Sentiment Pen: Recognizing Emotional Context Based on Handwriting Features

Jiawen Han
Keio Media Design
Yokohama, Japan
hanjiawen@kmd.keio.ac.jp

George Chernyshov
Keio Media Design
Yokohama, Japan
chernyshov@kmd.keio.ac.jp

Dingding Zheng
Keio Media Design
Yokohama, Japan
zheng208@kmd.keio.ac.jp

Peizhong Gao
University of Tokyo
Tokyo, Japan
kou@cyber.t.u-tokyo.ac.jp

Takuji Narumi
University of Tokyo
Tokyo, Japan
narumi@cyber.t.u-tokyo.ac.jp

Katrin Wolf
HAW Hamburg
Hamburg, Germany
katrin.wolf@haw-hamburg.de

Kai Kunze
Keio Media Design
Yokohama, Japan
kai@kmd.keio.ac.jp

ABSTRACT

In this paper, we discuss the assessment of the emotional state of the user from digitized handwriting for implicit human-computer interaction. The proposed concept exemplifies how a digital system could recognize the emotional context of the interaction. We discuss our approach to emotion recognition and the underlying neuro-physiological mechanisms. To verify the viability of our approach, we have conducted a series of tests where participants were asked to perform simple writing tasks after being exposed to a series of emotionally-stimulating video clips from EMDB[6], one set of four clips per each quadrant on the circumplex model of emotion[28]. The user-independent Support Vector Classifier (SVC) built using the recorded data shows up to 66% accuracy for certain types of writing tasks for 1 in 4 classification (1. High Valence, High Arousal; 2. High Valence, Low Arousal; 3. Low Valence, High Arousal; 4. Low Valence, Low Arousal). In the same conditions, a user-dependent classifier reaches an average of 70% accuracy across all 12 study participants. While future work is required to improve the classification rate, this work should be seen as proof-of-concept for emotion assessment of users while handwriting aiming to motivate research on implicit interaction while writing to enable emotion-sensitivity in mobile and ubiquitous computing.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; **Laboratory experiments**; *Text input*; *Systems and tools for interaction design.*

KEYWORDS

Affective Computing, Emotional Recognition, Handwriting Analysis

ACM Reference Format:
Jiawen Han, George Chernyshov, Dingding Zheng, Peizhong Gao, Takuji Narumi, Katrin Wolf, and Kai Kunze. 2019. Sentiment Pen: Recognizing Emotional Context Based on Handwriting Features. In *Augmented Human International Conference 2019 (AH2019), March 11–12, 2019, Reims, France.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3311823.3311868

AH2019, March 11–12, 2019, Reims, France

ACM ISBN 978-1-4503-6547-5/19/03…$15.00
https://doi.org/10.1145/3311823.3311868

1 INTRODUCTION

Writing, drawing, and painting have always been known to be a way to record, express and share one's feelings, emotions, thoughts and ideas. Anywhere from Bryullov's "The Last Day of Pompeii" to a heart drawn in the corner of a Happy Birthday note to a friend, emotions are the core of the message. In face-to-face communication, emotion can be expressed through mimic, voice, and gestures to indicate how pure words are meant. For example, the expression "you're really smart" can change from a compliment to irony using different pronunciations, voice tone, and mimic. In written communication, such emotional hints often get lost and the emotions related to messages, and consequently the entire message, is every so often misunderstood.

A large body of psychology and computer science research is investigating handwriting for person's identification as well as for emotion, cognitive state and personality feature detection [17, 18, 20, 31]. We believe that emotion recognition and communication could bring the user experience and satisfaction to a whole new level through adding information about the writers' emotions to the written messages. Such information would not only increase the information depth of written messages, it could also create empathy about the emotional state of a person which could be used to enrich a wider range of applications. Picture a smartphone that does not want to bother its owner with calls from an annoying coworker because the owner is tired and would rather enjoy the weekend, but lets through the call from a close friend. This may seem to be eery at this moment in time, but a smartphone can never be a true PDA (personal digital assistant) unless it can recognize the context of the situation as well as a human-assistant. As long as the inability to understand the context of the interaction persists, technology can never augment and change our lives as well as another human can.

The machines' poor level of understanding the emotional context has obvious reasons stemming from the history of the computing. Originally computers were a tool with a sole purpose of processing data fast. In this context expecting any emotional intelligence from a computer is as absurd as expecting support and compassion from a screwdriver.

However, in the past few decades computers have populated every house and every pocket. People use them to communicate with each other, express themselves, to share their experiences, and to satisfy their social needs. Technology plays a huge role in the social interactions in the modern era. If computers became so important for the social relationships between people, shouldn't they be better suited for the job? We believe that just as the first computers were suitably equipped for calculations, modern computers used for social interactions should be better suited for this use case. The importance of the emotions for UX (user experience) is also well recognized by the HCI community[14, 22, 27].

In fact, people are constantly expressing their emotions to the machines, e.g. anger and swearing at a program does not compile, or love and compassion for an old but beloved laptop that is struggling to boot the newest operating system. How richer the user experience could have been if computers could recognize and react to those emotional expressions?

