Investigating Differences in Gaze and Typing Behavior Across Writing Genres

Jun Wang, Eugene Y. Fu, Grace Ngai and Hong Va Leong

Department of Computing, Hong Kong Polytechnic University, Hong Kong

ARTICLE HISTORY
Compiled April 13, 2021

ABSTRACT
Writing is one of the most common activities undertaken on a computer, and the activity of writing has been widely studied. Given that writing is an intensively cognitive process, it makes sense that the type of writing that is being produced would have an affect on the writer's gaze and typing behaviors. However, only a few studies have explored this relationship. In this paper, we study the gaze-typing behaviors, specifically, the coordination between eye gaze and typing dynamics, of writers who are producing original articles in different genres: reminiscent, logical and creative. Our study focuses on Chinese typing, particularly via the Pinyin input method, which generates text via a two step method, and requires additional cognitive processes compared to typing in English. Our study involves 46 native Chinese speakers of varying ages from children to elderly. Our method deploys statistics- and sequence-based features to infer the mental state of the author during the writing process. The statistics-based features focus on modeling the overall gaze-typing behaviors during the process and the sequence-based features focus on the transition of the gaze-typing behaviors as the piece of writing progresses. Using a linear support-vector machine, we achieve an overall accuracy over 88% for the article-genre detection by using a leave-one-subject-out cross-validation evaluation.

KEYWORDS
Human-computer interaction; eye-gaze behavior; typing behavior; eye-hand coordination

1. Introduction

Writing tasks, such as writing instant messages, emails and other documents, form a large proportion of daily computer usage (Beauvisage, 2009). The activity of writing contains a cognitive and a generative (typing) process (Chukharev-Khudilaynen, 2014). During the cognitive process, a writer converts ideas to contextual sentences; during the typing process, a writer inputs sentences into the computer through the keyboard. The final output of the writing task is a piece of text, which can be classified into different genres such as correspondents (related to personal correspondence), technical writings, creative writings, and analytical writings (Gladis, 1993).

The cognitive process of writing has indeed been well studied in psychology. It contains three major parts: planning, translating and reviewing, and substantial interaction with the task environment, including rereading previously generated texts

CONTACT Eugene Y. Fu. Email: csyfu@comp.polyu.edu.hk

and typing on the keyboard (Flower & Hayes, 1981), and this interaction can be useful in inferring the cognitive process. For example, there has been work that investigates the possibility of determining the complexity of the writing task based on the writer's rereading behaviors (Torrance, Johansson, Johansson, & Wengelin, 2016; Van Waes, Leijten, & Quinlan, 2010) and the keyboard dynamics (Likens, Allen, & McNamara, 2017; Wallot & Grabowski, 2013). There are also some studies (Feit, Weir, & Oulasvirta, 2016; Johansson, Wengelin, Johansson, & Holmqvist, 2010; Papoutsaki, Gokaslan, Tompkin, He, & Huang, 2018) that use the gaze movement between the screen and the keyboard to predict the typer's typing skill level. Hernandez et al. (2014) detect mental stress based on the usage of mouse and keyboard. Huang et al. (2016) and Wang et al. (2019) determine whether a user is in high cognitive load based on the relative movement of gaze and mouse.

It is easy to imagine that writing different genres of texts involves different writing cognitive processes. For example, when a writer is journaling, the texts are often composed by narrative sentences recording memorable events; but a research paper usually contains sentences that are more formal and logical. However, to the best of our knowledge, there has not been a lot of work that explores whether, or how, these different cognitive processes affect the gaze-typing behaviors. In addition, most previous work in writing has been performed on writing in the English language. However, when the text concerned is to be generated on a computer, English is unique in the sense that there is a direct mapping between the typer's actions (e.g. the keys that are typed) and the desired output (e.g. the text that he/she wishes to generate). In other words, English texts can be directly input, letter for letter, on a standard keyboard. This is not true for languages such as Chinese, Kanji in Japanese, and Hindi, for example. In these languages, characters or words cannot be directly typed on a conventional keyboard. For example, typing in Chinese is a two-step process where an *approximation* of the target word or character is first *generated*, and a second *commit* step maps the typer's keystrokes to the final text. For example, in the Chinese pinyin input system, a user types in a phonetic approximation of the text. The computer then converts this to the actual Chinese characters. Since this mapping is often one-to-many, the user needs to choose the correctly generated text from a list. Even though language modelling algorithms are used to adaptively shuffle the most likely option to the top of the selection list, it is still reasonable to expect that the cognitive process of typing in Chinese would differ from typing in English. In this paper, we study writers' gaze and typing behaviors when they are generating texts on the computer using the Chinese Pinyin input method.

This study aims to investigate whether and how writing genres affect writers' writing behaviors when they are composing articles in Chinese. Two datasets are constructed. We collect data from 23 touch typists and 23 non-touch typists, who are native Chinese speakers, to investigate behaviors of writers with different levels of typing skill. Moreover, since we want our findings to be generalizable to every user, subjects recruited in this study are in multiple age groups, including children (18), college students (10) and elderly (18). In the experiment, subjects are required to compose articles in three writing genres respectively: reminiscent, logical, and creative. We collect subjects' eye gaze movements, mouse dynamics, keypress activities, and screen recordings in the experiments. Based on these signals, we extract features for writing genre detection.

Our features, which we call *gaze-typing features*, can be roughly divided into two types: statistics- and sequence-based. Statistics-based gaze-typing features are cross-modal features that combine the eye gaze location and the keyboard event data to capture both temporal and spatial information, which mainly describe the macro be-

haviors of a subject during the writing activity. Sequence-based gaze-typing features, on the other hand, capture the change in subject behaviors as the writing activity progresses. We then train a classifier to distinguish the writing genres automatically using only the eye gaze information and keyboard event information. Our final model is able to achieve an overall performance of 88.4% correctness when evaluated using a leave-one-subject-out cross-validation.

The contributions of this paper are therefore: (1) We identify a set of statistical- and sequence-based gaze-typing features, which appear to be indicative of different genres of writing (reminiscent, logical, creative) in Chinese generated on a computer using the Pinyin input method; (2) we explore how different cognitive processes of writing in different genres may affect the gaze-typing behaviors of subjects for both touch typists and non-touch typists from different age groups, including children, college students and elders; (3) based on these features, we develop an automatic classifier that is able to determine the writing genres that the writer is currently engaged in; and (4) demonstrate its effectiveness via experiments with human subjects with promising results. we believe our study would benefit future human-computer interaction studies, particularly, behavior-based user cognitive and mental state detection.

2. Related Work

Eye gaze and hand behaviors are important for daily human-computer interaction. We use our eyes to acquire content from the screen and the content is processed in our mind, which may engender a series of new instructions that are given to the computer through our hands via the mouse and the keyboard. As the *input* and *output* of the brain, eye gaze and hand behaviors may lend some insight into the mental state of the person.

There has been some previous work along these lines of utilizing eye gaze and hand behaviors to infer mental state of users. Bieget et al. (2010) found two fundamental gaze and cursor coordination strategies in search and selection tasks: (1) the user tends to move the mouse directly toward a target, when its approximate location is known; and (2) the user parallelizes searching and moving the mouse at a low speed otherwise. Rodden et al. (2008) analyzed gaze and cursor coordination in web searching tasks. They discovered that in most cases, mouse movements on web searching result pages are terminated with a click on some target, except for the mouse movements that follow horizontal eye movements or which highlight some particular texts. These type of mouse movements indicate that the user is processing the content.

There are also some studies that explore how affective states, especially mental stress, affect users' gaze and hand behaviors. Hernandez et al. (2014) infer mental stress based on hand behaviors. Using a special pressure-sensitive keyboard and a capacitive mouse, they exposed subjects to stressful task environments and found that when a user is under stress, he/she tends to type with greater force and hold the mouse with greater contact area. Huang et al. (2016) and Wang et al. (2019) infer the mental state by using gaze-mouse coordination. Huang et al. extract features from the small time window around each mouse-click to describe the movement of gaze relative to the cursor position. They use recursive mental math calculation to induce high cognitive load in subjects and discovered that when subjects are in high cognitive load, their gaze moves away from the click location *before* the click happens. Wang et al. illustrate that when a subject is mentally stressed, his/her gaze and mouse tend to approach to/depart from each other with higher speed.

