has not been explored adequately. Especially, there still lacks an automated method for accurately predicting mental workload using handwriting features. To solve the problem, we first conducted an experiment to collect handwriting data under different mental workload conditions. Then, a predictive model (called SVM-GA) on two-level handwriting features (i.e., sentence- and stroke-level) was created by combining support vector machines and genetic algorithms. The results show that (1) the SVM-GA model can differentiate three mental workload conditions with accuracy of 87.36% and 82.34% for the *child* and *adult* data sets, respectively and (2) children demonstrate different changes in handwriting features from adults when experiencing mental workload.

key words: *handwriting feature, mental workload, automated evaluation*

1. Introduction

Mental workload refers to the amount of mental demand imposed on a person by a particular cognitive task, and it can be attributable to the limited capacity of the person's working memory and his/her ability to process novel information [1]. The last decade has witnessed a growing interest in evaluating mental workload as a basic means of better understanding and improving human-computer interactions [2]–[4]. For example, Vertegaal and Chen used real-time feedback of mental workload to devise adaptive strategies of interruptions in a mobile phone [2]. Lin et al. [3] and Wilson and Sasses [4] have also emphasized that mental workload is a crucial user cost during interactions and should be carefully considered from both a health and safety point of view, since interaction techniques are becoming more and more pervasive. More importantly, it is highly desirable for intelligent training systems to evaluate learners' mental workload and keep it at an appropriate level in order to maximize their learning performance and motivation [5].

Current methods of evaluating mental workload can be mainly divided into three main categories: subjective, performance, and physiological measures. These technologies have been widely used with some success, but they have their own advantages and disadvantages (see Sect. 2 for details). In our opinions, an ideal evaluation methodology for mental workload should at least: (1) implicitly and continuously gather data; (2) require no extra equipment; (3) evaluate mental workload objectively, quantitatively and accurately; and (4) allow for automated evaluation and deployment in real-life scenarios. A little attention has recently been paid to behavioral measures in order to develop such an ideal method. Behavioral measures use changes in behavior patterns to indicate the levels of mental workload. The major advantage of behavioral measures is that behavioral data (e.g. handwriting process and speech) can often be collected implicitly without extra equipment, thereby increasing their applicability in real-world environments. Recently, the feasibility of employing changes in handwriting and verbal behavior to indicate mental workload has been identified by a few empirical findings [6], [7]. In the study, we are interested in handwriting behavior because it is pervasive in current educational environments.

Handwriting is a complex activity comprising cognitive, kinesthetic, perceptual and motor components, and it is often characterized by the information on pen-tip dynamics (velocity and acceleration), pen orientation (azimuth and altitude), pen-tip pressure, temporal (e.g. stroke duration on paper and in air) and spatial measures (e.g. stroke length and shape). Recently, a few studies (e.g. [8], [9]) have attempted to use machine learning techniques to model the relationship between mental workload and changes in handwriting features, but they still suffered from low classification accuracy. The low classification accuracy may mainly be attributed to one or more of the following issues: (1) the individual differences in handwriting patterns were not considered when handwriting features were extracted; (2) the features extracted to build the classifiers were limited to a narrow range; and (3) the algorithms developed were not effective enough for predicting mental workload.

The aim of this study was to develop a new automated model which can predict mental workload with relatively high accuracy. To achieve the goal, we first extracted extensive handwriting features from free text at two levels (i.e. sentence- and stroke-level features). Then, a hybrid model of Support Vector Machine and Generic Algorithm (hereafter SVM-GA) on the handwriting features was built and validated. Specifically, the study sought to answer the following research questions:

Research Question 1: can the SVM-GA model be used to predict mental workload of children and adults with

Manuscript received August 20, 2016.
Manuscript revised February 16, 2017.
Manuscript publicized May 30, 2017.

[†] The authors are with College of Computer Science, Sichuan University, No.24 South Section 1, Yihuan Road, Chengdu 610065, China.

[††] The author is with College of Tianfu, SouthWestern University of Finance and Economics, Jiuzhou Road, Mianyang 621000, China.

a) E-mail: lintao@scu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2016EDP7354

relatively high accuracy?

Research Question 2: whether and to what extent can classification accuracy be improved when individual differences in handwriting features are addressed?

Research Question 3: are there differences in handwriting patterns between children and adults when they are experiencing mental workload?

Based on the answers to these research questions, our contributions are to: (1) increase in-depth knowledge about what detailed features are predictive of mental workload for children and adults, and that complements what has been found in previous empirical studies and (2) develop a model on two-level handwriting features that can predict mental workload with relatively high accuracy.

2. Related Work

The most common techniques for mental load evaluation are subjective self-reports through questionnaires such as NASA-Task Load Index (NASA-TLX) [10]. Subjective techniques are good approaches to understanding the overall attitudes of participants, but they are usually administered post hoc or during break points, which may disrupt the normal flow of interaction. Moreover, subjective techniques tend to suffer from limited evaluative bandwidth and fail to evaluate mental workload at fine-grained level [11].