In this paper, we explore how an interactive system could recognize the emotional state of the users from their handwriting. Our approach is based on extracting features related to the fine motor performance related to procedural memory. Handwriting is a widespread example of such an activity. Handwriting is a traditionally non-digital way of recording information and conveying it to others. With the advent of the digital era, digital pens or styli allow users to digitize their writings and drawings instantly and store them on their digital devices. Digitized handwriting gives us access to precise handwriting data without requiring computer vision. Every stroke's data can be recorded and emotion-related features can be extracted as the user writes or draws.

Although the decline of handwriting is undeniable, handwriting still has it's place. Time journal published a piece on decline of handwriting in 1935, long before the digital revolution. In scientific literature decline of handwriting was mentioned as far back as 1965[10], also long before interpersonal text communication became widespread. On the other hand, recently released consumer market products, such as Apple Pencil and Microsoft Surface Pen attract a lot of attention and are commercially successful.

A survey of 168 subjects in front of Gongguan Station in Taipei shows that people overwhelmingly prefer handwriting for conveying sincere, emotional and personal messages, although for official communication they prefer typed documents. A 2010 online poll by Lifehacker.com showed that less than a quarter of the respondents take their notes exclusively by typing. 38.04% use only handwriting and 39.56% a combination of both methods. These results show that although handwriting is being replaced by typing in many areas, it is unlikely to disappear completely.

It is also worth noting the artistic value of writing and drawing by hand. Considering the important role that emotion plays in art, applying the presented approach for artistic activities may enable many interesting applications.

Contributions of this paper are the following:

- Describing the physiological mechanisms we intend to use for implicit emotion assessment.
- Presenting the feature set extracted from the digital handwriting recording that allows us to recognize the emotional condition of the writer.
- Describing and discussing the classification of digital handwriting into 4 classes of emotional conditions (High Valence - High Arousal, High Valence - Low Arousal, Low Valence High Arousal, Low Valence - Low Arousal).
- Presenting a user independent classifier with up to 66% classification accuracy for certain types of handwriting.

2 RELATED WORK

2.1 Theories of Emotion

The assessment and classification of emotion has received considerable attention in the field of psychology. Emotions can affect people's life in many aspects ranging from decision making to both mental and physical health. Though everyone seems to know what emotion is indistinctly, the fundamental concept of emotion is hard to define.

For decades, researchers have been continually developing alternative emotional measurement methods and technologies. There are several model of emotions that are widely accepted and validated. Mehrabian and Russel developed the PAD emotional state model to represent all emotions. the PAD model uses three dimensional scales: Pleasure/Valence, Arousal and Dominance [25]. Russell developed the Circumplex Model based on the PA part of PAD. In this model the emotions are distributed in a two-dimensional circular space, containing arousal and valence dimensions[28]. In this model, emotional states can be represented at any level of valence and arousal, or at a neutral level of one or both of these factors. PAD is also used by Lang and colleagues to develop a non-verbal pictorial self-assessment also known as SAM [24].

In the early 60s, Tomkins' proposed the eight basic emotion theory in which he defined eight basic emotions as follows. Two positive: Interest/ excitement and enjoyment/ joy, one neutral: Surprise/ startle, and five negative: distress/ anguish, fear/ terror, shame/ humiliation, contempt/ disgust and anger/ rage [32]. Based on Tomkin's affect theory, Lovheim proposed a new three-dimensional model for emotions and monoamine neurotransmitters also known as Lovheim cube of emotion [23]. In this model, the eight basic emotions located in each corner of an orthogonal coordinate system with Serotonin (5-HT, 5-hydroxytryptamine) is represented on the x-axis, noradrenaline (NE) on the y-axis and dopamine (DA) on the z-axis.

In scientific studies, emotion is best defined as a process that involves multiple responses rather than a single reaction[8]. Emotional processing involves both attentional engagement and behaviors preparatory to motivated action. Lang indicated there were at least three different systems in emotional response measuring: affective reports, physiological reactivity, and overt behavioral acts[21]. He also noted both physiological reactivity and overt behavioral acts often dictate clear preference for the scientific recordings[4]. As matter of fact, many prior studies demonstrated the possibility of emotion detection based on physiological signals such as Skin Conductance (SC also know as EDA), Heart Rate Variability (HRV), pupil dilation, respiration, temperature, etc. For instance, Skin Conductance (SC) is considered as one of the most sensitive markers and is frequently used to assess emotional arousal. Many studies found Skin Conductance Level (SCL) increases when emotional arousal increases[2, 3, 33].

2.2 Emotion Recognition

There are some attempts that researchers have made to improve the interaction between human and computer by helping computer to understand human's emotional context better. These attempts are based on different data sources. Shugrina presented an interactive emotional estimating and visualizing algorithms[30]. This approach is based on recognizing users' facial expressions through the detection of facial action units. Users' emotional ambiance can be estimated and later be mapped to the digital canvas after the rendering algorithm was performed.

Kirsch analyzed the results of three experiments and conducted a new experiment and developed the Sentic Mouse inspired by the results of emotional prediction theories[19]. The Sentic Mouse was an ordinary mouse augmented with a sensor to collect sentic data and proved to be able to capture valence information. Besides using mouse as the source of biometric characteristics, keyboard is also considered having the potential to help recognizing users' emotions as an input device. Clayton Epp conducted a research on determining user emotion by analyzing the rhythm of their typing patterns[11] and used the supervised machine learning algorithms for the classification. And the results proved that the keystroke dynamics is able to classify at least two levels of seven emotional states.