Writing on the computer is a complex task, which contains both cognitive and physical processes. Both eye gaze and hand movements are involved in the writing task. Butsch et al. (1932) contributed the first study to investigate the eye-hand behaviors of typewriting. They find that the gaze is always approximately 5–7 characters ahead of the hands. A similar phenomenon is also illustrated by Inhoff et al. (1997), which shows that eye gaze position is usually 3 character-spaces ahead of the actual character that is being typed. Logan et al. (1983) expanded the findings by determining three kinds of *span*, or attention of foci, in typing: stopping, eye-hand, and copying. The stopping span is for committing text and the eye-hand span is the temporal or pixel difference between the locations of the eye gaze and hand execution for activities such as mouse movements and keypresses. A special case of the eye-hand span when 40-odd characters were involved was also identified and named the copying span. However, all these findings are obtained from copy-typing tasks in which a subject simply copies words from a pre-prepared source. Compared with producing original texts on the computer, the copy-typing task omits the cognitive process of producing contextual sentences based on the writing goal, which would be expected to affect the gaze and hand behaviors.

Feit et al. (2016), Johansson et al. (2010) and Papoutsaki et al. (2018) take another step in investigating the differences of gaze and typing behaviors across touch typists and non-touch typists while producing their own texts. Feit et al. primarily focus on typing dynamics, Johansson et al. on rereading behaviors and Papoutsaki et al. on gaze movements. They show that compared with the touch typist, the non-touch typist uses fewer fingers to control the keyboard and types with significantly lower speed with a larger gaze movement along the y-axis, presumably when the person's gaze shifts from the screen down to the keyboard before pressing the key. The non-touch typist also rereads their own texts less frequently while writing.

Many other studies also use gaze and mouse behaviors to determine the complexity of writing tasks and the writing quality. Torrance et al. (2016) discover that while subjects are producing complex texts, they will spend more time rereading previously generated material and their fixation duration becomes longer for lexical processing. Waes et al. (2010) conduct an experimental writing task in which subjects are asked to correct an embedded error and also complete a sentence. As the task increases in complexity, subjects tend to first complete the sentence and then correct errors, even though sometimes they have already noticed the presence of the error. The cognitive load of subjects also increases and they fixate less on the partial sentence while reading. Likens et al. (2017) use fractal analysis to model the inter-keystroke intervals as a time series. Their findings suggest that writing pieces with higher quality are generated by typing processes with higher degree of auto-correlation in the inter-keystroke intervals.

Most previous studies investigating writer's typing behaviors have been done in the context of English typing and relatively little effort has been paid to non-English typing. Zheng et al. (2011) collected over 54 million error-correction operations in Chinese typing with Pinyin input method. and discovered that the errors caused by omitting some letters are always (around 50%) corrected by deletions (re-typing). Common errors include transposition errors caused by messing the typing order of the left and right hands, and substitution errors caused by mistyping phonic representations which are similar to and close to the correct ones on the keyboard, such as "m vs.n", and "z vs. c vs. s". Meena et al. (2016) and Joshi et al. (2004) focused on Hindi typing. They found that the large number of letters, complex characters in Hindi language, and special structure of Indic scripts increase the difficulty of typing Hindi on QWERTY keyboards. Users thus need much more training to type Hindi. Samura et al. (2009)

explored keyboard dynamics of typing free texts in Japanese. Their results suggested that keypress duration is an important feature for individual identification. In our study focuses on gaze-typing behaviors in Chinese typing, through which we target to determine the genre of the article which is being written. To our best knowledge, this is the first time that this problem has been investigated.

3. Identifying the Thinking Phases Through Gaze-typing Dynamics

According to psychology and linguistics, the cognitive process of writing consists of different thinking phases: planning, translating and reviewing (Flower & Hayes, 1981). The planning phase involves retrieving related information from long term memory and creative thinking; the translating phase converts ideas into language according to contextual logic, which mainly involves short-term memory; and the reviewing phase consists of two processes: evaluating and revising, which may lead to a new cycle of planning and translating.

Following this logic, we attempt to identify three temporal *windows* in gaze-typing behaviors when subjects are writing articles on the computer in Chinese using the Pinyin input method. The *thinking window* is a continuous period of time during which the subject does not type on the keyboard. The *typing window* is a period of time during which the subject is formulating text, and entering it on the keyboard. Intuitively, we believe that the thinking window likely consists of the planning and reviewing phases of cognitive activities, and the typing window consists mainly of the translating phase.

Previous work (M. X. Huang et al., 2016; Rodden et al., 2008) has shown that some elements of human cognition will impact the coordination between gaze location and hand (or mouse) activities. To capture this coordination, we define a third type of window: the *transition window*. This window is a short period of time between a thinking window and a typing window. We further define Type 1 as a transition from a typing window to a thinking window and Type 2 as a transition from a thinking window to a typing window.

When a writer inputs Chinese text using the Pinyin input method, a list of potential corresponding Chinese characters or phrases are generated. This list is presented to the writer in a pop-up box with potential candidate phrases that appears just below the caret (Figure 1). The writer either uses the number key to select a particular option or hits *enter* to select the first candidate phrase. The selection activity causes the candidate box to disappear and the selected Chinese character(s) to appear on the screen at the caret location. The caret then shifts to the end of the character(s) that were just generated.

We use the appearance of this candidates box to identify the time windows from the data (Figure 2). The typing window is triggered when the writer inputs one keystroke, thereby commencing the potential generation of text. The typing window continues as long as the candidates box is visible on the screen (i.e. while the writer has not yet selected the final text). The gap between two typing windows is then considered to be a thinking window. If a thinking window is shorter than 750 ms, which has been found to be the minimum time required to interpret 5 characters (Rayner, Smith, Malcolm, & Henderson, 2009), we merge the window, along with its two neighbouring typing windows, together into one continuous typing window. To validate the appropriateness of the 750 ms minimum length for a thinking window, we observe the behavior of subjects inside thinking windows with duration shorter than 750 ms. We find that

Figure 1.: (a) An example of the pop-up candidates box. The Latin text (pin'yin'shu'ru'fa) is the actual input typed by the writer. The Chinese characters below are the possible mappings to the actual text, identified by the system. Five potential mappings are identified. The writer can type the number (1-5) to select the correct text, or hit "enter" to select the most highly likely option (1, also in red). (b) After the selection, the candidate box disappears and the selected Chinese character(s) appears on the screen at the caret location. The caret then shifts to the end of the character(s) that were just generated)

across all subjects who participate in the experiment, which will be described in detail in Section 6, the average number of fixations inside these windows is 0.93, and only around 10% of the fixations are focused on the screen. Around 70% of the time, the subjects *glance* at the screen, and the gaze stays in the same place for a duration shorter than the minimum duration of the fixation, which we set at 170 ms. One possible purpose of the glance is to confirm that the chosen characters have indeed been generated and appended to the previous text. Around 20% of the time, the subjects do not even look at the screen. We therefore consider that a thinking window that is shorter than 750 ms can be considered to be a part of the previous typing window.

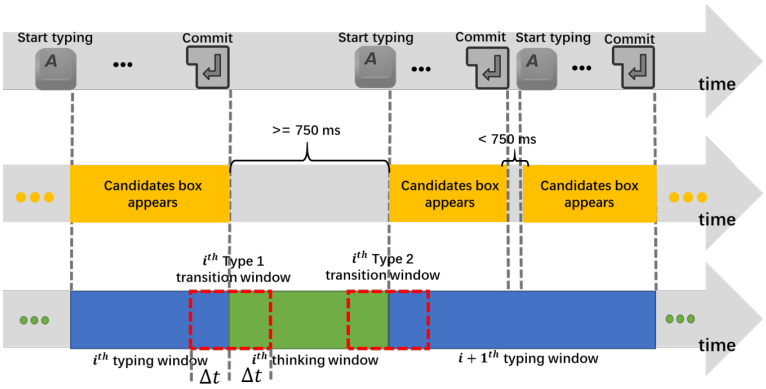

Figure 2.: The three types of time window and their correspondence with the appearance of the pop-up candidates box. Two adjacent typing windows are merged if the time gap is less than 750 ms. The duration for each transition window is $2\Delta t$, where Δt is 250 ms.

After identifying typing windows and thinking windows, we define transition windows (Tr) to be a 500 ms period spanning a thinking window and the adjacent typing window. The 500 ms parameter was chosen as it has been shown that the gaze usually starts to move 500 ms before the mouse moves, and the gaze always leads the mouse (J. Huang, White, & Buscher, 2012).

Fig 2 illustrates an example showing the three types of the time window and their relation to the writing activity. A whole writing process, thereby, can be considered as

a sequence of transitions between typing windows and thinking windows. We believe that different cognitive activities are involved in the thinking and typing windows, therefore, different behavior patterns should manifest in different types of windows.