Performance measures evaluate mental workload from users' observable task performance index such as time to complete a task, type and number of errors, and success rates [12]. The most commonly used technique for performance measures is the dual-task approach (i.e. secondary and primary tasks) [11]. Task performance on the secondary task is supposed to reflect the amount of mental workload imposed by the primary task. The dual-task approach is highly sensitive and reliable, but it is rarely applied in real-world scenarios due to the following drawbacks [5]: (1) the secondary task may interfere considerably with the primary task, especially if the primary task is complex or if cognitive resources are limited; and (2) they are not suitable for real-time, automated deployment since they are usually calculated post hoc.

Physiological measures (e.g. heart rate variability, galvanic skin response, and pupil response) are objective and quantitative methods. One major advantage of physiological measures is the continuous availability of physiological response data, allowing mental workload to be measured at a high rate (high bandwidth) and with a high degree of sensitivity, even in situations in which overt behavior is relatively rare [1], [13]. However, physiological measures typically are intrusive and require supplemental equipment. Moreover, processing of physiological signals is computationally intensive and human expertise tends to be required to interpret the resulting patterns from physiological data [1], [13].

Some efforts have been made to identify the feasibility of using handwriting information to detect mental workload by examining statistical differences in some handwriting features (e.g. mean duration of pen-tip on surface and in the air) under different load conditions [7], [8], [14]. Although the empirical findings have demonstrated promise for using handwriting features to detect mental workload, there is only a limited understanding of how to use changes in handwriting behavior to indicate mental workload. For example, there is still no general agreement on what detailed handwriting features are predictive of mental workload [6]–[8], [14]. The studies by Yu et al. [8] and Lin et al. [9] reported that pressure-related features were good indicators of mental workload, while Luria and Rosenblum [7] did not find significant differences in pressure-related features under three mental workload conditions.

In addition, a few studies attempted to use classification models to automatically predict mental workload with some success, but they still suffered from low classification accuracy. For example, back-propagation neural network (BPNN) [9] and Gaussian mixture models [8] have been built on a limited range of handwriting features to predict mental workload, producing classification accuracy of 76.27% and 75.4%, respectively. The relatively low classification accuracy may mainly be attributed to one or more of the following reasons. First, the large individual differences in handwriting patterns were not considered when handwriting features were extracted. Second, handwriting features are generally divided into two levels: stroke- and task-level features. To our knowledge, almost all previous studies have only focused on a single level (stroke-level or task-level), while the potential of the combined use of stroke- and task-level features in indicating mental workload has not been investigated. Finally, the classification algorithms adopted were not effective enough for predicting mental workload. At present, low accuracy has limited the use of these machine learning models in real-world applications.

3. SVM-GA Model for Predicting Mental Workload

Support vector machines (SVMs) with the Radial Basis Function (RBF) kernel were used predict mental workload. For an SVM, two issues should be addressed: how to select the optimal input feature subset for the SVM and how to set the best kernel parameters. These two issues are crucial because the feature subset choice influences the appropriate kernel parameters and vice versa [15], [16]. Therefore, obtaining the optimal feature subset and SVM parameters must occur simultaneously. In previous studies (e.g., [15]–[17]), genetic algorithms (GAs) have been used to select feature subsets and determine SVM parameters simultaneously and were proved to be useful for improving the performance of SVMs. However, the efficacy of such GA-based SVM models on handwriting features in predicting mental workload has not been explored. In this study, we developed SVM-GA models on handwriting features to predict mental workload, inspired by a previous study [15]. That is, the two parameters of the RBF (the penalty parameter C and the gammar (γ)) and feature subset were encoded as a binary string, and were simultaneously optimized by a GA. The main steps of the SVM-GA model are described as follows (see Fig. 1).

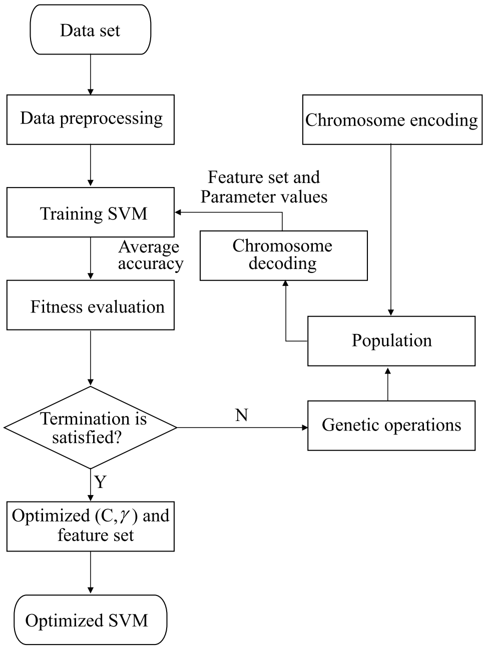

Fig. 1 The main steps of building the SVM-GA model.

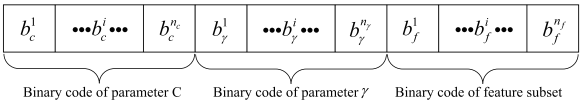

Fig. 2 A chromosome comprises three parts, parameter C, γ and feature subset.

