Keystroke Analysis for User Authentication

Khushi R. Kher, DoCSE SVNIT

Abstract

The utilization of keystroke dynamics for user authentication has been an area of substantial interest and development in the field of biometric security. This report offers an examination of keystroke analysis as a method for authenticating users, employing machine learning techniques. The report encompasses a comprehensive exploration of the theoretical underpinnings, and contemporary research in keystroke dynamics. Furthermore, it provides insights into the practical implementation of keystroke analysis for user authentication, elucidating the operational aspects and technical intricacies involved. Additionally, the report critically evaluates the limitations encountered within this authentication method, providing a detailed analysis of the challenges faced. The report concludes by outlining the potential of keystroke analysis in enhancing security measures and augmenting user experience. Overall, this seminar report aims to contribute to the discourse on keystroke dynamics, shedding light on both its advancements and limitations while envisioning its future prospects in the realm of user authentication.

1 Keystroke Exploration

As we venture deeper into the realm of user authentication, a remarkable subset of biometric authentication comes into focus: keystroke dynamics. This innovative approach harnesses the unique typing patterns of individuals to verify their identity. Keystroke analysis is a branch of biometric authentication that stands at the intersection of security and user convenience, offering a novel solution that capitalizes on the distinct rhythm and cadence of how individuals interact with their keyboards.

1.1 The Uniqueness of Typing Patterns

Keystroke dynamics explores the nuances of how individuals interact with their keyboards. It meticulously records various parameters such as the time intervals between keystrokes, the duration of key depressions (known as dwell times), and the delays between key presses (commonly referred to as flight times). This unique approach to biometric identification leverages the individuality of typing patterns, which are as distinct as a fingerprint or a retina pattern. Unlike conventional security measures that rely on what is typed, keystroke dynamics emphasizes how one types. This distinct characteristic renders it exceptionally robust and notoriously challenging to replicate. [6]

1. Compared to Traditional Biometrics: While fingerprints and retinal patterns are well-known biometric markers, keystroke dynamics offer a distinctive advantage. Unlike static biometrics, which remain constant throughout a person's life, typing patterns can evolve over time. Factors such as fatigue, mood, and even physical health can influence typing behavior. This adaptability makes keystroke dynamics a versatile biometric modality.
2. Robustness and Challenging Replication: The intricacies of typing patterns present a formidable challenge to would-be imposters. Attempting to replicate the exact timing and rhythm of an individual's typing is a complex feat. Furthermore, keystroke dynamics transcends language barriers and accommodates variations in typing speed, making it a truly universal method of identification.

1.2 Relevance in User Authentication

The relevance of keystroke analysis is underscored by its applicability in various domains, where security and user convenience are paramount [1]:

1. Enhanced Security: Keystroke analysis bolsters security by adding an extra layer of identity verification. Even if an attacker has stolen a password, they would need to mimic the user's typing rhythm to gain access, making unauthorized entry significantly more challenging.

2. User Convenience: Unlike traditional authentication methods that often require users to remember complex passwords or carry physical tokens, keystroke analysis seamlessly integrates into the user's natural behavior. Users need not remember additional credentials, enhancing the user experience.

3. Continuous Authentication: Keystroke analysis enables the concept of continuous user authentication. While traditional methods only authenticate users during login, keystroke analysis can continuously verify the user's identity throughout their session, offering continuous protection.

4. User assistance: In cases where users forget their passwords or experience difficulties logging in, keystroke analysis can serve as a fallback method. By recognizing their typing patterns, users can regain access without needing to remember their forgotten credentials

1.3 Advantages of Keystroke Analysis

Keystroke analysis offers several advantages in the realm of user authentication [3]:

1. Enhanced Security: Keystroke analysis is challenging to spoof, making it a robust security measure.

2. User-Friendly: Users appreciate the convenience of a system that adapts to their natural behavior.

3. Low Intrusiveness: Keystroke analysis can operate in the background without disrupting the user experience.

4. Versatile Applications: This technology can be applied in a range of contexts, from access control to online transactions.

1.4 Identification and Verification

In the realm of keystroke dynamics systems, two distinct operational modes exist, Identification mode and Verification mode. Identification mode involves the endeavor to ascertain an individual's identity by scrutinizing a biometric pattern derived from their keystroke dynamics. This process necessitates the collection of a substantial volume of keystroke data, which is then used to identify the computer user based on previously gathered information from keystroke dynamics profiles of all users. During this training phase, a biometric template is computed for each user [6].