Table 1.: Different types of thinking window and typing window based on gaze and typing activity

Window	**Type**	**Description**
Thinking window	Off-screen (O)	The subject looks away from the screen for the majority (>=50%) period of the window
	Reading (R)	The subject rereads the texts ahead of the caret
	Fixating (F)	The subject fixates at a place on the screen
Typing window	Less-press (L)	The subject presses fewer keys during the period of the window
	Pressing with uniform keypress intervals (U)	The subject presses several keys and time intervals between every two keypresses are similar in length
	Pressing with non-uniform keypress intervals (N)	The subject presses several keys and there exists at least one time interval between two keys, whose length is significantly greater than the others

3.1. Types of Thinking Window

A thinking window is a period of time between two typing windows when there is no typing activity. It has two main functions: 1) to review the texts that were just generated and 2) to think about what to write next. We expect that these two functions will generate different behavior patterns. For example, if a subject is in the reviewing phase, there is a higher probability that he/she may be rereading the already-generated texts, with more scanning behavior, and if a subject is in the planning phase recalling some writing material, we expect fixations with longer duration. Therefore, to better capture the changing of the cognitive activities, we differentiate the thinking window into 3 types: off-screen (O), reading (R) and fixating (F), based on the gaze behavior patterns.

A thinking window is determined as Type O, if a subject does not look at the screen for more than 50% of the time window. This thinking window appears more frequently when the subject is a non-touch typist. Two possible scenarios during which Type O thinking window may occur are when a subject is conceiving what to write next, or when a subject is recalling material. Alamargot, Chesnet, Dansac, and Ros (2006) show that when a subject is composing a text, long pauses are observed when he/she is contemplating "what to write next", or when he/she is considering the best way to express ideas. During this period, attention may not necessarily stay focused on the writing environment, which is referred as "averting the gaze". Therefore, Type O thinking windows exist in both the planning and translation phases.

Type F thinking windows are similar to Type O windows. During the window period, a subject focuses on the writing environment for a long period of time (long fixations), but without rereading the already-generated texts (lack of reading saccades).

A thinking window of Type R happens when a subject spends the majority ($\geq 50\%$) of the time rereading previously written texts. According to previous studies (Flower & Hayes, 1981; Klein, 1999), rereading often occurs when a subject externalizes his/her

ideas into text or reviews what he/she just writes. Thus Type R thinking windows appear in both the translation and reviewing phases.

3.2. Types of Typing Window

The main function of a typing window is to generate the actual text which was formulated in the last thinking window. Based on the typing behaviors, we define 3 different types of typing window: windows with lower keystroke frequency (L), windows with uniform keystroke intervals (U) and windows with non-uniform keystroke intervals (N). These windows capture different types of typing behavior patterns, which may reflect different mental states of a subject.

Type L typing windows are usually shorter in duration, as the keystroke frequency is lower, they contain fewer keypresses. We set the threshold to be not more than 4 keystrokes, including the final committing press that selects the character(s) to be generated. Considering that the average number of keypresses per typing window is 10.0, which is roughly equal to $3-5$ Chinese characters, these kinds of typing windows are fairly uncommon. The usual practice when typing in Chinese is to generate the approximation of a sequence of Chinese characters in the same candidates box before committing. As shown in Table 2, many phrases generated in Type L typing windows are functional phrases, especially auxiliary words, which are often used with a main verb to express tense, aspect, modality, voice, emphasis, etc. and may reflect the mental state of the subject.

Table 2.: Types of phrases generated in the type L typing window

Type	**Percentage**	**Type**	**Percentage**
Auxiliary word	23%	Link verb	3%
Preposition	7%	Pronoun	1%
Conjunction	7%	Other	54%
Adverb	4%		

Type N and Type U typing windows contain more than 4 keystrokes. The distinction between them is that Type N typing windows contain at least one interval between successive keypresses which lasts significantly – at least 3 standard deviations (over all keypress intervals of the subject) – longer than the others. This distinction attempts to capture pauses in writing, which indicate cognitive processing (Wallot & Grabowski, 2013). Table 1 lists all the types of thinking windows and typing windows with their descriptions.

4. Extracting Statistics-based Gaze-typing Features from Time Windows

We analyze statistics-based gaze-typing features both at window level and session level, and along the temporal and spatial dimensions. The process of extracting statistics-based gaze-typing features is shown in Figure 3. We define a session as the activity collected during the time it takes to compose a given article. The time windows are then extracted from the session using the appearance and disappearance of the candidates box as indicators. A session therefore consists of multiple thinking windows, typing windows and transition windows. We further differentiate the thinking window and typing window into different types based on the gaze and typing activities as

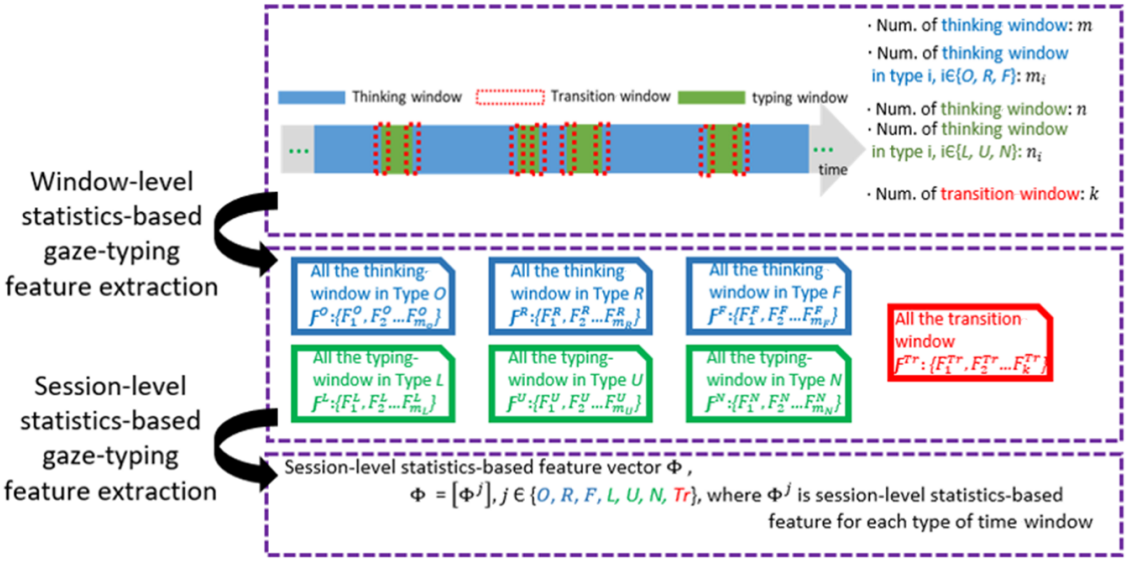

Figure 3.: Overview of Feature Extraction of Statistics-based Gaze-typing Features

shown in Table 1. For each type of thinking window, typing window and transition window, different sets of features are extracted to generate a window-level feature vector F_i^j, where $j \in \{O, R, F, L, U, N, Tr\}$ indicates the type of feature vector and i indicates that the feature vector is extracted from the i^{th} time window in type j of the session. The feature vectors of the same type are then aggregated to form the session-level feature vector $\phi^j, j \in \{O, R, F, L, U, N, Tr\}$. Appending the session-level feature vectors for different types of thinking window and typing window together gives us the final overall session-level statistics-based gaze-typing feature vector ϕ, where $\phi = [\phi^j]$.

4.1. Window-level Features for the Thinking Window

After defining three different types of thinking window based on the gaze behavior patterns during the window period, we can construct the window-level features to capture behavioral differences when generating articles in different genres.

Since there are no keyboard events during the thinking window, by definition, thinking window features are related to the eye gaze. In Type O thinking windows, we want to model behavior that characterizes a subject's formulating ideas for additional content while not focusing on the writing environment. However, since we cannot detect the gaze position reliably when the subject's gaze is off-screen, the only feature (f_1^O) that we can define is the duration while the subject's gaze is off the screen, as shown in Table 3. This feature gives us a sense of the length of the pause while the subject either recalls the material that will be generated next, or while he/she translates ideas into texts. We define a time period as being an off-screen gaze if 1) the eye tracker cannot capture any eye gaze inside the screen area and 2) the duration of the period is equals to or longer than 400 ms, which is the average duration of an eye blink (Schiffman, 1990).

