Smart Media Journal / Vol.13, No.7 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2024.13.7.18

# A Review of Public Datasets for Keystroke-based Behavior Analysis

Kolmogortseva Karina, Soo-Hyung Kim, Aera Kim

**Abstract**

One of the newest trends in AI is emotion recognition utilizing keystroke dynamics, which leverages biometric data to identify users and assess emotional states. This work offers a comparison of four datasets that are frequently used to research keystroke dynamics: BB-MAS, Buffalo, Clarkson II, and CMU. The datasets contain different types of data, both behavioral and physiological biometric data that was gathered in a range of environments, from controlled labs to real work environments. Considering the benefits and drawbacks of each dataset, paying particular attention to how well it can be used for tasks like emotion recognition and behavioral analysis. Our findings demonstrate how user attributes, task circumstances, and ambient elements affect typing behavior. This comparative analysis aims to guide future research and development of applications for emotion detection and biometrics, emphasizing the importance of collecting diverse data and the possibility of integrating keystroke dynamics with other biometric measurements.

Keywords : affective computing | emotion detection | data collection | biometric authentication | mood analysis

## I. INTRODUCTION

Information sources that can provide information about the emotional state of a person can be: textual, audio, physical, visual, and behavioral. These sources include a person's environment, his experiences, as well as his psychological and physical state, which can describe his emotional state. By understanding these conditions, we can prevent unpleasant situations such as stress, fatigue, depression, etc. Emotions are dynamic, so they need to be measured constantly. Therefore, emotion detection is one of the trends in the field of AI recently.

Biometrics is a unique set of data about a person, with the help of which a person can be identified; it cannot be lost or stolen, as it is always with a person. There are several types of biometrics: physiological and behavioral. Physiological biometrics include factors such as hand geometry and pupils. Behavioral biometrics is also called passive because the user does not have to do anything while working, just behave as usual.

Biometric authentication is based on the behavioral traits of the individual being identified, as opposed to traditional authentication techniques, which usually rely on static identifiers like passwords or PINs. The use of human physiological characteristics as a means of identification has become widespread.

* This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS- 2023-0219107), and the Institute of Information & communications Technology Planning & Evaluation (ITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (ITP-2023-RS-2023-025629) grant funded by the Korea government(MSIT).

Manuscript : 2024.06.17
Revised : 2024.07.22, 2024.08.02
Confirmation of Publication: 2024.08.05
Corresponding Author: Soo-Hyung Kim e-mail: shkim@jnu.ac.kr

The problem with physiological information is that it requires secure storage, as the user needs to share his personal information. Advances in AI have opened up new possibilities for authentication and identification of users using an analysis of the user's interactions with the device.

User behavior can be used to identify individuals, and this requires storing large amounts of data. The stored data is used to develop average human behavior, so it improves the accuracy of identification when the user is tired, drunk, in a hurry or in another state.

The methodology based on measuring key dynamics [1] follows many of the same principles as studying handwriting, but modern technology allows for many different ways for the user to measure. The simplest way to extract user data from keystrokes is to measure keystroke time, which is the time between pressing two consecutive keys, and to measure keystroke time, which is the amount of time that a single key is directly physically touched. Analyzing this data, it is possible to authenticate and identify the user, as well as measure his stress level, emotional state, and fatigue.

Using the dynamics of the keys, we can analyze not only the content, but also the typing technique, combining different features, and even combining with physiological data is possible. Systems can determine whether the person typing the text is a genuine user or an imposter by identifying patterns in time and the user's “typewriting.”

As mentioned above, keypress dynamics can also be used to determine a person’s emotional state.

The study's primary goals are as follows:

- to provide a comprehensive analysis of four widely-used datasets (BB-MAS [2], Buffalo [3], Clarkson II [4], and CMU [5]) for keystroke dynamics research;
- to understand what kind of information might be more useful for different type of tasks;
- to determine each data set's benefits and drawbacks to direct future studies and applications.

These patterns are influenced by an individual's motor talents, mental state, habits, and environmental circumstances. The written characters and the moment of each keystroke are typically included in the logged data. AI compares the typed pattern with a stored biometric template during verification to verify whether a user is on the list of authorized users.

Recognition of emotions is predicated on subtle differences in behavior patterns in different emotional states. For example, when the person is stressed, he may type slower and use the “Backspace” button more often, while when feeling at ease, they may type faster and more smoothly.