When a pattern is presented for identification, it undergoes a meticulous comparison against every known template, resulting in a score or distance indicating the similarity between the pattern and the template. The system attributes the pattern to the individual with the most akin biometric template. To safeguard against impostor patterns (i.e., those from individuals unknown to the system), the similarity must surpass a specified threshold. If this threshold is not met, the pattern is promptly rejected.

Notably, in identification with keystroke dynamics, the user must be identified solely based on their keystroke patterns, without any additional information. This reliance on inherent biometric features makes it a potential applications in user authentication and cybersecurity.

Conversely, in the verification case, the system verifies a person's identity against their individual template. Keystroke verification techniques can be categorized as static and dynamic or continuous.

Static verification methods analyze keystroke verification characteristics only at specific instances, offering heightened security compared to conventional username/password authentication, particularly during the user login sequence. In contrast, continuous verification monitors the user's typing behavior continuously throughout their interaction. This ongoing process involves the regular monitoring of the user while they type on the keyboard, enabling real-time analysis.

This dual approach to keystroke dynamics systems, encompassing both identification and verification modes, demonstrates the versatility and potential of biometric-based authentication methods in enhancing computer security.

1.5 Keystroke Analysis Methods

- Static at login: Static keystroke analysis authenticates a typing pattern based on a known keyword, phrase, or some other predetermined text. The captured typing pattern is then compared against a profile previously recorded during system enrolment. It supplements traditional username/password login by checking the timing of keystrokes for the username and/or password components [7].
- Periodic dynamics: Dynamic keystroke analysis authenticates a user based on their typing during a logged-in session. The captured session data is compared to an archived user profile to determine deviations. In a periodic configuration, the authentication judgment can be intermittent, either as part of a timed supervision or in response to a suspicious event. This method doesn't depend on specific text entries and can authenticate based on any input [7].
- Continuous dynamics: Continuous keystroke analysis extends data capturing to the entire duration of the logged-in session. This offers significantly more data for the authentication judgment. An impostor may be detected earlier in the session compared to periodic monitoring. However, this method requires additional processing, increasing computational overhead [7].
- Keyword-specific: Keyword-specific keystroke analysis extends continuous or periodic monitoring to consider metrics related to specific keywords. This could be an extra measure incorporated into a monitoring system to detect potential misuse of sensitive commands. For example, monitoring the keystroke metrics of a user attempting to execute critical commands [7].
- Application-specific: Application-specific keystroke analysis further extends continuous or periodic monitoring. It allows for the development of separate keystroke profiles for distinct applications. For example, profiling a user separately for their word processing application and their email client [7].

1.6 Measures of effectiveness

Similar to other biometric systems, the effectiveness of keystroke analysis is assessed based on two key metrics: the False Acceptance Rate (FAR) and the False Rejection Rate (FRR). The FAR pertains to instances where impostors are incorrectly identified as legitimate users. Conversely, the FRR refers to cases where the system mistakenly labels a legitimate user as an impostor. These rates are mutually exclusive, meaning that achieving optimal levels for both measures simultaneously is not feasible. Striving for an "equal error" scenario is not a practical compromise. Therefore, a decision must be made regarding which rate should take precedence, and this choice will vary based on whether static or dynamic authentication is employed [2].

In a static authentication scenario, the primary focus is on minimizing the FAR. This is because a successful impostor could potentially gain unchecked access for an entire session, posing a significant security risk. On the other hand, in a dynamic authentication scenario, there is a broader window for detecting impostors. Consequently, the priority shifts towards minimizing the FRR. Rejecting legitimate users during a session could have more significant consequences than occasional false login failures. Another crucial consideration in dynamic scenarios is the speed at which the system can provide an accurate identity assessment.

2 Deep Dive

From each keyboard keypress, we derived twelve features by combining the information from the keypress itself with the timing of the preceding and subsequent keypresses. For each couple of keys, the following data was collected [6]:

- Hold 1: The duration for which the first key is held down.
- Press-Press: The time interval between the pressing of the first key and the pressing of the second key.
- Release-Press: The time duration between the release of the first key and the pressing of the second key.
- Release-Release: The time elapsed between the release of the first key and the release of the second key.
- Hold-2: The duration for which the first key is held before the second key is operated.
- Total Time: The overall period between pressing the first key and releasing the second key.