2.3 Handwriting Analysis

Handwriting analysis is found to be useful for biometric authentication and personality characteristics evaluation and emotion prediction.

Christian Hook introduced a device used for verification and identification of individuals by means of handwritten items[16]. Biometric test samples are compared against reference templates through the classification and matching algorithm using extracted features from the 3D pressure signals, which demonstrate a remarkable potential for authentication system.

Kedar has reviewed the processing steps in his paper which were Image Pre-processing, Feature Extraction and Classification [18]. Handwriting characteristics,such as baseline, slant, pen-pressure, were also mentioned as analyzing features for the recognition of emotional outlays and other personality traits. Janet Fisher focused on the ability of automatic handwriting analysis in identifying traits in violent behavior[13]. Multiple samples from incarcerated violent offenders were compared with non-violent offenders. The results were found to be helpful in violent behavior prediction.

Michael Fairhurst worked on the enhancement of handwriting's forensic value to predict users' specific emotional state, which were the "happy" and the "stressed"[12]. For data processing, classification is performed by using KNN, Jrip and SVM classifier. SVM classifier performed best for the prediction in both cases with accuracy close to 80% in the "happy" emotion state and 70% in the "stressed" emotion state.

2.4 Neural Substrate of Emotion

SC, heart rate, respiratory rate, and pupil dilation, are all controlled by the autonomic nervous system (ANS). ANS is the part of the peripheral nervous system that controls involuntary functions that are critical for survival. Within the brain, the ANS is located in the medulla oblongata in the lower brainstem. During emotional stimulation, projections from the central nucleus or the bed nucleus of the stria terminalis (BNST) to the ventral tegmental area appear to mediate increases in dopamine metabolites in the prefrontal cortex [15].

Most of the methods of implicit assessment of the emotional state of the subject without self-assessment rely on recording physiological signals related to the ANS activity. This approach seems very reasonable, as there are multiple works demonstrating how emotions influence the ANS[9]. This, combined with the lack of voluntary conscious control over the ANS, makes ANS a very interesting source of information regarding the cognitive or emotional state of the subject.

In this paper we are approaching the emotion recognition from a slightly different angle by trying to extract features from handwriting. Handwriting can be seen as a fine motor task highly dependant on procedural memory (responsible for "muscle memory"). The reliance on procedural memory rather than deliberate conscious control of every hand muscle movement presumes the low degree of conscious control over the handwriting process. Which would make it possible for unconscious manifestation of certain emotion-related features. This could be explained by the change in neurotransmitter quantities associated with emotional experiences, particularly the monoamine family: serotonin, dopamine and noradrenaline[23]. In addition to this there are several studies demonstrating the effects of ANS activity and neurotransmitter balance on motor activity in humans as well as in birds and other mammals[1, 7, 29].

Based on this, we are making an assumption that the experienced emotions can manifest through the motor tasks, such as handwriting in our case. The research question for this study is whether there is a correlation between these manifestations and the emotional condition of a given subject. Furthermore, if the correlation exists can it be generalized to a wider population or is it user-dependant.

3 EXPERIMENTAL SETUP

To verify our hypothesis we conducted a series of experiments. The first study was a preliminary test with a small number (N=4) participants. The results of the preliminary study were used to test the recording setup, get a sense of what kind of data is it possible to record using our apparatus and what would be the appropriate approach to data processing as well as gather qualitative user feedback. Second study uses the same apparatus as the first one with some changes in the tasks performed by the users. Methodology,

Figure 1: Participant during the experiment

apparatus and the tasks performed by the participants are described further on in this section.

3.1 Method Overview

In order to induce certain emotions we use a series of emotionally-charged video clips. Due to short length of the clips we group them in sets of 4 clips for each possible combination of valence and arousal: High Valence with High Arousal (HV-HA), High Valence with Low Arousal (HV-LA), Low Valence with High Arousal (LV-HA), and Low Valence with Low Arousal (LV-LA). The length of each set is 160 seconds. After watching each series of clips participants were asked to fill in a self-assessment form and to perform a series of writing tasks. Dominance was not included in the test because the high dominance videos were overlapping with high arousal and/or valence. Another reason for leaving out dominance is the test session length, adding high and low dominance to the existing 4 sets would require 8 testing sessions that would result in increased fatigue of the participants, which could possibly affect the data.

3.2 Materials

This section describes the materials used as emotion stimulus, emotion measurement and parts of the handwriting tasks.

3.2.1 The Emotional Movie Database. Film clips from the Emotional Movie Database (EMDB)[6] was used as emotion stimulus to elicit emotional states of the participants. EMDB is a database of 52 non-auditory film clips with different ratings of valence, arousal and dominance based on the self-assessment manikin (SAM)[4]. The duration of each film clip is 40 seconds. In this study, 16 film clips are selected and classified into 4 groups based on the arousal levels and valence levels due to their rating scores provided. Since the SAM scores for the film clips are split by gender, we prepared separate sets for male and female participants.