In Type R thinking windows, a subject is mainly rereading previously-generated texts. We thus design the first part of the feature set (f_1^R) to describe the text that is being reread by the subject. If the location of the text that is being read is close to

Table 3.: F^O: Features describing the behavior in the type O thinking window

Feature	Meaning	Formulation
f_1^O	Gaze off-screen duration	Sum of the duration when gaze is off-screen

the caret, it is likely that this text was just generated in the previous typing windows and the subject is likely to be in a reviewing phase. However, if the location that is being read is 2 or 3 sentences away from the caret, then the subject may be translating his/her ideas into a sentence that integrates with the previous text.

Another feature (f_2^R) measures the amount of text reread by the subject. We define the distance between the reread texts and the caret as the number of pixels from the midpoint of the reread text to the position of the caret along the text line. Figure 4 illustrates an example. The green line shows the reread texts, and the distance to the caret is denoted by the red dash line. We also extract the features ($f_4^R - f_5^R$) to describe the fixation including the number of fixations and average duration of fixations. Table 4 describes this set of features.

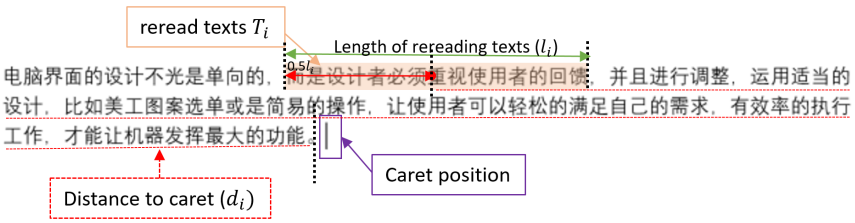

Figure 4.: Illustration of the features that describe the reread texts.

Table 4.: F^R: Features describing the behavior in the type R thinking window

Feature	Meaning	Formulation
f_1^R	Distance of the reread texts	Average distance between all the reread texts to the caret during the time window ($Mean(d_i)$)
f_2^R	Length of the reread texts	Total length of the reread texts ($Sum(l_i)$)
f_3^R	Rereading duration	Total duration spent in rereading already-generated texts
f_4^R	Number of fixations	Total number of fixations in the time window
f_5^R	Duration of fixations	Average duration of the fixations in the time window

For Type F thinking windows, we want to capture behavior patterns similar to Type O windows, but modelling the act in which a subject fixates on the screen without rereading previously generated texts. Besides feature (f_1^F), which measures the duration of the pause, we also extract features (f_2^F) to describe the location of the fixation relative to the caret, as shown in Figure 5, and features (f_4^F, f_5^F) that describe the fixation.

We observe from our data that there are time periods during which the writer seems to stare at a small area for an extended period of time. This behavior generates a lot of fixations within that small area. This also seems to be correlated with thinking

behavior on the part of the writer, as they do not seem to correspond to reading behavior. We therefore define these *stare points* (sp) as areas with radius of 50 pixels or less (two Chinese words take up 100 pixels) with several fixation points.

The distance between each stare point and the caret is measured from the center of the stare point to the center of the caret, and the duration of the i^{th} stare point (sp_i) t_i is defined as the total duration of all fixations in stare point sp_i. Table 5 shows all the features with meaning and formulation.

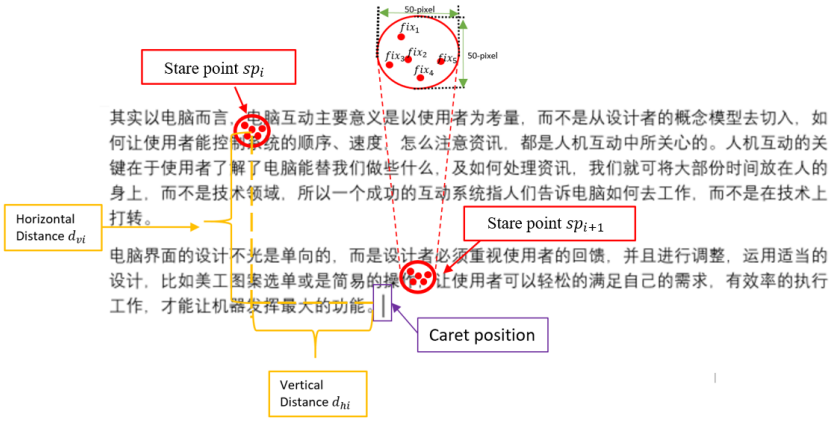

Figure 5.: Illustration of the features that describe the staring point.

Table 5.: F^F: Features describing the behavior in the type F thinking window

Feature	Meaning	Formulation
f_1^F	Horizontal distance to the caret	Average horizontal distance of fixations to the caret position ($Mean(d_{hi})$)
f_2^F	Vertical distance to the caret	Average vertical distance of fixations to the caret position ($Mean(d_{vi})$)
f_3^F	Total duration spent in staring and thinking	Total duration spent in staring at screen and thinking
f_4^F	Number of fixations	Total number of fixations in the time window
f_5^F	Duration of fixations	Average duration of the fixations in the time window

4.2. Window-level Features for the Typing Window

The definition of the typing window is a period of time during which a subject is typing on the keyboard, which we believe corresponds to the activity of translating ideas into language and input into computer. As previously described in Section 3, a typing window contains keystrokes, which are processed by the system through a series of pop-up candidates boxes. Similar to the thinking window, we define different types of typing window based on typing patterns and design different groups of features to model the behavior patterns.

Table 6 shows the features extracted from Type L typing windows. These windows contain few keyboard presses, which we observe usually correspond to the generation of functional characters or phrases. The language modelling inside the keypress-to-character conversion mapping sorts commonly-seen characters or phrases to the top,

Table 6.: F^L: Features describing the behavior in the type L typing window

Feature	Meaning	Formulation
f_1^L	Duration	Duration in which the pop up window is visible on the screen
f_2^L	Keypress interval	Average interval between every two keypresses

which means that the writer often only needs to type the first character instead of the whole phonetic mapping. For example, the phonetic mapping for "I" and "we" are *"wo"* (我) and *"wo men"*(我们) respectively. Since these words are so often used, the software will generate these words as soon as the writer types "w", without waiting for the following "o". These characters, because they are so commonly used, are usually generated proficiently and at a high speed. Features ($f_1^L - f_2^L$) are designed to capture the impact of the different cognitive activities on the generation of these common terms.

Type U and Type N typing windows contain more keypresses. This allows us to extract more complex features to model the behavior patterns. Wallot and Grabowski (2013) illustrate that, compared with simple typing, generating text creates more complex keystroke activity, which manifests in two ways: 1) longer keypress intervals, reflecting longer pauses in writing and 2) increased number of editions and deletions. Our features are designed to describe these two aspects of behavior, as described in Table 7 and Table 8.

Table 7.: F^U:Features describing the behavior in the type U typing window

Feature	Meaning	Formulation
f_1^U	Number of keystrokes	Total number of keystrokes during the window period
f_2^U	Keypress interval	Average interval between every two keypresses
f_3^U	Recurrence	Total number of deletes and edits performed during the window period
f_4^U	Duration	Duration in which the pop up window is visible on the screen

Table 8.: F^N: Features describing the behavior in the type N typing window

Feature	Meaning	Formulation
f_1^N	Number of keystrokes	Total number of keystrokes during the window period
f_2^N	Keypress interval	Average interval between every two keypresses
f_3^N	Pause duration	Total duration of intervals, in which the duration is 3-deviations away from the average
f_4^N	Recurrence	Total number of deletes and edits performed during the window period
f_5^N	Duration	Duration in which the pop up window is visible on the screen

4.3. Window-level Features for Transition Windows

By definition, the transition window is a short transition period between a thinking window and a typing window. During this time, a subject has either started to type on

the keyboard, and is thereby in the process of translating ideas into text, or has finished typing and has therefore entered the reviewing or thinking phase. M. X. Huang et al. (2016) has shown that gaze-hand patterns around the transition point are indicative of the cognitive state of the human being. We therefore follow their work in extracting similar features around the transition point.

One noticeable gaze-hand transition in this experiment is that subjects sometimes need to look at the keyboard to locate keys, resulting in much shifting of the gaze between screen, keyboard, and the candidate box. This allows us to extract the following information:

- When a subject starts typing (Type 2 transition window), we compute the time difference between the first keypress and the gaze starting to move downwards towards the keyboard. If the first keypress happens first, then the time difference is positive. If we do not observe this kind of behavior during the time period, then the feature is set to 0.
- When a subject finishes typing (Type 1 transition window), we compute the time difference between the last keypress (committing to a textual string from a list of candidates) and the gaze starting to move away from the candidates box area. Similarly, if we do not observe this kind of behavior during the time period, then the feature is set to 0.