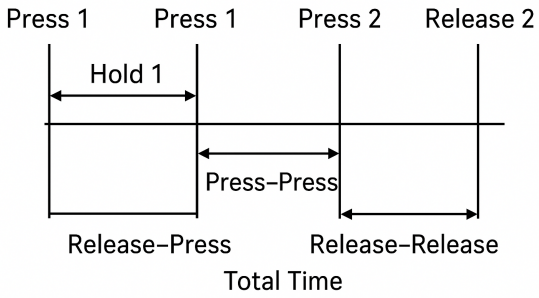

Figure 1: Extracted Data

The analyzed metrics were gathered based on the concept of a digraph, which represents the time interval between pressing two buttons. For each of these six features, we calculated and stored a slope, indicating the difference between the measurement of that feature in the current digraph and the subsequent digraph. Consequently, for each word, we can extract L * 12 features, where L represents the length of the word minus one [6].

This implementation utilizes the "Keystroke Dynamics Challenge 1" dataset available on Kaggle [5]. The dataset records the typing behavior of 110 users. Each user made eight attempts to input their password, and the dataset includes the captured timestamps of their keystrokes.

2.1 Verification Process

The verification process aims to confirm if the provided user identity aligns with the claimed one. This verification workflow involves the following steps:

- The user is prompted to enter the username and password previously chosen during enrollment.
- If the submitted (username, password) combination exists within our dataset, the user is requested to re-enter the password for confirmation.
- The keystroke pattern observed during the second password entry is compared against the patterns recorded during the enrollment phase.
- A similarity score is computed and returned using one of the aforementioned methods.

Additionally, every test entry is stored. Users are encouraged to attempt imitating impostor behavior by inputting (username, password) pairs that don't correspond to their actual credentials. This evaluation process assists in establishing an appropriate threshold for the model to make a conclusive decision.

2.2 Analysis

The L2 norm (Euclidean norm) is a mathematical metric commonly employed in machine learning to assess the magnitude or distance between vectors in multi-dimensional space. Keystroke-based authentication evaluates the similarity between a new password entry and a user's registered typing patterns by computing the Euclidean distance between their feature vectors. A reduced L2 distance signifies greater similarity, demonstrating consistent typing patterns. This norm accurately reflects nuanced differences in keystroke dynamics and is preferred for its discriminative power, resilience, and compatibility with machine learning systems, however it may be susceptible to outliers and encounter difficulties in high-dimensional environments [8].

The Naive Bayes algorithm is a probabilistic classifier derived from Bayes' theorem, which computes the probability of a data item belonging to a class based on its attributes. Assuming feature independence, it calculates probabilities efficiently, rendering it ideal for high-dimensional data such as keystroke dynamics. This study employed Gaussian Naive Bayes (GNB), wherein each feature is represented by a normal distribution characterized by its mean and variance. The method derives these parameters from training data and forecasts the most likely class for fresh inputs based on the calculated likelihoods.

Naive Bayes provides benefits like simplicity, rapidity, and proficient management of small or high-dimensional datasets, rendering it suitable for real-time user authentication. Nevertheless, its fundamental drawback resides in the independence assumption, which may not be valid for coupled keystroke features, thereby diminishing accuracy. Nonetheless, both the L2 norm and Naive Bayes offer complementary advantages—distance-based and probabilistic viewpoints—that enhance the efficacy of differentiating authentic users from impostors in keystroke-based authentication systems [4].

2.3 Evaluation

The Verification module assesses user authenticity by establishing optimal thresholds for two models based on False Acceptance Rate (FAR) and False Rejection Rate (FRR). FAR assesses unwanted access, and FRR evaluates the denial of legitimate users; attaining equilibrium between the two guarantees both security and usability. In the L2 Norm model, decreased distance scores beneath the threshold signify valid authentication, while Naive Bayes depends on elevated probability values. The Equal Error Rate (EER), at the intersection of False Acceptance Rate (FAR) and False Rejection Rate (FRR), underscores model efficacy, with the L2 Norm demonstrating enhanced precision. Furthermore, the Half Total Error Rate (HTER), which is the harmonic mean of the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), offers a singular metric of system reliability, thereby validating the efficacy of the L2 Norm in differentiating authentic users from impostors.