3.2.2 The Self-Assessment Manikin. The Self-Assessment Manikin (SAM)[24] was used to measure emotion states of participants after watching each film clip. SAM is a non-verbal pictorial self-report assessment associate with the reporter's affective reaction to a stimuli. It was designed to measure three aspects- pleasure, arousal and dominance.

3.2.3 The Affective Norms for English Words. The Affective Norms for English Words(ANEW)[5] was used as written materials to refer in the handwriting task. ANEW is a database of approximately 600 hundreds English words with different ratings of valence, arousal and dominance based on the self-assessment manikin(SAM). The most neutral words were selected separately for male and female participants based on the SAM scores.

3.3 Apparatus

The film clips were presented on a 49 inch 4K TV. For each condition we used 4 film clips with no pauses in between. Total duration of emotional stimulation for each condition was 160 seconds. Handwriting tasks were performed on a first generation Apple iPad pro 12.9 inch tablet with Apple pencil stylus. The software running on the tablet was recording each touch event and streaming the data wirelessly to the experimenter's PC that was simultaneously streaming the video clips to the TV.

The following information was recorded on the iPad: X and Y coordinates of the stylus, OS time with millisecond resolution, altitude angle, azimuth angle, force, touch phase (touch start, touch moved, touch holds, tracking lost, finger/stylus lifted). The software on the experimenter's Computer was appending the PC timestamp to each sample received for synchronization with the tasks performed and video clips, as the task and video timing was controlled by the PC. All the data received from iPad as well as timestamps related to tasks and video clips were recorded.

3.4 Preliminary Test

Preliminary to the experiment we have conducted a similar test with 4 participants to validate and verify the setup, gather qualitative feedback, observe the participants' behaviour during the study and get a better understanding of the data being recorded. The preliminary test procedure was as follows. Each participant had to give written inform consent to participate in the study, was warned about possibly ethically or morally disturbing film clips in the test and was informed about their right to stop the test at any moment in time.

In total the test was made out of 5 sessions of nearly identical structure: watching the video series, self assessment using SAM, then performing three written tasks. The first task was to re-write the list of 15 random emotionally neutral words appearing on the TV screen on the tablet. Second task was identical to the first, except the word list was different and participants had 30 seconds to finish the task. The time limit was introduced to force the participants to write fast, which was presumed to require higher reliance on muscle memory and less conscious control over the handwriting. In the last task participants were asked to write a line of the following characters: o, /, |, -, :, +. The motivation for the last test was to reduce the complexity of the strokes to in order to simplify the further analysis.

This session structure was repeated for each video category (HV-HA, HV-LA, LV-HA, LV-LA), except the baseline session which was done prior to the sessions with video clips, that did not include any videos. The order of the sessions with videos was counterbalanced.

Table 1: 16 stroke features selected as most suitable for classification

Name	Description
ALT_ANGL	Altitude(in radians) of the stylus
AZ_ANGL	Azimuth(in radians) of the stylus
PATH_AZ_ANGL	Total angular path of azimuth angle
SPEED_AL_MIN	Minimal speed of the altitude angle (from 25ms windows)
SPEED_AZ_MIN	minimal speed of the azimuth angle (from 25ms windows)
SPEED_MAX	Maximum of stroke speed
SPEED_STD	Standard deviation of stroke speed
SPEED_X_MIN	Minimal speed on the X axis (from 25ms windows)
SPEED_Y_MIN	Minimal speed on the Y axis (from 25ms windows)
FORCE	Average force of the stylus touch
FORCE_SD	Standard deviation of stroke force
PATH_X	Total stroke path on the X axis
PATH_Y	Total stroke path on the Y axis
TIME	Duration of the stroke
TIME_FROM_LAST	Time from the last stroke
SAMPLES	Number of samples in the stroke

The recorded handwriting data was split into separate strokes, for each stroke the following features were calculated: start and end time, duration in time, stroke length in pixels, speed SD, number of samples, average force, force SD, average altitude angle and its SD, average azimuth angle and its SD. Using a KNN classifier we achieved up to 51% accuracy with 4 groups (HV-HA, HV-LA, LV-HA, LV-LA) for user-independent model and up to 66% for user-dependant. Features related to altitude and azimuth angle seem to have the strongest effect on the classification accuracy. Although the number of participants was too small to consider, it supports the viability of our approach and apparatus. This pretest also showed the poorest results for the task number one, which was later removed from the final study. By observing the behavior of the participants it was noted that sometimes they get distracted and not pay attention to the videos, which was also addressed in the final study. Due to the small number of the participants we will omit the in-depth discussion of the results, as this test was intended to serve as a technical test of the setup. However its influence on the final study design is worth mentioning.

3.5 Experiment

3.5.1 Participants. Thirteen volunteers aged 19 to 40 years (mean:24, SD:5 ;female: 8, male: 5) were recruited for this study. Every participant received a 1000 JPY gift-card as a compensation for their time, since the whole experiment takes each participant 30 to 40 minutes.

3.5.2 Procedure. The procedure of the final experiment is very close to the preliminary test. To summarize, the structure of the study was the following.

(1) Written informed consent.
(2) Baseline SAM test and recording session without film clips.
(3) Four recording sessions, one session for each film clip category (HV-HA, HV-LA, LV-HA, LV-LA). Order of the sessions is counterbalanced to avoid ordering effects.