### 1.1 Role of AI in Keystroke Dynamics

AI plays an important role in analyzing keystroke-based behavior, using advanced algorithms to analyze and interpret complex typing patterns. Machine learning and deep learning models allow us to accurately recognize emotional states by subtle changes in keystrokes. And not only improves the accuracy of emotion identification systems but also facilitates real-time analysis and adaptation. For example, AI-based models can integrate keystroke data with other biometric signals to create reliable emotion recognition systems that enhance user safety and improve mental health monitoring The scalability and automation offered by AI can change the sector, allowing for large-scale future advances in AI will continue to push the boundaries of what is possible in keystroke dynamics research and emotion recognition. AI is highly relevant to this research topic because it can effectively handle and analyze ambiguous data like keystroke dynamics and emotional states and can process vast amounts of data quickly and efficiently, enabling the real-time analysis of user emotions.

## II. RELATED WORK

Research on emotion recognition based on how a person types is becoming more popular by combining biometrics and human-computer interactions. A systematic review of the literature (SLR) conducted by [6] emphasizes the effectiveness of this analysis, because even a small change in emotions causes a response in human behavior, including in muscles, and the speed of typing depends on what state the person is in. By conducting this analysis, the authors present a comprehensive view of the effectiveness of using keyboard dynamics. They are researching methodologies that have been used over the past 10 years and admitted the lack of data sets related to emotion recognition by keystroke dynamics.

This study [7] shows the correlation between keyboard strokes and human emotions. Many applications have utilized keystroke dynamics and mouse movements to infer users' emotions. These studies address various challenges in data collection, data representation, and classifier training. The authors point out that data collecting is time-consuming because participants must convey their feelings in a natural way. The findings suggest that it might not be the best idea to create a single multi-class universal model that works for all users and emotional states. Emphasizing the greater people's variability of expressing some emotions through the dynamics of keystrokes, while other emotions can be expressed more universally. Collecting larger and more representative datasets and studying optimal sets of functions for different emotional states can improve the accuracy of emotion detection.

This research [8] has concentrated on differentiating the mood of the text—between positive and negative opinions—based on typing patterns. Participants wrote opinions about their good and bad learning experiences, and this information was logged. Additionally, the research analyzed keystroke dynamics in relation to the Pleasure-Arousal-Dominance (PAD) model, providing valuable insights into the emotional states of participants. The study also employed the Self-Assessment Manikin (SAM) scale to find out the participants' self-reported degrees of dominance, arousal, and pleasure. This comprehensive approach showed the possible connections between keyboard patterns and emotional states, including differences in arousal and pleasure depending on the kind of opinion stated. These metrics, particularly the significant differences in pleasure and arousal levels, offer a nuanced understanding of emotional expression through typing behavior. The included characteristics of keystroke dynamics were the number of keystrokes per second, and the frequency of using the specific keys, such as spacebar, backspace and delete keys. These features were critical in identifying variations in typing behavior for different emotional states. Overall, although keystroke dynamics provide insight into emotional states during human-computer interaction (HCI), further studies involving a variety of data, such as age, gender, technical skills, typing experience, and fatigue, are needed to expand these findings.

## III. Comparative Analysis of Datasets

### 3.1 CMU Dataset

Dataset Description

The CMU dataset [5] stores data from 47 people who entered the 10-digit password “.tie5Roanl”. The time series data include H.key – hold times of the key, DD.key1.key2– keydown-keydown and UD.key1.key2 – keyup-keydown.

Acquisition Process

Data collection was controlled by a laboratory environment. Each user typed the password 50 times in 1 session, 8 times, for a total of 400 lines per user.

Data Characteristics

Data Characteristics are detailed in the table 1.

Table 1. Data Characteristics of CMU Dataset.

| Characteristics | Description | Example Entry |
|---|---|---|
| Subject | Identifier for each subject (s002 to s057) | S002 |
| Session | Session in which the password was typed (1 to 8) | 1 |
| Repetition | Repetition of the password within the session (1 to 50) | 1 |
| Hold Times (H) | Time from key press to release | 0.1491 |
| Keydown-Keydown Times (DD) | Time from pressing one key to pressing the next | 0.3979 |
| Keyup-Keydown Times (UD) | Time from releasing one key to pressing the next | 0.2488 |

31 columns present the timing information for the password.