According to this information, we extract two features as shown in Table 9.

Table 9.: F^{Tr}: Window-level features extracted from the transition window

Feature	**Meaning**	**Formulation**
f_1^{Tr}	Time taken in looking towards the keyboard	For type 2 transition windows: Time between first keypress and gaze moving toward the keyboard
f_2^{Tr}	Time taken in looking away from the candidates box	For type 1 transition windows: Time between the last keypress and gaze moving away from the candidates box area

4.4. Building Session-level Statistics-based Gaze-typing Features

Session-level statistics-based gaze-typing features are used to model the overall gaze-typing behaviors in a session, which is the activity collected during the entire time of composing a given article. We believe these statistical session-level features can represent the macro behavior of a subject. Therefore, we extract two types of session-level features based on statistics from the window-level features: the average behavior and the variation inside a session.

For example, a session consists of m thinking windows, which includes m_O Type O, m_R Type R and m_F Type F thinking windows, where $m = m_O + m_R + m_F$. There are also n typing windows, which includes n_L Type L, n_U Type U, and n_N Type N typing windows, $n = n_L + n_U + n_N$. k transition windows are also extracted from the session.

A window-level feature vector is extracted from each time window based on its type as introduced in Sections 4.1, 4.2 and 4.3. We construct ϕ^j, a session-level statistics-based feature vector of Type j, where $j \in \{O, R, F, L, U, N, Tr\}$, by computing the mean value and standard deviation for each feature from the window-level feature vector F^j across all the time windows in Type j during the session.

For instance, the session-level statistics-based feature vector ϕ^R would be calculated as $[\mu_1, \sigma_1, \ldots, \mu_5, \sigma_5]$, where μ_i and σ_i are the mean and standard deviation of the i^{th} feature in the window-level feature vector F^R across all Type R thinking windows and i $\in [0, 5]$. Session-level statistics-based feature vectors for other types of windows can be extracted in the same way. The final overall session-level statistics-based feature vector ϕ is built by concatenating all types of session-level feature vectors together.

5. Extracting Sequence-based Gaze-typing Features From Session

Our statistics-based features are used to model the gaze-typing behavior patterns inside each type of time window. In contrast, the sequence-based features are designed to model the change of a subject's behaviors across the session, which we hypothesize can distinguish between different writing genres.

To build the sequence-based features, we first construct the behavior-transition sequence for each session, which captures the whole of the behavior transition exhibited by a subject across an entire session. We then extract "indicative" subsequences, or patterns, from this behavior-transition sequence.

The details of the process are described in this section.

5.1. Modelling the Behavior Transition within a Session

As defined in Section 4, a session represents the activity during the entire process of composing an article, which can be represented as a sequence of transitions between thinking windows and typing windows. We also show how we categorize thinking windows and typing windows into 6 types, that are designed to capture distinctive behaviors. Following on this, we model the change in behavior of a writer during the whole process of writing an article through the transition over the different types of time windows within a session.

For instance, the i^{th} session ($Sess_i$) contains m thinking windows and n typing windows. Since thinking windows and typing windows appear alternately, thus $|m - n| = 1$ or $m = n$. Given this, we generate a session-level behavior sequence $s_i = \{state_i\}_{i=m+n}$, where $state_i \in \{O, R, F, L, U, N\}$ corresponds to the type of the i^{th} time window in $Sess_i$. The label of s_i is the genre of $Sess_i$, which can be "Reminiscent", "Logical" or "Creative".

Table 10 presents an overview of the behavior information, including average length of the behavior sequence, and the distribution of the various behavior types within the sequences, for each of the writing genres in our dataset.

Table 10.: Overview of behavior types for different genres of writing

	Length of behavior sequence	State ratio					
		O	R	F	L	U	N
Reminiscent	259	15.7%	11.4%	23.6%	13.8%	12.0%	23.4%
Logical	245	14.6%	12.4%	23.8%	13.5%	11.3%	24.4%
Creative	237	14.9%	11.8%	23.8%	13.2%	12.1%	24.2%

5.2. Extracting Indicative Patterns from the Behavior Sequences

A pattern is a subsequence of behaviors, which can be regarded as a series of actions. For example, a pattern "$F \Rightarrow U$" is commonly seen in our dataset, and it describes the situation whereby a subject stares at the screen for a while to think, followed by typing text on the keyboard with uniform keypress intervals. However, this pattern is so frequently seen that it appears in all behavior sequences across different genres of writing. In this sense, it is not indicative as its presence does not provide distinguishing information between the different genres of writing.

An "indicative" pattern is therefore a subsequence which occurs differently across behavior sequences from different genres of writing. In order to judge the degree of "indicativeness", we define a weighting scheme that assigns an appropriate weight to each pattern to imply the amount of genre information provided by that pattern. Inspired by the work from text categorization (Debole & Sebastiani, 2004; Lan, Tan, & Low, 2006; Ramos et al., 2003), our pattern weighting scheme comprises of 3 components: pattern frequency (pf), relevance frequency (rf) and trend distance weightings (td). The weighting (w) can be computed as:

$$w = pf \times rf \times td \tag{1}$$

5.2.1. Relevance Frequency

Relevance frequency (rf) was originally proposed by Lan et al. (2006) for text categorization. In traditional text categorization problem, the rf factor is a supervised term, weighted with their indicativeness, which can be roughly interpreted as their power of discriminating the documents into positive and negative categories.

We map the original problem to our sequence classification task by viewing patterns and behavior sequences to terms and documents. We map each of the genres *reminiscent*, *logical* and *logical* to *positive* and the two other genres to *negative* in turn. For example, our dataset contains articles composed in one of three genres: reminiscent, logical and creative. Therefore, to compute the rf value of pattern p for the reminiscent genre, then S_+ contains all the behavior sequences exhibited when the writer is composing articles in the reminiscent genre, and S_- contains all the behavior sequences appearing in the logical and creative genres. Given all behavior sequences with positive labels (S_+) and all sequences with negative labels (S_-), then the relevance frequency of a pattern p can be computed as:

$$rf(p, S_+, S_-) = log(2 + \frac{|\{s \in S_+ : p \in s\}|}{|\{s \in S_- : p \in s\}|}) \tag{2}$$

where $|\cdot|$ returns the number of elements in the set.

5.2.2. Pattern Frequency

The relevance frequency formula gives higher weights to patterns that occur very infrequently in one class and more frequently in the other class. However, there is a possibility that it will identify rare patterns, which occur only once or twice in the entire dataset. These patterns are not helpful for our purpose as they may not be generalizable.

We therefore include the pattern frequency factor to balance the indicativeness with

generalizability. Pattern frequency (pf) measures how frequently a pattern p occurs in a behavior sequence. Since the length of the sequence may vary from session to session, the pattern frequency is normalized by the length of the sequence. Given a behavior sequence s, the pattern frequency of a pattern p can be computed as:

$$pf(p, s) = log(n_{p,s}/len(s)) \tag{3}$$

where $len(\cdot)$ returns the length of the sequence and $n_{p,s}$ is the number of occurrences of pattern p in the behavior-transition sequence s.

5.2.3. Trend Distance Weighting

The process of writing an article is dynamic, and as such, the writing behavior may change during the writing process. For example, when a subject writes a reminiscent article, recall behavior may appear more frequently at the beginning than the end of the writing.

Figure 6 presents an example. We have behavior sequences s_1 and s_2, belonging to different genres of writing, both of which contain 15 occurrences of Patterns p_1 and p_2. On the surface, it appears that p_1 and p_2 are not particularly discriminative. However, when we consider the different stages of writing, it can be seen that p_1 and p_2 have very different appearance patterns – p_1 appears more frequently towards the beginning of s_1, and more frequently towards the end of s_2.

These kinds of difference cannot be readily captured by tf and rf factors. Hence, we need a new factor to capture this difference.

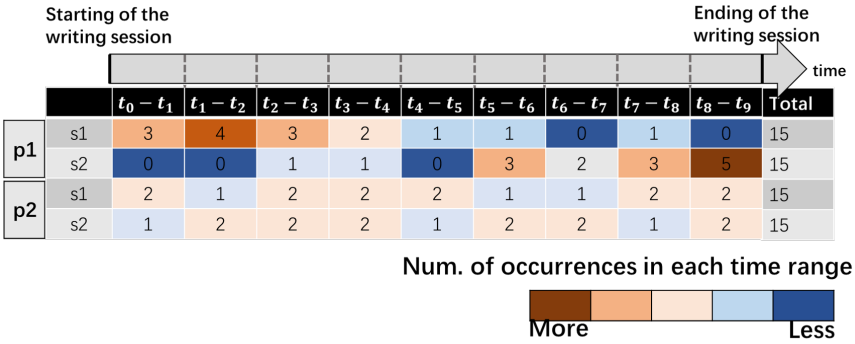

Figure 6.: Examples of two patterns which have same total occurrence times but have different trend distance weighting.