(1) Data preprocessing. Each feature is linearly scaled to the range [0, 1] by the max-min scaling technique [18] and this can prevent the handwriting features in greater numeric ranges from dominating those in smaller numeric ranges.

(2) Encoding chromosomes and initializing GA. The binary encoding system was used to represent the chromosome. The chromosome comprises three parts, C, γ and the feature subset (see Fig. 2).

In Fig. 2, $b_C^1 \sim b_C^{n_C}$ represents the binary code of parameter C, $b_\gamma^1 \sim b_\gamma^{n_\gamma}$ represents the binary code of parameter γ, and $b_f{}^1 \sim b_f{}^{n_f}$ represents the feature mask. n_C is the number of bits representing parameter C, n_γ is the number of bits representing parameter γ, and n_f is the number of bits representing the features. n_C and n_γ can be chosen according to the required calculation precision. For the chromosome representing the feature set, the bit with value '1' represents that the feature is selected, and '0' indicates that the feature is not selected. The parameter settings for the GA are as follows: population size (500), crossover rate (0.7), mutation rate (0.02), n_c (20), n_r (20), and n_f (the number of the features).

(3) Decoding chromosomes. The chromosome representing the genotype of each parameter (C, γ) was decoded into the phenotype (parameter values) by Formula (1) [15]:

$$p = \min_p + \frac{\max_p - \min_p}{2^n - 1} \times d \tag{1}$$

where $\min_p$, $\max_p$, p, d and n refer, respectively, to the minimum and maximum values of a parameter and the phenotype, decimal value and length of a bit string.

(4) Evaluating fitness. The fitness of an individual in the population was calculated according to the average classification accuracy of 10-fold cross validation, which is the overall number of correct classifications from 10 iterations divided by the total number of samples (n) in a data set. The fitness value (F) is calculated by Formula (2):

$$F = \sum_{i=1}^{n} \mathrm{H}_i/\mathrm{n} \tag{2}$$

where H_i is 1 if the predicted value of the SVMs equals the actual class label, otherwise H_i is zero.

(5) Terminating GA. If the termination condition is satisfied, the GA stops; otherwise, the GA continues with the next generation. The termination criteria of the GA are that the generation number reaches 600 or that the fitness value does not improve during the last 100 generations.

(6) Performing genetic operations. Two-point crossover, roulette wheel selection, and elitism replacement techniques were used as the genetic operators.

4. Experiment

An experiment was conducted to collect the samples of handwriting behavior that had been accurately labeled as the behavior captured under different mental workload conditions.

4.1 Participants

Two hundred participants (100 females and 100 males) were recruited for the experiment. One hundred of the participants are university students aged 18 to 31 (mean age = 22.83, SD = 2.24), and the other one hundred participants are students aged 9 to 12 (mean age = 11.39, SD = 1.12) in elementary schools. Before the experiment, all participants were required to complete a background questionnaire about their experience with handwriting devices and English learning, along with personal information such as sex and handedness. English is not the first language for any participants, but all of them have learnt English over three years. All participants are right handed with normal cognitive and physical ability. Participants were compensated for their participation in the experiment.

4.2 Tasks

Each participant experienced four experimental conditions comprised of one baseline condition (transcribing a sentence) and three sentence-making task conditions. Inspired by Ransdell and Levy's experiments [19], we chose sentence-making training as the experimental task. Specifically, after seeing a set of randomly selected words on the screen, participants were required to write down the composed sentences based on these given words in a single line or multiple lines. All words were from the participants' textbook. They were advised to use the given words, but not necessary in the given order.

The task difficulty was manipulated by the number of the given words and the tasks become more difficult with the increase in the number of the given words. Participants completed the sentence-making tasks under three difficulty conditions (based on one, two and three given words) and these conditions were expected to cause three different levels of mental workload, respectively. Every time the participant pressed the "start" key, the given words were displayed on screen for a limited time before disappearing. The duration for displaying the words was one second for one-word cases, two seconds for two words and three seconds for three words. The participants were required to remember the displayed words and write a sentence with the given words. There was no time limit for writing, but participants were not allowed to write down the words before writing the sentence.

4.3 Procedures

The experiment was divided into four phases: a welcome phase, a practice phase, a task phase, and a debriefing phase. During the welcome phase, all participants were required to sign a consent form with a detailed description of the experiment, its duration, and its research purpose. Each participant also filled out a background questionnaire.

At the outset of the practice phase, instructions were read to each participant describing the task rules, and each participant was given a brief tutorial on how to complete the task. Participants were then allowed to practice for several minutes to ensure that they mastered the skill of using the experimental device. In addition, the experiment coordinator confirmed with participants that they understood the meaning of each word which might appear in the tasks.

During the task phase, each participant was first required to normally transcribe one sentence as her/his baseline condition. They then completed three sentence-making tasks with different difficulty. The three task conditions were randomly presented to participants to avoid the order effects. After each task condition, participants had about 7 min to rest and complete a questionnaire rating mental workload to that task condition using an online NASA-TLX questionnaire system. After a participant finished one sentence and proceeded to the next one, the writing space was cleared automatically. At the conclusion of the experiment, participants were debriefed on their impressions of the experiment.