3 Applicability

While individual results based on a single insertion may not sufficiently guarantee user acceptance or rejection, the scenario would be different if users' keystroke patterns were detected and stored for every conceivable digraph. In such a case, a keyboard listener could quietly monitor a PC user's typing behavior and initiate a screen lock or request a password entry if these patterns consistently deviate from the user's profile for an extended period. This proposed feature has the potential to strengthen security while utilizing a minimally invasive method that capitalizes on this distinctive metric.

Moreover, a system akin to the suggested one, exploiting behavioral biometric traits, could function as a form of two-factor authentication. In this setup, users would be required to furnish both a behavioral characteristic, like typing a password, and a physical one, such as facial recognition or fingerprint authentication, to access a system or application. Overall, the integration of multiple biometric features in a recognition system can substantially enhance security and accuracy, making it considerably more challenging for potential attackers to breach security using only one feature.

4 Limitations

- Behavioral Variability: Typing patterns fluctuate as a result of stress, weariness, or environmental factors.
- Adaptability and Context: Various gadgets or keyboards modify typing behavior.
- Incomplete User Profile: Critical elements such as pressure and error rates are frequently overlooked.
- Impractical Large-Scale Implementation: Customized thresholds require substantial resources.
- Privacy Issues: Ongoing keystroke surveillance may appear intrusive.
- Optimal Threshold Dependency: The absence of user-specific thresholds diminishes accuracy.

5 Conclusion

Keystroke analysis as a user authentication approach signifies a substantial improvement in augmenting security and dependability within biometric systems. It is crucial to tackle issues like accuracy, vulnerability to impersonation, and adaptability for ongoing advancement, with joint initiatives among researchers, developers, and cybersecurity specialists propelling the enhancement of advanced algorithms and adaptive machine learning models. Equally significant is the focus on ethical practices, user awareness, and regulatory compliance to safeguard privacy, cultivate trust, and encourage the proper implementation of keystroke analysis in various applications.

5.1 Future Scope

The future of keystroke analysis is in its incorporation into multifactor authentication systems, especially in augmenting two-factor authentication. The integration of keystroke dynamics with additional biometric modalities, such facial recognition, fingerprint analysis, or iris scanning, enhances overall system security and mitigates risks linked to single-factor approaches. This synergy adds a layer of verification, reducing the risks of impersonation and unlawful access while delivering a more robust and user-friendly authentication solution. Current research aims to increase these integrations, and collaborative initiatives among biometrics, authentication technologies, and regulatory compliance are anticipated to position keystroke-enhanced multifactor authentication as a crucial component in the security framework of the digital era.

References

[1] M. Ahmed, A. N. Mahmood, and J. Hu. "Keystroke Dynamics for User Authentication". In: *Keystroke Dynamics for User Authentication*. 2012. URL: https://dl.acm.org/doi/10.1145/581271.581272.

[2] V. M. Baaijen, D. Galbraith, and K. d. Glopper. "Keystroke Analysis: Reflections on Procedures and Measures". In: *Journal Name* (2012). URL: https://journals.sagepub.com/doi/abs/10.1177/0741088312451108.

[3] A. Bierman and J. P. Degabriele. "Blind Password Registration and User Authentication". In: *Blind Password Registration and User Authentication*. 2009.

[4] S. M. Furnell et al. "Applications of Keystroke Analysis for Improved Login Security and Continuous User Authentication". In: *Book/Conference Title*. 2000. URL: https://link.springer.com/content/pdf/10.1007/978-1-5041-2919-0_25.pdf.

[5] *Keystroke Dynamics Challenge 1, Kaggle*. URL: https://www.kaggle.com/competitions/keystroke-dynamics-challenge-1/overview.

[6] M. Raffaele and T. Battistini. "KeyStroke Dynamics Authentication System". In: *KeyStroke Dynamics Authentication System* (2023).

[7] D. Shanmugapriya and G. Padmavathi. "A Survey of Biometric Keystroke Dynamics: Approaches, Security and Challenges". In: *arXiv preprint* (2009). URL: https://arxiv.org/ftp/arxiv/papers/0910/0910.0817.pdf.

[8] N. Sharma. *Importance of Distance Metrics in Machine Learning Modelling*. Accessed: 2025-10-24. 2022. URL: https://towardsdatascience.com/importance-of-distance-metrics-in-machine-learning-modelling-e51395ffe60d.