Each session, except the baseline recording that did not include film clips, is structured as follows. The orders of writing tasks in each session remains the same for all the participants.

(1) 4 EMDB film clips, 160 seconds in total. Order of the clips is randomized each time.
(2) SAM test for perceived Valence, Arousal and Dominance levels.
(3) Task 1. Rewrite the words from the screen with time limit of 30 seconds. After the time limit the words disappear. The test is intended to force the participants to write fast which would increase the reliance on the procedural memory and weaken the conscious control over the fine motor performance. This task is aimed to gather normal handwriting data.
(4) Task 2. Free doodling while watching the same video clips as in P.1 of this list. Order of clips is random. During this task we expect to record data related to drawing.
(5) Task 3. Write a line of the following characters: o, /, |, -, :, +. the test is added to record very basic and repetitive strokes which could simplify the analysis of the stroke features.

The main difference from the preliminary test is the replacement of the task 1 with task 2 and for the task two the participants were ask to freely doodle on the tablet while watching the film clips again. The duration of the new task 2 is 3 minutes. For the baseline condition recording participants were asked to doodle in absence of any video stimuli.

The total duration of the test with each participant was between 30 and 45 minutes. The experiments were conducted in a sound-proof audio recording studio, with TV screen covering most of the participant's field of view. The only objects on the desk in front of the participant were SAM test sheets, tablet with a stylus and a charging cable.

3.5.3 Feature Extraction. In total throughout 5 session with 13 participants we recorded 1442789 samples of the pen position that were split into 14282 strokes based on the touch phase data provided by the tablet. Recordings for a few test sessions were lost or discarded due to technical or network reasons. For further analysis we used only the sessions that were reliably recorded from start to finish.

For each stroke was described by the following parameters, 44 in total. Later we bring the number down to 16 most usable features (See Table 1).

- Stroke start and end time from experimenter's PC and tablet.
- Stroke length in pixels (total and on X and Y axis separately) and in millimeters.
- Stroke duration.
- Stroke SD of the distances between each sample in the stroke.
- Number of samples in the stroke.
- Average force and its SD.
- Average altitude and azimuth angles, their SD and total angular path.
- Time from the last stroke.
- Features extracted by splitting the stroke into 25ms windows with 50% overlap:
  - Top window speed on X and Y axis, on both axes combined, angular speed of azimuth and altitude angles and maximal force.
  - Lowest window speed on X and Y axis, on both axes combined, angular speed of azimuth and altitude angles and minimal force.

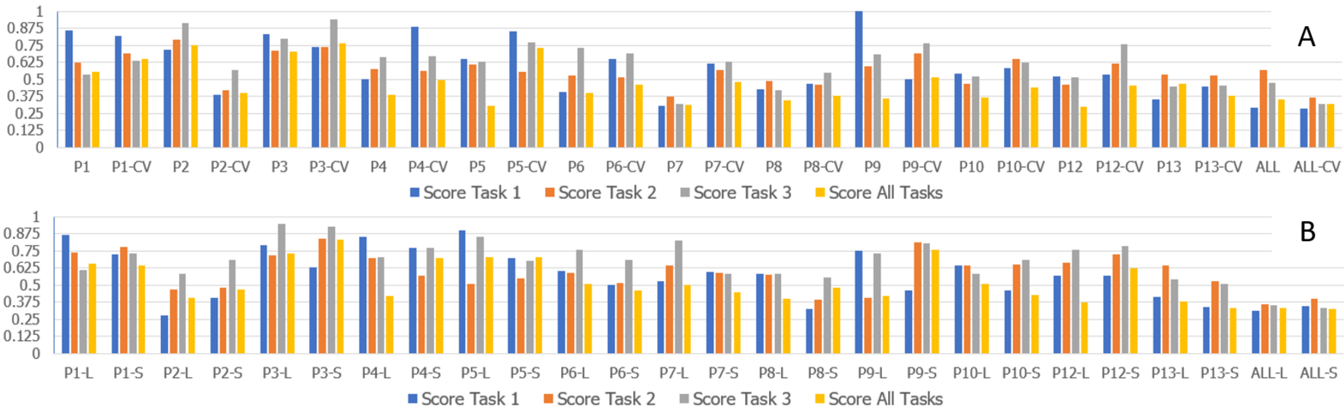

Figure 2: Classification scores for user-dependent model. A. Test scores sing all strokes, CV - cross validation scores for each participant's model. B. Models using only short (S) and long (L) strokes. "Score All tasks" is the task-independent classifier, "ALL" represents data for user-independent classifier.