4.4 Data Collection

Handwriting process was recorded using a WACOM DTZ-1200W tablet at 142 Hz. The device can provide almost the

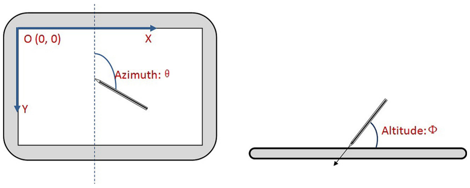

Fig. 3 Handwriting data captured by a WACOM tablet.

Table 1 Sentence-level features of a sentence.

Category	Description	Abbr.
Dynamics features (DF)	Average velocity of writing a sentence (i.e. average writing velocity)	AV
	Standard deviation of writing velocity	SDV
	Average writing velocity in X direction	AVX
	Standard deviation of writing velocity in X direction	SDVX
	Average writing velocity in Y direction	AVY
	Standard deviation of writing velocity in Y direction	SDVY
	Maximum writing velocity	MV
	Maximum writing velocity in X direction	MVX
	Maximum writing velocity in Y direction	MYY
	Average writing acceleration	AA
	Average writing acceleration in X direction	AAX
	Average writing acceleration in Y direction	AAY
	Maximum writing acceleration	MA
	Maximum writing acceleration in X direction	MAX
	Maximum writing acceleration in Y direction	MAY
	Average pen altitude	AAL
	Standard deviation of pen altitude	SDAL
	Maximum pen altitude	MAL
	Average pen azimuth	AAZ
	Standard deviation of pen azimuth	SDAZ
	Maximum pen azimuth	MAZ
Temporal Features (TS)	Total duration of writing a sentence	TD
	Duration of pen-tip in the air	DIA
	Duration of pen-tip on writing surface	DOS
	Ratio of DIA to DOS	RATS
	Count of the pen-tip pauses longer than 300 (ms) (sensible pauses) between two successive moves	SPC
Pressure Features (PS)	Average pen-tip pressure	AP
	Maximum pen-tip pressure	MP
	Standard deviation of pen-tip pressure	SDP
Spatial Features (SF)	Total distance traveled by pen-tip on writing surface	TDS
	Total distance traveled by pen-tip in X direction	TDSX
	Total distance traveled by pen-tip in Y direction	TDSY

Table 2 Stroke-level features of a sentence.

Category	Descriptions	Abbr.
Temporal Features (TF)	Average stroke duration	S_AD
	Standard deviation of stroke duration	S_SDD
	Total stroke duration	S_TD
Spatial Features (SF)	Average stroke length	S_AL
	Standard deviation of stroke length	S_SDL
	Total stroke length	S_TL
	Average stroke height (height refers to the direct distance in the y-axis from the lowest to the highest point of a stroke.)	S_AH
	Standard deviation of stroke height	S_SDH
	Maximum stroke height	S_MH
	Average stroke width (width refers to the direct distance in the x-axis from the left side of a stroke to the right side.)	S_AW
	Standard deviation of stroke width	S_SDW
	Maximum stroke width	S_MW
Dynamics Features (DF)	Average stroke angular velocity (angular velocity refers to the degrees through which pen travels per second during writing a stroke.)	S_AAV
	Standard deviation of stroke angular velocity	S_SDAV
	Maximum stroke angular velocity	S_MAV
Pressure Features (PF)	Average stroke pressure	S_AP
	Standard deviation of stroke pressure	S_SDP
	Maximum stroke pressure	S_MP

same writing experience as would be found in the pencil-and-paper writing. Furthermore, the use of the device is so simple that children can be familiar with it during a few minutes in the practice phase. Actually, at the debriefing phase, all participants also reported that the use of the device imposed no or little load on them.

Handwriting raw data are a series of samples along the pen trace with time-stamps and include the coordinate positions (x, y), pressure (0-1024 levels) of the pen-tip (p), and the altitude and azimuth of the pen (θ, ϕ), as shown in Fig. 3. Handwriting features for each sentence were calculated from these data sources.

4.5 Handwriting Features

Handwriting features were obtained at both the stroke-level and sentence-level (i.e. task-level) for each sentence. Handwriting features at each level can be roughly divided into four categories: temporal, spatial, dynamics and pressure, as listed in Tables 1 and 2. Specifically, sentence-level features were obtained by calculating statistics such as average and standard deviation of handwriting measures across the sentence (see Table 1).

A stroke is defined as the curve created by the movement of pen-tip on the writing surface, in which pen-tip pressure of all sampled data points is greater than 50 (non-scale units). We ruled out outlier (very long or very short) strokes which seemed extreme, as compared with other strokes. That is, strokes of less than 50ms or more than 850ms duration were deleted. A sentence includes numerous strokes. Stroke-level features of a sentence were extracted by calculating some statistics of handwriting measures across all strokes of the sentence. Table 2 lists the extracted stroke-level features for each sentence.

Given the individual differences of handwriting behavior, the raw data of handwriting were normalized per participant per feature using Z-scores [18]. For each participant, the baseline samples were used to calculate means and standard deviations per feature; then the samples under three manipulation conditions were normalized using those values. The features extracted from the raw and normalized data are called the raw and normalized features, respectively. Both the raw and normalized features were investigated to determine if accounting for individual differences by normalizing data in this way would yield higher classification accuracy.