Applications and Insights

The CMU dataset is ideal for studying the evolution of typing behavior over time and for applications in user authentication and behavior analysis under controlled conditions.

### 3.2 Buffalo Dataset

Dataset Description

Three sessions of data collection from 148 individuals comprise the Buffalo dataset. Every meeting lasted roughly fifty minutes. The participants were divided into two groups: 75 always typed on the same keyboard, and 73 alternated between three keyboards.

Acquisition Process

Data was gathered in a laboratory. The tasks that the participants completed included answering questions, describing paintings, and transcribing a portion of Steve Jobs' lecture. Information about keys pressed and released, utilizing various keyboards, and doing various tasks are all included in the dataset. Everybody's gender is also included.

Data Characteristics

Data Characteristics are showed in the

table 2.

Table 2. Data Characteristics of Buffalo Dataset.

| Characteristics | Description | Example Entry |
|---|---|---|
| Key | Name of the pressed key | A |
| Event | Key event (key down or key up) | KeyDown |
| Timestamp | Time in milliseconds | 63578429792961 |
| Filename Details | Includes username, keyboard type, session, task number | Provided |

Details about Demographics

The analysis can be aided by the age and gender details included in the dataset.

Applications and Insights

This data set contains both free and fixed text tasks, as well as gender information, which makes it possible to analyze the typing behavior of different types of tasks for different groups of people. This is especially useful for studies predicting age and gender.

### 3.3 BB-MAS Dataset

Dataset Description

117 individuals took part in the data collection in lab and real-world settings, using different devices (desktop, phone, tablet).

Acquisition Process

Data was collected while users were doing activities, such as typing, browsing, walking, climbing stairs, and climbing upstairs.

Data Characteristics

Data Characteristics are presented in the table 3.

Table 3. Data Characteristics of BB-MAS Dataset.

| Characteristics | Description | Example Entry |
|---|---|---|
| Event ID (EID) | Unique identifier for each event | 0 |
| Key | Pressed key | t |
| Direction | 0 for key down, 1 for key up | 0 |
| Time | Date-time format | 2019-04-14 18:09:41.538 |

Applications and Insights

This dataset opens the possibility of considering how different typing changes depending not only on the device but on changes in physiological indicators.

### 3.4 The Clarkson II dataset

Dataset Description

This dataset was expanded by collecting not only keystroke data but also mouse movements, mouse clicks, and data on the software background and foreground processes. The data shows no limitations on human-computer interactions. Also, some piece of information includes "noise" because participants played computer games.

Acquisition Process

The data was collected under uncontrolled conditions, so that the data was more varied, with participants using different types of keyboards, but the data about keyboard type is not provided. Since the data was gathered in a natural environment there is not a free text, but just a natural working process.

Data Characteristics

Data Characteristics is showed in the table 4.

Table 4. Data Characteristics of Clarkson II dataset.

| Characteristics | Description | Example Entry |
|---|---|---|
| User ID | Unique identifier for each participant | 4302075 |
| Time Stamp (ticks) | Time of the event recorded in ticks (100ns per tick) | 636172286538589004 (2016-12-13 12:24:13) |
| Action Type | Describes action 'KeyDown' (0) and 'KeyUp' (1) | 0 |
| Key Name | Name of the pressed key | A |

Details about other collected data

The data was collected in completely uncontrolled conditions, so the data was more diverse as participants used different types of keyboards, but the data about keyboard type was not provided.

Since the data was gathered in a natural environment, it reflects authentic working processes rather than structured free-text tasks.

Applications and Insights

The distinctive feature of this dataset is that it is closer to the real user workflow. Besides Keystrokes, there are also mouse clicks and data on the software background and foreground processes, all this information can be mapped to a specific task performed by the user, which also gives a particular context.

### 3.5 Comparative Analysis

The comparative analysis of data sets reveals many subtleties of data collection methods, their characteristics, what data is best to collect, and what aspects to pay attention to. Each data set has strengths and weaknesses that differ from other data, features that guide the direction of research on emotional states.

Keystroke time data is recorded in the CMU dataset in a controlled laboratory environment. Because it offers uniform data collecting, it is perfect for long-term research on how typing habit changes over time. High data consistency is ensured by stringent guidelines to prevent data input errors during data collecting.