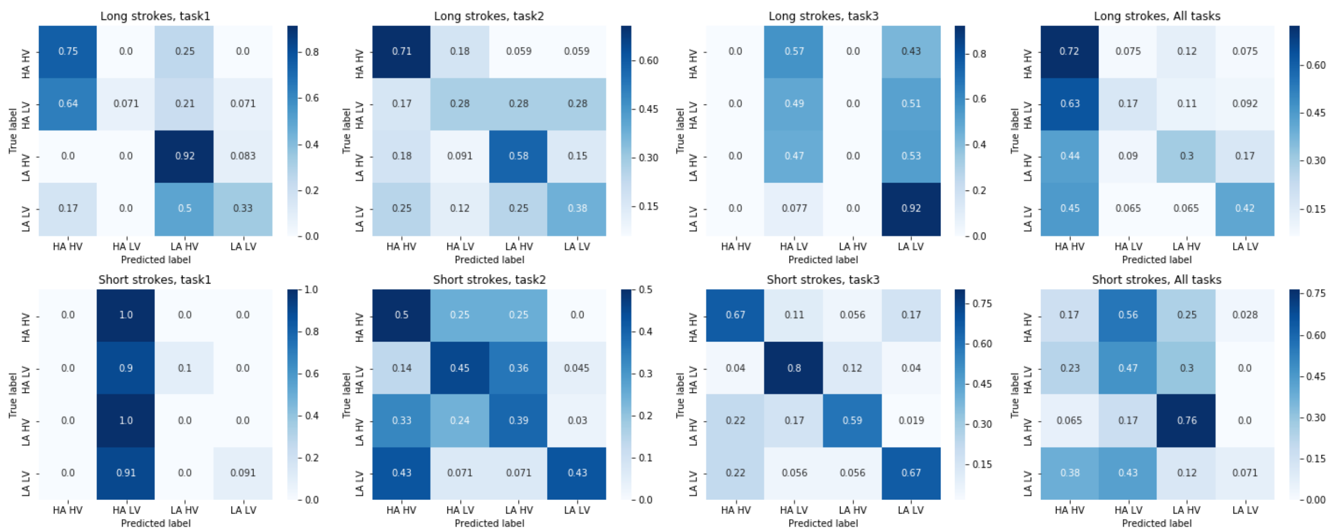

Figure 3: Confusion matrices for user-independent model. The stroke data was split into short (bottom) and long (top) strokes before classification.

- Maximal speed increase between neighboring windows on X and Y axis, on both axes combined, maximal angular speed increase of azimuth and altitude angles and maximal force increase.
- Maximal speed decrease between neighboring windows on X and Y axis, on both axes combined, maximal angular speed decrease of azimuth and altitude angles and maximal force decrease.

Excluding stroke start and end time we analyzed the remaining 40 features in order to test whether there is any relation between the handwriting features and the self-reported emotional state of the participant recorded using the SAM. In order to do this we used the C-Support Vector Classification (SVC) from scikit-learn library [26] with valence, arousal and dominance ratings as labels.

Preliminary data analysis showed significant ($p < 0.000$) difference between one-to-one ANOVA test of multiple features (e.g. angles, pressure, time from the last stroke, etc.) of strokes rated as high valence, arousal or dominance and strokes rated versus strokes rated as low valence, arousal or dominance.

In order to select the most promising features we have calculated the Gini Importance for each feature for each of the three tasks separately. Then selected and combined top 10 features from each of the tasks. The resulting list contains 16 features and is shown in the Table 1. The details on the SVC and its performance is discussed in the following section.

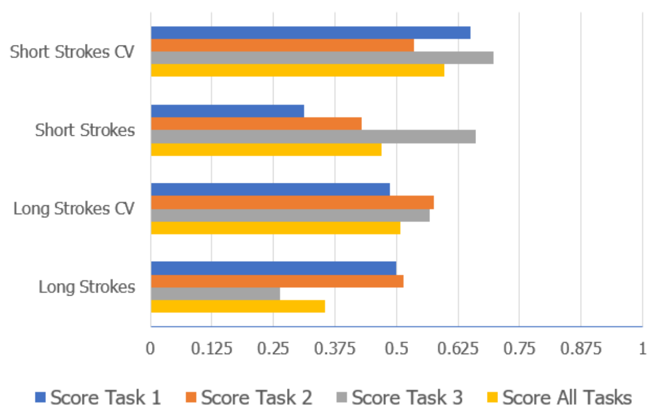

Figure 4: Classification scores for user-independent model. CV - cross validation results. The stroke data was split into short and long strokes before classification.

4 RESULTS AND DISCUSSION

To assert the effect of the chosen emotional stimuli, we calculated the correlation between the SAM scores reported in the EMDB and those self-reported. For valence scores the correlation was 0.905. For arousal scores the correlation was 0.959. This confirms that the effect of emotional stimulation was present and as expected.

The classifier we used is SVC from scikit-learn library. Before training, we split all the data into train data, and test data (train to test ratio is 9 : 1). In order to make sure that the classifier does not overfit on the train data, we used a 5-fold cross-validation for the train data and calculated the average of the 5 accuracy scores as the training score of the classifier (Fig.2.A (CV)). When training the classifier, we tuned it for the best training score by shifting the hyperparameter C and gamma to optimize each classifier. We then test the classifier on the test data and calculate the test score. (Fig.2.A).

4.1 User-dependent Model

The user-dependent models were made using the type of the video content as the label (HV-HA, HV-LA, LV-HA, LV-LA). Due to data loss for participants 9 and 11 results for task 1 of participant 9 are irrelevant, and participant 11 was completely removed from the analysis.

The classifier scores for all strokes combined, short and long separately and cross validation scores are presented on Fig.2. The results represent data for all 12 participants individually and the last chart for all participants combined. For each participant we present results split into 3 groups: for all strokes, for short and long strokes. In each group we have classifiers for tasks 1, 2, 3, and all tasks combined.