5. Results

Handwriting data and subjective ratings were collected from 200 participants, but the data from four students of elementary schools and two university students had to be discarded because they failed to complete all three tasks. The collected data were divided into two data sets: *child* (96 elementary school students) and *adult* (98 university students). All analyses were conducted on each data set separately. In the study, mental workload participants perceived was first evaluated through the overall NASA-TLX score to confirm whether the designed task conditions induced different mental workload levels. Then, the SVM-GA models were, respectively, built on the raw and normalized handwriting features to classify different mental workload levels. The development platform for our models was Intel Core CPU i7-3770 (3.4 GHz, 4 cores), 8G RAM, Windows7 operating system. The development environment was MATLAB (2012a) and the software of SVM was Libsvm (3.1).

5.1 Subjective Evaluation for Mental Workload

For the *child* data set, one-way ANOVA analyses showed that there were significant differences in overall NASA-TLX scores across the three task conditions ($F_{(2,285)} = 163.21$, $p < 0.001$). Post-hoc comparisons also showed that the three-word tasks induced the greatest mental workload, followed by the two-word task and one-word task (see Table 3). For the *adult* data set, the same pattern (see Table 3) was also found ($F_{(2,291)} = 157.64$, $p < 0.001$). In addition, when examined individually, the patterns were almost consistent for all participants. At the debriefing phase, almost all participants also reported that they felt it easy to compose sentences with a single word, and challenging or especially challenging for the three-word tasks.

The above results validate our experimental task design

Table 3 Significant differences in the self-reported mental workload across the three tasks were confirmed. (* represents $p < 0.01$)

Data	One-word	Two-word	Three-word	F	One-Two	One-Three	Two-Three
sets	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*		t	t	t
Child	59.61(6.94)	68.84(9.82)	78.28(11.33)	163.21*	7.53*	17.12*	11.34*
Adult	49.36(5.55)	60.99(6.34)	72.08(9.73)	157.64*	6.61*	16.75*	9.13*

Table 4 The ten normalized features in descending order of importance ranked by information gain.

Child	1	2	3	4	5
	DIA	SPC	S_AD	AV	S_AP
	6	7	8	9	10
	AAL	S_SDP	SDAZ	S_AAV	AP
Adult	1	2	3	4	5
	S_SDD	SPC	DOS	RATS	AVY
	6	7	8	9	10
	S_AAV	S_SDP	S_MP	S_AL	S_SDW

and confirm that the three task conditions indeed can elicit different mental workload levels, which provides an important benchmark for building the SVM-GA model to predict mental workload.

5.2 Information Gain Analyses

To understand the relative importance of each feature for building a classifier, an analysis of information gain was conducted. Take the example of the normalized features. Table 4 lists the top ten features in descending order of importance and there are large differences in them between the *child* and *adult* data sets.

The results in Table 4 provide positive support for **Research Question 3** and show something rather interesting: (1) temporal features clearly offered the best promise for detecting mental workload for each data set among four categories of features (i.e. dynamics, temporal, and spatial and pressure features), but the *child* and *adult* data sets have different detailed features (*child*: DIA, SPC and S_AD; *adult*: S_SDD, SPC, DOS and RATS); (2) two pen-orientation features (AAL and SDAZ) were respectively ranked sixth and eighth among the top ten features for the *child* data set, but none of pen-orientation features was found among the top ten important features for the *adult* data set; (3) dynamics and pressure features also provide useful information for detecting mental workload, but there were differences in the detailed features and their order of importance between the child and adult data sets (*child*: S_AP, S_SDP, S_AAV and AP; *adult*: AVY, S_AAV and S_MP); and (4) two spatial features (S_AL and S_SDW) were respectively ranked ninth and tenth for the *adult* data set, but none of spatial features was found among the top ten important features for the *child* data set. The findings for the raw features are similar to those for the normalized features.

5.3 Accuracy of SVM-GA Models

The average classification accuracy of the SVM-GA models on the child and adult data sets were calculated to answer **Research Question 1**. The results show: (1) for the *child* data set, the SVM-GA models could, respectively, classify three mental workload levels with accuracy of 87.36% and 79.52% on the normalized and raw features; and (2) for the *adult* data set, the SVM-GA models could, respectively, obtain the accuracy of 82.34% and 78.69% on the normalized and raw features.

Average classification accuracy of the two data sets was compared (t-tests). The results showed that, whether for the normalized or for raw features, the average classification accuracy of the *child* data set was significantly higher than those for the *adult* data set (normalized: $t_{18} = 3.03$, $p = 0.002$; raw: $t_{18} = 2.21$, $p = 0.047$).