Based on this analysis, we propose a new weighting factor which we call the *trend distance weighting*, which takes into account the occurrences across the whole process of writing an article. We first assume that a writing process consists of Π stages. As a simplifying assumption, we also assume that the duration for all stages are equal. This allows us to partition the behavior sequence into Π behavior subsequences.

Figure 7 shows how we generate our behavior subsequences. We first divide the session into Π stages of equal duration (red dotted box). The behavior subsequence s^{par_i} is then simply the sequence of time window types of the windows that appear in the partition. In our example, the i^{th} stage consists of 5 time windows with types U,O,U,F,L. The behavior subsequence s^{par_i} is therefore $U => O => U => F => L$. Likewise, s^{par_i+1} is $O => U => O => N => R => N$.

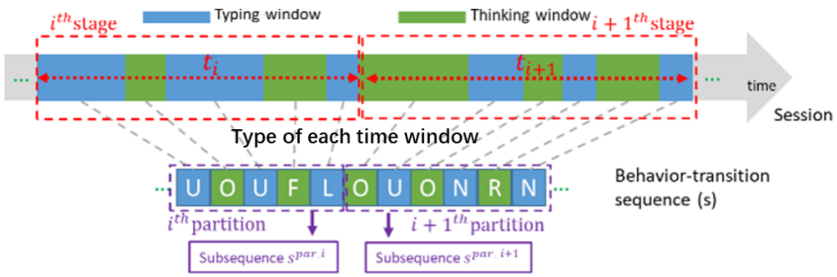

Figure 7.: Generating behavior subsequences from session data

To compute the trend distance weighting, first we count the number of occurrences of the pattern in each partition. Given a behavior-transition sequence s_i with Π partitions, the number of occurrences of the pattern p in each partition can be expressed as a vector $\mathbf{N}^p_{s_i}$ and $\mathbf{N}^p_{s_i} = [f_{p,s^{part\_1}}, f_{p,s^{part\_2}}, ..., f_{p,s^{part\_\Pi}}]$. Given all behavior-transition sequences with positive labels (S_+) and all sequences with negative labels (S_-), the trend distance weighting (td) of a pattern p can be computed as:

$$
\begin{aligned}
td(p, S_+, S_-) &= Euclidean(Q^p_{S_+}, Q^p_{S_-}), where \\
Q^p_{S_+} &= \frac{\sum_{i=1}^{|S_+|} \mathbf{N}^p_{s_i} / \left\|\mathbf{N}^p_{s_i}\right\|_1}{|S_+|}, s_i \in S_+ \\
Q^p_{S_-} &= \frac{\sum_{i=1}^{|S_-|} \mathbf{N}^p_{s_i} / \left\|\mathbf{N}^p_{s_i}\right\|_1}{|S_-|}, s_i \in S_-
\end{aligned}
\tag{4}
$$

5.3. Sequence-based Gaze-typing Features

We have so far defined a weighting scheme to select indicative patterns, which represents some gaze-typing behaviors that may potentially distinguish different writing activities based on the exhibited behaviors. When a subject writes an article in one of the reminiscent, logical, creative writing genres, he/she is more likely to show behaviors that are indicative of that genre. This means that extracted patterns that are indicative for a particular genre should occur more frequently in behavior sequences generated from writing sessions corresponding to the genre.

We use a bag-of-words model (Wang et al., 2019; Wang, Liu, She, Nahavandi, & Kouzani, 2013) to generate the sequence-based gaze-typing features from the behavior sequences. We select the k highest-weighted patterns as our indicative patterns, or *word*, and represent each behavior-transition sequence by the occurrence frequencies of the *words* contained, in a bag-of-words approach (Wang et al., 2019, 2013). If k patterns are selected and each behavior-transition sequence contains Π partitions, then the size of the sequence-based gaze-typing feature vector is $k \times \Pi$ and the value of the i^{th} entry is the number of occurrences of the $(\lfloor (i-1)/\Pi \rfloor + 1)^{th}$ pattern in the $(i - 3 \times \lfloor (i-1)/\Pi \rfloor)^{th}$ partition of the sequence.

6. Experiment

The aim of this study is to investigate gaze-typing behaviors across different genres of writing. We, therefore, construct our own datasets so that they satisfy three

requirements. First, subjects need to produce original articles of different genres. Second, the datasets need to include subjects from different age groups. Third, subjects need to type in a non-English language (we selected Chinese, using Pinyin as the input method). This section will introduce details of dataset construction, experiment design and data distribution.

6.1. Collecting Datasets

There are 46 subjects engaged in this study. 18 subjects belong to the child age group (age $8-12$, $M = 9.85$, $STD = 1.46$), 10 subjects to the college student age group (age $22-29$, $M = 24.6$, $STD = 2.46$) and 18 to the elder age group (age $55-67$, $M = 60.75$, $STD = 4.05$). All subjects were compensated for their time with a 200 RMB supermarket coupon, and parental consent was obtained for all the child subjects. A pre-experiment survey showed that all subjects were familiar with using computers and were at least able to type using two hands. All the subjects were native Chinese speakers and their normal mode of textual input uses the Chinese Pinyin input method described in Section 3.

Figure 8 shows the experiment environment. The experiment was carried out in a conventional office setting. The setup consisted of a 22" LCD monitor at 1600×1050 resolution with Microsoft Word running in full-screen mode. A Tobii EyeX eye tracker was attached to the bottom of the screen and a full-size QWERTY keyboard used for input. During the experiment, subjects were required to sit around $60cm$ away from the screen in a comfortable typing position. The subject's eye gaze location on the screen, as detected by the EyeX tracker, was logged at $60Hz$. The mouse cursor position was also captured at $100Hz$. All keypress events were also logged. Screen recordings were also taken at $30Hz$.

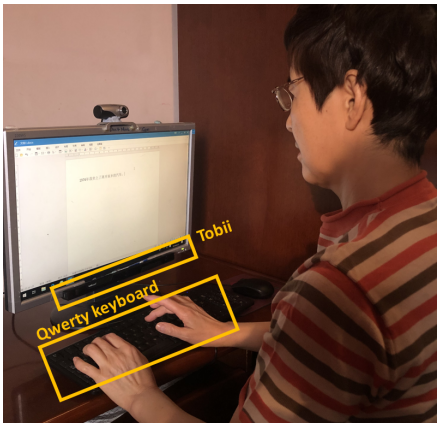

Figure 8.: Experiment environment

6.2. Experiment Design

We hypothesize that when people are in a writing task, their cognitive process and load correlate to the genre of their writing. For example, editing a scientific article and writing a narrative essay are completely different writing tasks, with different cognitive processes, that may manifest in different behaviors. Therefore, we asked our subjects to compose three articles as follows:

- **Reminiscent** — A memorable event that happened several years ago. The event should be described vividly and explicitly, with the objective of taking the reader back in time to experience the event.
- **Logical** — Write a set of instructions, using connective words (i.e. not in list form) to teach a new ability. Examples are playing bridge, or 2-digit number multiplication or division. The assumption is that the reader does not have pre-knowledge about this ability.
- **Creative** — Write an essay on a fantasy event, such as a day in the far future, or the life on a moon colony.

For each article, the subjects were instructed to write would be about one page in length, in around 30 minutes. If a subject did not finish within 30 minutes, he/she would be reminded of the time, but the experiment would continue till he/she finished writing the article. The order of the genres was randomized for each subject. The font size was set to 18 DenXian with triple-line spacing, so that the eye tracker could locate fixations and saccades accurately. Before starting writing the first article , every subject was given enough time to adapt to the keyboard and computer settings. Between every two tasks, there was a 15 minutes break to avoid fatigue. After each break, the eye tracker was recalibrated.

Experiment sessions in which the subjects wrote too little (i.e. they got stuck), or which otherwise did not meet our length requirement, were removed. In total, data collected from one elderly subject and two child subjects were removed due to the length and one elderly subject's data was removed due to the equipment issue (the eye tracker somehow failed to detect her eyes accurately).