The Buffalo dataset comes one step closer to real-world typing conditions, including different keyboard types and tasks. It can help in analyzing typing behavior as it includes fixed and free text types. Demographic information could open up avenues for other research into human emotions and perhaps target a specific group of people.

The BB-MAS dataset features more comprehensive data collection because data was collected from computers, tablets, and phones, both in the laboratory and real environment. It models human behavior, as it includes data that a person performs physical activities, such as going downstairs and walking along corridors. This set makes it possible to assess human typing behavior under various physiological and contextual conditions.

Clarkson II is unique because it includes keystroke data, but also mouse movements and clicks, and software information. This set reflects the user's real work environment well, offering many options for analyzing user behavior in the work environment using the user's computer. The presence of "noise" from activities such as computer games also enhances realism. So "noise" can be useful for developing stress recognition models or any models where researchers need to take into account user behavior in real life.

Table 5 summarizes these datasets.

Table 5. Comparison of three datasets.

| Aspect | CMU Dataset | Buffalo Dataset | BB-MAS Dataset | Clarkson II Dataset |
|---|---|---|---|---|
| Number of Users | 47 | 148 | 117 | 103 |
| Acquisition Procedure | Detailed timing information of keystrokes | Laboratory sessions across three sessions | Keystroke, gait, and swipe data from multiple devices | Passive logging software |
| Acquisition Environment | Not specified | Laboratory setting | Laboratory and real-world settings | Uncontrolled, natural setting |
| Keyboard | Desktops | Desktops | Deskto | Desktops |

| Type | | | ps, phones , tablets | |
|---|---|---|---|---|
| Variety of Keyboards | Yes | Yes | No | Yes |
| Keyboard Information | Not provided | Provided | Provid ed | Not provided |
| Distribution of Samples | Primarily graduate students | Participant s from diverse backgroun ds | Partici pants from diverse backgr ounds | Not provided |
| Additional Information | Not provided | Gender | Age, gender, height, ethnicit y, langua ges, etc. | Mouse clicks, mouse movement s, and active programs |
| Key Name | Provided | Provided | Provid ed | Provided |
| Type of text | Fixed | Fixed and Free | Fixed and Free | Free |

## Ⅳ. Discussion

The comparative analysis of datasets is an important innovation regarding different approaches and how they work to identify emotions based on keyboard dynamics. All data sets are presented as separate advantages and disadvantages, which demonstrate the various methods and factors necessary for effective emotion recognition.

The CMU dataset is most suitable for long-term studies of the evolution of typing behavior because it is collected under regulated environment and provides consistent and reliable data. The lack of adaptability in everyday situations where users may encounter different emotional states and input settings limits their use.

In contrast, the Buffalo dataset provides a more realistic typing environment, including different types of keyboards and tasks, as well as demographic information. This diversity provides a more complete analysis of typing behavior in different contexts and user groups. But even working in a controlled laboratory imposes certain limitations that may not fully reflect the spontaneity of natural typing.

By including data from multiple devices and environments, BB-MAS datasets provide a more complete view and more accurately reflect real-world scenarios. The inclusion of physical activity adds another level of complexity, allowing the evaluation of typing behavior in various physiological conditions. This diversity makes the BBMAS datasets especially valuable for research aimed at understanding how external factors influence typing dynamics.

The Clarkson II dataset stands out because it contains information about mouse movements, clicks, and software collected in an uncontrolled natural environment. This dataset provides a context for analyzing user behavior in a real-world work environment and allows us to understand how various tasks and actions affect typing patterns. Realistic "noise" from playing games enhances realism and provides resources for developing models that explain everyday distractions and stressful situations.

The main focus of these 4 datasets is common biometric recognition. Research [8] has integrated biometric recognition and emotional state analysis, exploring how emotional factors influence typing patterns. Creating a complete dataset with information for both biometric and emotional recognition is the aim of this study.