The results also showed that data normalization (Z-score) which accounts for individual differences is generally helpful for improving the performance of the SVM-GA models, but that this benefit is more pronounced for the *child* data set. Specifically, using normalized features resulted in an increase (just 2.65%) in classification accuracy for the *adult* data set. In contrast, normalized features resulted in a great increase (7.84%) in classification accuracy for the *child* data set. We conducted paired t-tests on the classification accuracies between the raw and normalized features for the *child* and *adult* data sets separately. The results showed that the improvement of accuracy for both the *child* and *adult* data sets was significant (*child*: $t_9 = 3.47$, $p < 0.001$; *adult*: $t_9 = 2.52$, $p = 0.03$). These results answer **Research Question 2** and show that individual differences should be taken into consideration when developing machine learning techniques to predict mental workload, especially if mental workload of children is of interest.

Considering that the normalized features can produce better results than the raw features for both the *child* and *adult* data sets, we take the example of the normalized features to describe the feature subset selected by the GA, as listed in Table 5. The results in Table 5 show: (1) for both the *child* and *adult* data sets, the feature subsets included four categories of features (i.e. dynamics, temporal, spatial and pressure features) and this identified the inference by Luria and Rosenblum [7] that different handwriting measures can provide distinct information on mental workload; and (2) for the *child* data set, two pen-orientation features (i.e. AAL and SDZA) were selected as the input of its SVM-GA model, but the feature subset for the *adult* data set did

Table 5 The normalized features selected by the GA for each data set. *DF*, *TF*, *PF*, and *SF* represent four categories of handwriting features (i.e. dynamics, temporal, pressure and spatial).

	Sentence-level features			
	DF	*TF*	*PF*	*SF*
Child	AV,AAL,AAX,SDAZ	DIA,SPC	AP,SDP	
Adult	AVY	DRAT	SDP	
	Stroke-level features			
	DF	*TF*	*PF*	*SF*
Child	S_AAV	S_SDD	S_AP	
Adult	S_AAV	S_AD	S_AP	S_AL,S_SDW

Table 6 Performance comparison between SVM-GA and SVM-Grid models on the normalized features.

	Accuracy (%)		**Average Time (Seconds)**	
	Child	*Adult*	*Child*	*Adult*
SVM-GA	87.36	82.43	682.12	572.36
SVM-Grid	76.21	71.72	479.23	337.91
p-values	0.011	0.013	0.016	0.003

not include any pen-orientation features. The results also provide positive answers to **Research Question 3**.

5.4 Performance Comparison between SVM-GA and SVM-Grid

The Grid algorithm is also a common method for searching for the best C and γ when using SVMs with the RBF kernel function. We compared the classification accuracy and the average time of searching for the best pair (C and γ) of the GA-based approach with those of the Grid algorithm. Considering that doing a complete grid-search may be time-consuming, an improved Grid algorithm [20] with a grid space (C:[2^{-5}, 2^{15}]; γ:[2^{-15}, 2^{3}]) was developed.

Similar to the SVM-GA models, the SVM-Grid models on the normalized features had higher accuracy than on the raw features for both the *child* and *adult* data sets. Therefore, t-test was used to compare the performance (i.e., accuracy and average search time) between the SVM-GA and SVM-Grid models on the normalized features (see Table 6). It can be seen from Table 6 that (1) the SVM-GA obtained higher average accuracy than the SVM-Grid, producing the increase of 11.15% and 10.71% on the *child* and *adult* data sets, respectively; and (2) for the *child* and *adult* data sets, the average time of searching for the pair (C and γ) of the Grid algorithm was 202.89 and 234.45 seconds shorter than that of GA-based approach, respectively. The results showed that the average searching time of the GA-based approach is slightly inferior to that of the Grid algorithm (all p-values < 0.05), but it significantly improved the classification accuracy (all p-values < 0.05). Note that the software environment for the two approaches and the predefined searching precision of the Grid algorithm may affect the computational time.

6. Discussions

This study designed a hybrid model of GA and SVM (SVM-GA) on handwriting features to predict mental workload and it served the purpose well. Specifically, our results showed that its classification accuracy on the normalized features was significantly higher than those on the raw feature for each data set, producing the accuracy of 87.36% and 82.74% on *child* and *adult* data sets, respectively. The accuracy was also vastly superior to that obtained by the SVM-Grid models. There is a considerable improvement in classification accuracy compared with the result (76.27%) in a recent study [9] that adopted the same experimental design and task. These results show that the SVM-GA models are highly effective for classifying mental workload levels when individual differences are addressed, especially when mental workload evaluation of children is of interest. These results answer **Research Question 1**. While the classification accuracy for mental workload may not be comparable to the accuracy obtained by physiological data, the current approach has advantages over physiological measures because it requires no additional hardware, is unobtrusive, and is less computationally intensive.

The SVM-GA models show good performance in the study, which is partly attributed to their good capability to model complex behavior patterns. For example, the potential of pressure-related features in detecting mental workload was not confirmed by MANOVA analyses in the study by Luria and Rosenblum [7], but their values for evaluating mental workload were clearly identified by the SVM-GA model. Pressure-related features were selected to construct the SVM-GA models for both the *child* (AP, SDP and S_AP) and *adult* (AP, and S_AP) data sets. The findings were also supported by the results from information gain analyses, which showed that several pressure-related features (e.g., AP, S_MP and S_SDP) ranked as the top ten important features for evaluating mental workload for both the *child* and *adult* data sets. These findings also further confirm the judgment made by Luria and Rosenblum [7] that not all handwriting features relate linearly to mental load and nonlinear models should be further considered when developing techniques to predict mental workload. In fact, we also built other common classifiers (along with the corresponding feature selection wrappers) to predict mental workload, including Decision Trees (C4.5), BPNNs, k-Nearest Neighbor, AdaBoost, and Linear Discriminant Analysis. However, the SVM-GA models consistently obtained the best results for both the *child* and *adult* data sets.