6.3. Datasets

Our experiments resulted in 138 instances, each of them representing around 30 minutes of composing/typing activity from 46 subjects (18 in child age group, 10 in college student age group, 18 in elder age group). Among these instances, 46 instances were classified as reminiscent, 46 as logical and 46 as creative.

We note that one of the biggest impacts on the gaze-typing behavior comes from the typing skill. Touch typing is a style of typing in which the subject relies on muscle memory to locate keys. Non-touch typists need to look at the keyboard to locate the keys. Therefore, the gaze and typing data of the non-touch typists exhibit more dramatic displacements along the y-axis and lower typing speed. These differences in the gaze-typing behaviors are far more marked than the differences induced by the article genres.

We therefore separate the data based on typing skill for easier analysis. For each subject, we measure the time spent typing by the subject while looking at the keyboard by summing up all the typing windows during which the subject's eye gaze is away from the screen, and compute the ratio r of that time to the sum of all typing windows.

Figure 9 shows the cumulative distribution function of r for all the subjects. In this study, we choose $r = 0.5$ as the threshold to distinguish touch typists and non-touch typists. If $r >= 0.5$, implying that the subject needs to look at keyboard while typing more than half of the time, then the subject was classified as a non-touch typist. Otherwise, the subject was classified as a touch typist.

We finally determine 23 touch typist subjects and 23 non-touch-typing subjects based on the overall ratio r across three articles for each subject. Detailed composition of the datasets is shown in Table 11. As expected, most of the subjects in the child-

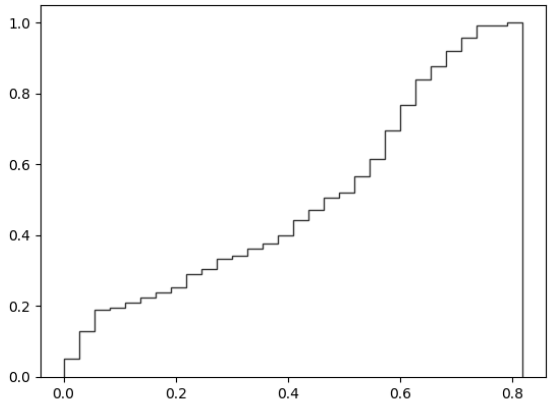

Figure 9.: Cumulative distribution function of r for all subjects

age group were non-touch typists, and most of the subjects in the college age group were classified as touch typists. Around 67% of subjects in the elderly-age group were classified as touch typists, as far as we knew, they worked with computers before they retired.

Table 11.: Detailed composition of the datasets from each age group

Dataset	Num. of subjects in Child-age group	Num. of subjects in College student group	Num. of subjects in Elderly-age group
Touch typist	1	10	12
Non-touch typist	17	0	6

Each data instance, which represents the typing activity over one article, contains mouse and eye gaze positions in the form of a series of $< t_{gaze}, x_{gaze}, y_{gaze} >$ tuples and keyboard events in the form of a series of $< t_{key}, key_{name} >$ tuples. We applied two-phase heuristic filters (Vargha & Delaney, 1998) to remove the impulse noise from the eye tracking data. The processed eye tracking positions was then passed through the Dispersion-Threshold identification algorithm (Salvucci & Goldberg, 2000) with the dispersion as 35 pixels and the minimum fixation time as 170 ms to identify eye fixations, which were presented in the form of a series of $< t_{fix}, d, x_{fix}, y_{fix} >$ tuples, where t_{fix} is the timestamp when the fixation starts, d is the duration of the fixation and x_{fix}, y_{fix} is the coordinate of the fixation position. A series of caret positions are also extracted in the form of $< t_{caret}, x_{caret}, y_{caret} >$, which means that at moment t_{caret}, the caret is located at $< x_{caret}, y_{caret} >$ on the screen. This pre-preprocessed data is used to model gaze-typing behaviors at a given moment t. At moment t, based on the keyboard events data, we can determine whether a subject is in a thinking or typing phase. Keyboard events can also be used to model the typing dynamics. Combining with fixation and caret positions allows us to deduce whether a subject is rereading the previous written texts, or just staring at a place and thinking.

6.4. Data Distribution

During the experiment, three articles in different genres were written by each subject. In order to not disturb the cognitive process of writing, we did not impose many detailed constraints, such as word choices, average sentence length or number of paragraphs. The subjects are also allowed to delete or edit previously-entered text during the writing. For better understanding of the data, this section presents some descriptive statistical analysis of the different writing behavior observed in our dataset.

Table 12.: Number of words among different writing genres

	Num. words in Reminiscent	Num. words in Logical	Num. of words in Creative	Num. of words in All genres
Touch typist	266.5	222.0	227.3	237.4
Non-touch typist	237.2	199.0	208.3	216.3

Table 13.: Typing speed in words per minute among different writing genres

	Typing speed (WPM) Reminiscent	Typing speed (WPM) Logical	Typing speed (WPM) Creative	Typing speed (WPM) All genres
Touch typist	46.9	44.4	42.2	44.5
Non-touch typist	16.2	15.5	16.7	16.1

Table 12 shows the length of articles in different genres generated by touch typist and non-touch typists. According to the table, touch typists tend to generate articles in longer length. For different genres of articles, reminiscent articles are the longest and creative articles are shortest. Table 13 illustrates the typing speed across different groups. Obviously, touch typists type much faster than non-touch typists. Figure 10 shows the vocabulary usage for both the touch typists and the non-touch typists. The majority of the vocabulary used for both groups belong to the top 1000 most-frequently used Chinese characters[1]. One interesting finding is that logical writing appears to require more varied vocabulary, compared with other writing genres (i.e. a smaller proportion of the generated characters belong to the top-n most-frequently used characters).

Table 14.: Ratio of each type of time window for different typist groups

		Avg. num. of time windows in a session	State ratio					
			O	R	F	L	U	N
Touch typist	Reminiscent	304	11.9%	11.3%	27.1%	13.2%	14.0%	22.5%
	Logical	282	10.7%	13.3%	26.9%	14.0%	11.5%	23.6%
	Creative	273	11.8%	11.8%	26.5%	12.6%	13.5%	23.8%
Non-touch typist	Reminiscent	220	18.8%	11.4%	20.7%	14,2%	10.3%	24.5%
	Logical	201	19.3%	11.5%	20.0%	12.9%	11.1%	25.3%
	Creative	194	18.7%	11.9%	20.5%	13.8%	10.4%	24.7%

Table 14 shows the ratio of each type of time window introduced in section 1 for touch and non-touch typists. As expected, compared with touch typists, the proportion of Type *O* typing window is higher for non-touch typists since they often need to look at the keyboard while typing.

[1] http://lingua.mtsu.edu/chinese-computing/statistics/char/list.php

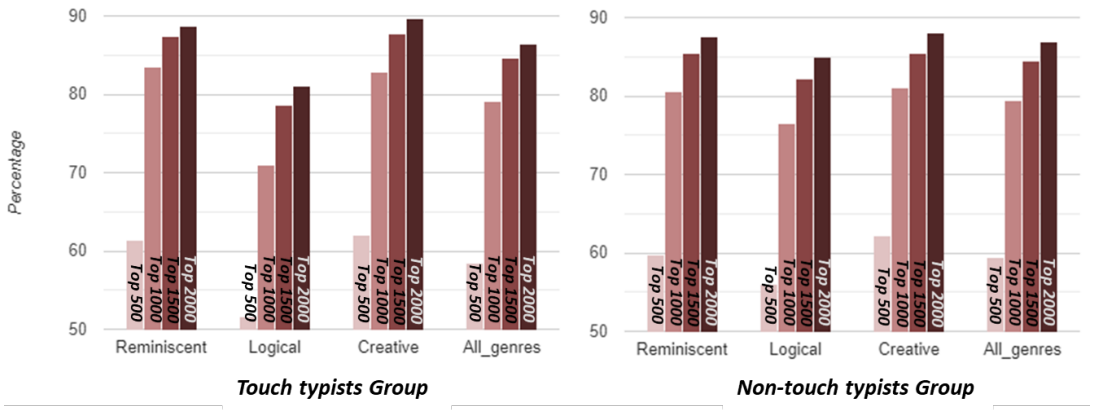

Figure 10.: Vocabulary usage across different writing genres for touch typists and non-touch typists: percentage of vocabulary in the essay belonging to top N frequently used Chinese characters, where N equals 500, 1000, 1500 and 2000

6.5. Data Verification

In our experiment, subjects were asked to write one article in each of three genres: reminiscent, logical, and creative respectively. For proper analysis, it is important to know whether the subjects were actually able to follow instructions and generate the appropriate articles for the requested genres, i.e. the content of the articles match the expected genre. This is especially important in the case of the child subjects, who may have less experience writing logical articles. To this end, we recruited two experts, who are high-school Chinese teachers from mainland China, to review the articles written by our subjects. We asked the experts to read the articles and label each of them as *reminiscent*, *logical*, *creative*, or *hard to decide*. Given an article, the experts label it only based on its content without knowing any prior knowledge, such as the expected genre. The articles are presented in random order to the experts.