## Ⅴ. Conclusion and Future Work

## Conclusion

In addition, the awareness of researchers on this issue is growing, and we hope that more databases will be publicly available. But this can lead to another threat, which is that many experiments are conducted with different databases. In addition, all databases are suitable for different purposes. The identity database is not optimal for verification algorithms. Just as the fixed-text algorithm are incompatible with free-text databases. The study underscored the necessity for developing new datasets that capture emotional states through keystroke dynamics, as existing datasets are insufficient for this purpose. Taking into account current research and comparing four existing datasets, key features necessary for typing patterns to accurately reflect emotional states have been identified. These results highlight how important it is to have a rich and relevant dataset to properly advance emotion recognition in artificial intelligence systems.

## Future Work

Based on the results of this study, future research may be directed in several promising directions. Firstly, it is needed to broaden participant diversity. To enhance the generalizability of findings and capture a broader range of typing behaviors and emotional responses, future research should prioritize the involvement of participants from diverse departments and backgrounds. Furthermore, future analysis can explore how an individual's mental state may change, how much they are different from the normal one during the day and across different days of the week (e.g., from Monday to Friday) or what kind of task takes more energy. Understanding these fluctuations in user conditions can provide valuable insights into productivity patterns and psychological health. Additionally, user information might have a big impact on typing behavior and emotional reactions, so they also should be included. These characteristics include age, gender, cultural background, computer ability, and habits like left-handedness. Neglecting these characteristics may result in incorrect categorization.

## REFERENCES

[1] C. Epp, M. Lippold, Mandryk, “Identifying Emotional States Using Keystroke Dynamics”, *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (CHI 2011)*, pp. 715-724, 2011.

[2] Amith K. Belman, Li Wang, Sundaraja S. Iyengar, Pawel Sniatala, Robert Wright, Robert Dora, Jacob Baldwin, Zhanpeng Jin, Vir V. Phoha, November 20, 2019, "SU-AIS BB-MAS (Syracuse University and Assured Information Security - Behavioral Biometrics Multi-device and multi-Activity data from Same users) Dataset ", *IEEE Dataport,* doi: https://dx.doi.org/10.21227/rpaz-0h66.

[3] Sun, Y.; Hayreddin, C.; Shambhu, U. “Shared Keystroke Dataset for Continuous Authentication,” *In 2016 IEEE International Workshop on Information Forensics and Security (WIFS); IEEE: Abu Dhabi, United Arab Emirates,* 2016.

[4] Christopher Murphy, Jiaju Huang, Daqing Hou, Stephanie Schuckers. "Shared Dataset on Natural Human-Computer Interaction to Support Continuous Authentication

Research," *IJCB 2017*, Colorado Denver, 2018.
[5] K. S. Killourhy and R. A. Maxion, “Comparing anomaly−detection algorithms for keystroke dynamics,” in 2009 *IEEE/IFIP International Conference on Dependable Systems & Networks,* pp. 125–134, 2009.
[6] Maalej, Aicha, and Ilhem Kallel. "Does keystroke dynamics tell us about emotions? A systematic literature review and dataset construction," *2020 16th International Conference on Intelligent Environments (IE). IEEE*, Aug., 2020.
[7] A. Kolakowska, “Recognizing emotions on the basis of keystroke dynamics,” in Proc. 8th Int. Conf. Hum. Syst. Interact. (HSI), Jun., pp. 291–297, 2015.
[8] Kołakowska, A. and Landowska, A., 2021. “Keystroke Dynamics Patterns While Writing Positive and Negative Opinions,” *Sensors,* vol. 21, no. 17, pp. 5963, 2021.

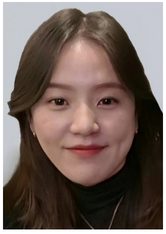
Authors

Kolmogortseva Karina

She received B.S. degree from the Department of Information science and Computer Engineering, TUSUR University, Russia, in 2020. She is currently pursuing the M.S. degree with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. She joined the Pattern Recognition Laboratory, in September 2023. Her research interests include emotion recognition using keystroke dynamics, biometric authentication, and the analysis of behavioral patterns.

Soo−Hyung Kim (Member, IEEE)

He received the B.S. degree in computer engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, in 1988 and 1993, respectively. Since 1997, he has been a Professor with the School of Artificial Intelligence, Chonnam National University, South Korea. His research interests include video understanding, multi−modal emotion recognition, medical image analysis, and document image processing.



Aera Kim

She received the B.E. degree from the Department of Biotechnology at Chonnam National University in 2013. Currently, she is working in the AI Convergence Department as a research staff.