In addition to the good modeling capability of the SVM-GA model, the obtained high classification accuracy may be attributed to the feature extraction technique at two levels (i.e. sentence-level and stroke-level features). The sentence-level features can provide the global handwriting information during tasks and the stroke-level features can describe handwriting pattern at a more detailed level. We believe that the features at the two levels can complement each other well when used to predict mental workload. To our knowledge, this is the first attempt to combine sentence- and stroke-level features to build the automated model of evaluating mental workload, and has shown the potential of them in predicting mental workload.

The SVM-GA models were built on both the raw and normalized features to determine whether accounting for individual differences would improve their classification performance. The baseline condition and Z-score method were used to normalize the raw data to address individual differences in handwriting features, and the classification accuracy on the normalized features was significantly higher than those on the raw features for each data set, providing a positive support for **Research Question 2**. These results open a new door for how to improve classification accuracy when classifiers are built on handwriting feature to predict mental workload.

It is also important to note that the nature of the features selected for the SVM-GA models of children and adults may be not same, answering **Research Question 3**. Take the example of the normalized handwriting features. For the *child* data set, two pen orientation-related features (AAL and SDZA) were selected as the input of its SVM-GA model, while the feature subset for the *adult* data set did not include any pen orientation-related features; this shows that handwriting information on pen orientation may make distinct contributions to predict children's mental workload. In contrast, two spatial features (S_AL and S_SDW) were selected to construct the SVM-GA model of adults, while the SVM-GA model for children did not include any spatial features. In addition, information gain analyses also showed that while pressure and dynamics features were useful for evaluating mental workload, they had different importance for the *child* and *adult* data sets. Therefore, we believe that classifiers need to be trained for adults and children separately to obtain better performance due to the differences in handwriting patterns between them.

The SVM-GA model also obtained significantly higher accuracy for the *child* data set compared with the *adult* data set, producing the increases of 1.01% and 5.02% on the raw and normalized features, respectively. The higher accuracy on the *child* data set may be attributable to the following two reasons. First, compared to the adults, the children experienced higher mental load for the same experimental task due to their relatively low level of proficiency in English or encountering more difficulty during the experimental procedure; the higher load might cause more obvious changes in handwriting behavior, which made it possible to obtain a higher accuracy. Second, handwriting performance becomes automatic with time, and children tend to have a lower automatic level than adults. The less automatic handwriting is, the more variability there is in handwriting behavior [7], [21], [22]. For children, the dis-automatization as a result of mental workload may be more obvious than adults, enabling mental workload of children to be easily detected. Due to the limitation of our experimental design, our results could not conclude that changes in handwriting behavior caused by mental workload for children are more detectable than those for adults, but this phenomenon have proposed an interesting research problem that should be carefully examined in future work.

7. Conclusions and Future Work

To our best knowledge, machine learning techniques have rarely been created on handwriting features for mental workload evaluation. The exploratory study extends the state of the art by (1) illustrating a new methodology of combining SVM and GA algorithms to automatically evaluate mental workload levels with relatively high classification accuracy; (2) showing the differences in handwriting patterns between children and adults when they are experiencing mental workload; (3) highlighting the importance of addressing individual differences in handwriting features when classifiers are adopted to predict mental workload; and (4) providing knowledge about what detailed handwriting features are predictive of mental workload for children and adults, respectively.

The opportunities for future work are vast. The high classification accuracy obtained by the SVM-GA model can spur this method into practical applications in future work. For children, most handwriting activities require the focused attention, and the automated evaluation for mental workload can help us to determine the mental efforts they are experiencing, and further give us hints on improving their writing performance or developing adaptive writing interfaces.

Our further research will also work towards improving the generalizability of our findings. Mental workload in the study was induced by sentence-making tasks in laboratory settings and participants were drawn from the population at elementary schools and a university. Although our experimental results show the potential of using SVM-GA and two-level handwriting features to evaluate mental workload, its effectiveness should be further validated across multiple kinds of experimental tasks (e.g., writing numbers) and fine-grained mental workload levels in real-world situations. In addition, the size of the data set in the study is relatively small for machine learning and a large number of samples could allow for more accurate classification results.

Acknowledgments

This study was supported by the Science and Technology Supporting Program, Sichuan Province, China (No. 2013GZX0138 and No.2014GZ0154), the Research Foundation of Sichuan University for Young Teachers (No. 2015SCU11050).

References

[1] A.F. Kramer, "Physiological metrics of mental workload: A review of recent progress," Multiple-task Performance, ed. D.L. Damos, pp.279–328, Taylor and Francis, London, 1991.

[2] D. Chen and R. Vertegaal, "Using mental load for managing

interruptions in physiologically attentive user interfaces," Proc. 22th Conf. on Hum. Factors Comput. Syst., Vienna, Austria, pp.1513–1516, April, 2004.