For all strokes combined, the accuracy for 1 out of 4 classification averaged between all tasks for all 12 users is 58.4% (Score Task 1,2,3 on Fig.2A), 43% for the task-independent classifier (Score All Tasks on Fig.2A). Data for short and long strokes separately is presented on Fig.2B. For short strokes the accuracy averaged between all tasks for all 12 users is 62%, task-independent is 57%. For tasks 1, 2 and 3 it is 54%, 61% and 69% respectively. For long strokes the accuracy averaged between all tasks for all 12 users is 65%, task-independent is 50%. For tasks 1, 2 and 3 it is 64%, 60% and 70% respectively. The accuracy of user independent classifier is 28%, 57% and 47% for the tasks 1, 2, and 3 respectively. Accuracy of task-independent user-independent classifier was 35%.

4.2 User-independent Model

In order to build the classifier we used the valence and arousal scores of the SAM test as labels. The scores were grouped into "high" (6-9 points) and "low" (1-4 points) categories. The data scored at 5 points of valence or arousal was omitted. This assures that the actual arousal and valence levels of the participant correspond to the handwriting, since not all the participants had intended emotional response to the video stimuli.

The confusion matrices of each condition are presented on Fig.3, and the score results of this classification are presented on Fig.4. We ran classification on the data sets from each task and on all three tasks combined. It was found that the classification precision changes greatly if we use only short or only long strokes. The stroke data was split into short and long strokes in relation to the median for each data set used. Short strokes gave particularly good results for task 3, as it required participants to draw mostly short lines. Accuracy for this test is reaching 66% for 1 in 4 groups classification. Using long strokes showed better results for tasks 1 and 2 with accuracy of 50 and 51% respectively. Surprisingly, the accuracy of classification of the short stroke data set for all 3 tasks was higher than for long stroke data set (47% for short and 35% for long).

On the downside, although this approach provides a reliable labeling, we have to exclude all the data that did not fit into one of the four labeled groups: 1. High Arousal and Valence. 2. High Arousal low Valence. 3. Low Arousal and high Valence. 4. Low Arousal and Valence. After applying this labeling to the stroke data, the data for some of the participants was not sufficient to build a reliable classifier. Thus this approach was used only for user-independent model that was trained on the data from all the users.

4.3 Results Summary

In both, user-dependent and user-independent models length-based separation of the strokes clearly resulted in increased accuracy. The possible explanation for this may be that certain features manifest themselves differently depending on the type of the strokes and tasks. Meaning that check marks or crosses may convey information related to the emotional states different from long strokes used for drawing or doodling. Especially for task3 (simple and repetitive characters), it showed better performance on short strokes, as it required participants to write short lines. For different tasks, it may be necessary to train different models for the classification.

The differences in performance of user-dependent and user-independent model suggest that the differences between individual manifestations of the emotional states in fine motor performance are present and have to be considered for such applications.

5 CONCLUSIONS AND FUTURE WORK

This paper presents the conducted experiments and analysis of the collected emotionally-labeled handwriting data. Out of 44 extracted features, we selected 16 most promising features according to their Gini Importance. Based on this feature set, we presented a series of classifiers, using the type of the video content as the label for the user-dependent model and the SAM scores as the label for the user-independent model. For user dependent 4-class classifier, we got the classification accuracy up to 70% for certain tasks. For user independent 4-class classifier, we got the classification accuracy up to 66% for a certain writing task.

This work shows the potential to build a system capable of recognizing people's emotions, which could give us a better insight into how the human computer interaction could evolve and even how we could improve the communication between people.

In total we gathered 270 MB of handwriting data. In the future we plan to polish our classification algorithms and try several other approaches to data analysis, such as deep learning. It would be very interesting to apply this method to more artistic activities, such as painting or drawing. In order to do this we plan to record additional data, since the current data set contains only a limited number of types of handwriting. We are also interested in developing a prototype display of emotional condition based on the recognition of the user's emotional state from handwriting and physiological signals in real time, which may enable people to be more aware of both, their emotional feelings and that of the surrounding people.

ACKNOWLEDGMENTS

The work is supported by the Japan Science and Technology Agency CREST program under Grant No: JPMJCR16E1.

REFERENCES

[1] Kiyoshi Ando, Chris E. Johanson, Lewis S. Seiden, and Charles R. Schuster. 1985. Sensitivity changes to dopaminergic agents in fine motor control of rhesus monkeys after repeated methamphetamine administration. *Pharmacology Biochemistry and Behavior* 22, 5 (1985), 737 – 743. https://doi.org/10.1016/0091-3057(85)90522-2

[2] John L Andreassi. 2010. *Psychophysiology: Human behavior and physiological response*. Psychology Press.

[3] Wolfram Boucsein. 2012. *Electrodermal activity*. Springer Science & Business Media.

[4] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[5] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report. Citeseer.

[6] Sandra Carvalho, Jorge Leite, Santiago Galdo-Álvarez, and Óscar F Gonçalves. 2012. The emotional movie database (EMDB): A self-report and psychophysiological study. *Applied psychophysiology and biofeedback* 37, 4 (2012), 279–294.