[3] T. Lin, A. Imamiya, and X. Mao, "Using multiple data sources to get closer insights into user cost and task performance," Interact. Comput., vol.20, no.3, pp.364–374, May 2008.

[4] G.M. Wilson and M. Angela Sasse, "From doing to being: Getting closer to the user experience," Interact. Comput., vol.16, no.4, pp.697–705, Aug. 2004.

[5] F. Paas, J.E. Tuovinen, H. Tabbers, and P.W.M. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," Educ. Psychol., vol.38, no.1, pp.63–71, March 2003.

[6] F. Chen, N. Ruiz, E. Choi, J. Epps, M.A. Khawaja, R. Taib, B. Yin, and Y. Wang, "Multimodal behavior and interaction as indicators of cognitive load," ACM Trans. Interact. Intell. Syst., vol.2, no.4, pp.1–36, Dec. 2012.

[7] G. Luria and S. Rosenblum, "A computerized multidimensional measurement of mental workload via handwriting analysis," Behav. Res. Methods, vol.44, no.2, pp.575–586, June 2012.

[8] K. Yu, J. Epps, and F. Chen, "Cognitive load evaluation with pen orientation and pressure," Proc. 24th ACM Symp. on User Interface Softw. Tech., pp.1–4, Alicante, Spain, Oct. 2011.

[9] T. Lin, T. Xie, Y. Chen, and N. Tang, "Automatic cognitive load evaluation using writing features: An exploratory study," Int. J. Ind. Ergon., vol.43, no.3, pp.210–217, May 2013.

[10] S.G. Hart and L.E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in Human Mental Workload, eds. P.A. Hancock and N. Meshkati, pp.139–183, Elsevier Science, Amsterdam, 1988.

[11] W.W. Wierwille and F.T. Eggemeier, "Recommendations for mental workload measurement in a test and evaluation environment," Hum. Factors: J. Hum. Factors Ergon. Soc., vol.35, no.2, pp.263–281, June 1993.

[12] E. Galy, M. Cariou, and C. Melan, "What is the relationship between mental workload factors and cognitive load types?," Int. J. of Psychophysiol., vol.83, no.3, pp.269–275, March 2012.

[13] R.L. Mandryk and M.S. Atkins, "A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies," Int. J. Hum-Comput. Stud., vol.65, no.4, pp.329–347, April 2007.

[14] K. Yu, J. Epps, and F. Chen, "Cognitive load evaluation of handwriting using stroke-level features," Proc. 16th Int. Conf. on Intell. User interfaces, Palo Alto, CA, USA, pp.423–426, Feb. 2011.

[15] C.-L. Huang and C.-J. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," Expert Syst. Appl., vol.31, no.2, pp.231–240, Aug. 2006.

[16] H. Frohlich, O. Chapelle, and B. Scholkopf, "Feature selection for support vector machines by means of genetic algorithm," Proc. 15th Int. Conf. on Tools Artif. Intell., Sacramento, California, USA, pp.142–148, Nov. 2003.

[17] M. Zhao, C. Fu, L. Ji, K. Tang, and M. Zhou, "Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes," Expert Syst Appl, vol.38, no.5, pp.5197–5204, May 2011.

[18] J. Han, M. Kamber, and J. Pei, Data mining: Concepts and techniques, Morgan Kaufmann Publishers, San Francisco, 2011.

[19] S. Ransdell and C.M. Levy, "Writing, reading, and speaking memory spans and the importance of resource flexibility," The Cognitive Demands of Writing: Processing Capacity and Working Memory in Text Production, ed. M. Torrance and G. Jeffrey, pp.99–113, Amsterdam University Press, 1999.

[20] C.W. Hsu, C.C. Chang, and C.J. Lin, "A practical guide to support vector classification," Tech. Rep., Department of Computer Science, National Taiwan University, vol.67, no.5, 2003.

[21] B.C.M. Smits-Engelsman and G.P. Van Galen, "Dysgraphia in children: Lasting psychomotor deficiency or transient developmental delay?," J. Exp. Child Psychol., vol.67, no.2, pp.164–184, Nov. 1997.

[22] J. Wann, "Handwriting disturbances: Developmental trends," in Themes in motor development, eds. H.T.A. Whiting and M. Wade, pp.207–223, Springer, Berlin, 1986.

Zhiming Wu is a postgraduate in Sichuan University, graduated from East China Normal University with a degree of bachelor in Software Engineering, and is studying Human-Computer Interaction to complete a doctor degree of Computer Science in Sichuan University.

Hongyan Xu received the M.S. degree in computer science from Shanghai Maritime University in 2008, and is studying Human-Computer Interaction to complete a doctor degree of Computer Science in Sichuan University. He is a teacher of College of Tianfu, Southwestern University of Finance and Economics.

Tao Lin received the M.S. degree in computer science from Sichuan University, China, in 2003, and the PhD degree in information science from University of Yamanashi, Japan, in 2007. He also worked at Waseda University, Japan, as a visiting lecturer. Currently, He is a professor at Sichuan University, China.