[7] Christian Collet, Franck Di Rienzo, Nady Hoyek, and Aymeric Guillot. 2013. Autonomic nervous system correlates in movement observation and motor imagery. *Frontiers in Human Neuroscience* 7 (2013), 415. https://doi.org/10.3389/fnhum.2013.00415

[8] Lang Peter J Davis, Michael. 2003. *Handbook of psychology, Volume 3: Biological psychology*. John Wiley & Sons, Inc. 405–439 pages.

[9] P Ekman, RW Levenson, and WV Friesen. 1983. Autonomic nervous system activity distinguishes among emotions. *Science* 221, 4616 (1983), 1208–1210. https://doi.org/10.1126/science.6612338 arXiv:http://science.sciencemag.org/content/221/4616/1208.full.pdf

[10] E. A. Enstrom. 1965. The Decline of Handwriting. *The Elementary School Journal* 66, 1 (1965), 22–27. https://doi.org/10.1086/460256 arXiv:https://doi.org/10.1086/460256

[11] Clayton Epp, Michael Lippold, and Regan L Mandryk. 2011. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 715–724.

[12] Michael Fairhurst, Meryem Erbilek, and Cheng Li. 2014. Enhancing the forensic value of handwriting using emotion prediction. In *Biometrics and Forensics (IWBF), 2014 International Workshop on*. IEEE, 1–6.

[13] Janet Fisher, Anish Maredia, Anita Nixon, Nerissa Williams, and Jonathan Leet. 2012. Identifying personality traits, and especially traits resulting in violent behavior through automatic handwriting analysis. *Proceedings of Student-Faculty Research Day, CSIS, Pace University* (2012).

[14] Jodi Forlizzi and Katja Battarbee. 2004. Understanding Experience in Interactive Systems. In *Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques (DIS '04)*. ACM, New York, NY, USA, 261–268. https://doi.org/10.1145/1013115.1013152

[15] Lee E Goldstein, Ann M Rasmusson, B Steve Bunney, and Robert H Roth. 1996. Role of the amygdala in the coordination of behavioral, neuroendocrine, and prefrontal cortical monoamine responses to psychological stress in the rat. *Journal of Neuroscience* 16, 15 (1996), 4787–4798.

[16] Christian Hook, Juergen Kempf, and Georg Scharfenberg. 2004. A novel digitizing pen for the analysis of pen pressure and inclination in handwriting biometrics. In *International Workshop on Biometric Authentication*. Springer, 283–294.

[17] Roy A Huber and Alfred M Headrick. 1999. *Handwriting identification: facts and fundamentals*. CRC press.

[18] SV Kedar, DS Bormane, Aaditi Dhadwal, Shiwali Alone, and Rashi Agarwal. 2015. Automatic emotion recognition through handwriting analysis: a review. In *Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on*. IEEE, 811–816.

[19] Dana Kirsch. 1997. *The Sentic Mouse: Developing a tool for Measuring Emotional Valence*. Technical Report. MIT Media Laboratory Perceptual Computing Section.

[20] Richard J Klimoski and Anat Rafaeli. 1983. Inferring personal qualities through handwriting analysis. *Journal of Occupational Psychology* 56, 3 (1983), 191–202.

[21] Peter J Lang. 1969. The mechanics of desensitization and the laboratory study of human fear. *Behavior therapy: Appraisal and status. New York: McGraw-Hill* (1969), 160–191.

[22] Effie Lai-Chong Law, Virpi Roto, Marc Hassenzahl, Arnold P.O.S. Vermeeren, and Joke Kort. 2009. Understanding, Scoping and Defining User Experience: A Survey Approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 719–728. https://doi.org/10.1145/1518701.1518813

[23] Hugo LÃűvheim. 2012. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses* 78, 2 (2012), 341 – 348. https://doi.org/10.1016/j.mehy.2011.11.016

[24] P.Lang M.Bradley. 1994. Measuring Emotion: The Self-Assessment Semantic Differential Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25 (1994), 49–59. Issue 1.

[25] Albert Mehrabian. 1980. Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies. (1980).

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[27] Rosalind W. Picard. 1999. Affective Computing for HCI. In *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I - Volume I*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 829–833. http://dl.acm.org/citation.cfm?id=647943.742338

[28] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[29] Deanna M Sasaki-Adams and Ann E Kelley. 2001. Serotonin-Dopamine Interactions in the Control of Conditioned Reinforcement and Motor Behavior. *Neuropsychopharmacology* (2001). https://doi.org/10.1016/S0893-133X(01)00240-8

[30] Maria Shugrina, Margrit Betke, and John Collomosse. 2006. Empathic Painting: Interactive Stylization Through Observed Emotional State. In *Proceedings of the 4th International Symposium on Non-photorealistic Animation and Rendering (NPAR '06)*. ACM, New York, NY, USA, 87–96. https://doi.org/10.1145/1124728.1124744

[31] Sargur N Srihari, Sung-Hyuk Cha, Hina Arora, and Sangjik Lee. 2002. Individuality of handwriting. *Journal of forensic science* 47, 4 (2002), 1–17.

[32] Silvan Tomkins. 1962. *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company.

[33] Marieke van Dooren, Joris H Janssen, et al. 2012. Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & behavior* 106, 2 (2012), 298–304.