It is not hard to imagine that different parts of an article may be categorized as different genres. For example, an article presents detailed instructions for cooking a dish (logical), may also include content which touches on the writer's memory and life experience (reminiscent), such as "when I first tried this, I ...". To take this into consideration, in addition to giving one overall genre label, we also asked the experts to rate the genres (*reminiscent*, *logical*, *creative*) for each article by distributing 5 points across the three genres. For example, if an expert thinks that a given article contains about 80% of logical content, and roughly 20% of reminiscent content, he/she would be expected to give the article the overall label of *logical* and rate the genres as **Reminiscent: 1, Logical: 4, and Creative: 0**.

Table 15 presents the results of the expert review. Our results suggest that all the articles are correctly written for the requested genres, even for the articles written by our child subjects. Content-wise, most of the children and young adults write logical articles with instructions for mathematical operations such as 2-digit number multiplication or division, while most of the older adults write articles on cooking or card games. According to the experts, all the logical articles are well written with reasonable and clear logical steps. It is also interesting to observe that reminiscent and creative articles also have some logical content, as can be seen in Table 15, where

Table 15.: Results of the expert review – Detailed ratings of articles written by subjects from different age groups

	Reminiscent writing			Logical writing			Creative writing		
	R.	L.	C.	R.	L.	C.	R.	L.	C.
All	**4.21**	0.73	0.06	0.13	**4.79**	0.08	0.02	0.95	**4.03**
Child	**4.24**	0.76	0.00	0.04	**4.86**	0.10	0.00	0.88	**4.12**
College	**4.09**	0.60	0.31	0.00	**4.91**	0.09	0.01	0.85	**4.14**
Elder	**4.25**	0.75	0.00	0.28	**4.66**	0.06	0.00	0.87	**4.13**

∗ **R.**: reminiscent genre; **L.**: logical genre; **C.**: creative genre

around 15% of the content in the reminiscent and creative articles was judged as belonging to the logical genre by the experts. This may be due to the fact that the writers feel the need to systematically present their narrations in order to make them believable or convincing.

Table 16.: P-values of ANOVA tests on article genre ratings for different age groups

	Reminiscent	Logical	Creative
P-value	0.13	0.11	0.15

Table 16 shows the p-values of one-way analysis of variance (ANOVA) tests (Howell, 2012) on the article genre ratings of different age groups. The resulted p-values suggest that there is no significant difference ($p > 0.05$) in the genre ratings across different age groups. In other words, the experts judge that all the subjects, including the children and elderly, were able to generate the appropriate articles in the requested genres.

7. Experimental Evaluation

We evaluate our statistics-based and sequence-based gaze-typing features on the task of detecting the genre of an article that a subject is currently working on. In this section, we first analyze features to understand gaze typing behaviors across different genres and then we build our article genre detection model based on the analysis results. The detection model is evaluated on the datasets that we construct and the performance will be reported at the end of the section.

7.1. Understanding Statistics-based Gaze-typing Features

In section 4, statistics-based gaze-typing features are extracted from different types of time window. Since our objective is to build a user-independent model, we want our features to be effective at capturing behavior differences across different subjects. However, for different subjects, the range of a feature can be entirely different. For example, some subjects are used to generating a series of Chinese characters in one pop-up candidates box and then revising them by correcting mistypings. On the other hand, some subjects are used to typing phrase by phrase or even character by character. This means that the range of the features: number of keystrokes (f_1^U and f_1^N) are completely different and the raw features f_1^U and f_1^N are not generalizable across

users. To solve this problem, we apply min-max normalization for all statistics-based gaze-typing features across different sessions of a same subject to mitigate the effect of user variation. After normalization, the ranges of all the features are within $[0, 1]$. Since the scope of normalization is across all the sessions in different article-genres of a same subject, so the normalized features are still capable of capturing the differences between the different article-genres and can be compared across different subjects and used in a user-independent fashion.

To better understand the gaze-typing behaviors, we analyze the window-level statistics-based gaze-typing features in different article-genre sessions by answering 2 questions: 1) whether there is a significant difference between different article-genre groups for each feature and 2) how they are different.

First, we group all the window-level features with the same type together and then they are divided into 3 groups: reminiscent, logical and creative, based on the genre of their corresponding article. A Kruskal Wallis H test (Vargha & Delaney, 1998) is then performed to test whether features in the three groups originate from the same distribution. In other words, if the test shows that a particular feature is significantly different, it means that feature can potentially capture the differences between writing articles in different genres. Kruskal Wallis H test is a non-parametric method, which is the extension of Mann–Whitney U test (Nachar et al., 2008) to support multiple groups (more than 2) comparison. Compared with the one-way analysis of variance test, Kruskal Wallis H test does not need the population to be normally distributed, nor does it assume that standard deviations of the groups are all equal.

To see out how these features are different across different article-genres, we apply the Dunn's test with Bonferroni correction (Dinno, 2015), an non-parametric post hoc test, on the features shown significant by the Kruskal Wallis H test. Table 17 lists all the significant features by the Kruskal Wallis H test with their p-values of Kruskal Wallis H test and p-values with correction of Dunn's test. For both p-values, if $p \leq 0.05$, then it will be considered as *significant* under such test. The mean values of all significant features are also shown in the table for comparison across different groups.

Rereading behaviors differ between the writing reminiscent and creative articles. When a subject composes an article in the creative genre, he/she tends to spend more time in rereading already-generated texts with more fixations, but each fixation is shorter in duration compared with composing an article in the reminiscent genre. Intuitively, the results make sense. Rereading behaviors appear more frequently in the translating phase and reviewing phase. Compared with reminiscent writing, composing an article in the creative genre requires a subject to continually ensure that the plot is reasonable. Therefore, it makes sense that they spend more time rereading the texts, and reread longer chunks of text.

We also observe some *pause* behaviors, when a subject stares at a position on the screen for a while during the typing period. During this time, the pop-up candidates box window remains on the screen, but there is little gaze movement and no keypresses. These *pauses* appear less often while composing reminiscent articles. One possible reason is that reminiscent writing is less complex compared with others. It is known that the frequency and duration of these *pause* behaviors are positively correlated with the complexity of the writing task (Wallot & Grabowski, 2013).

We apply the Kruskal Wallis H test to the data from the non-touch typists in a similar fashion. In their cases, most of the significant features are extracted from the Type F thinking window and the Type L and U typing windows. Significant features from the thinking phase include: the total duration of staring and thinking (f_3^F) and

Table 17.: Kruskal Wallis H test and Dunn's test results of significant statistics-based gaze-typing features for touch typists

Significant feature	P-value with correction of Dunn's test			P-value of Wallis H test	Reminiscent (Normalized mean)	Logical (Normalized mean)	Creative (Normalized mean)
	Reminiscent vs. Logical	Reminiscent vs. Creative	Logical vs. Creative				
Length of the reread texts in the type R thinking window (f_2^R)	1.00	1.00	**0.02**	**0.02**	0.30	0.27	0.32
Rereading duration in the type R thinking window (f_3^R)	0.31	**0.01**	**0.01**	**0.01**	0.13	0.14	0.17
Number of fixations in the type R thinking window (f_4^R)	0.07	**0.01**	1.00	**0.01**	0.10	0.12	0.14
Duration of fixations in the type R thinking window (f_5^R)	1.00	**0.02**	**0.01**	**0.01**	0.31	0.31	0.28
Total duration of staring and thinking in the type F thinking window (f_3^F)	**0.01**	**0.01**	1.00	**0.01**	0.12	0.13	0.13
Duration of typing in the type L typing window (f_1^L)	**0.01**	1.00	**0.01**	**0.01**	0.10	0.12	0.09
Keypress interval in the type L typing window (f_2^L)	**0.01**	0.92	**0.01**	**0.01**	0.10	0.11	0.09
Pause duration in the type N typing window (f_3^N)	**0.02**	**0.01**	1.00	**0.01**	0.08	0.10	0.11
Duration of typing in the type N typing window (f_5^N)	0.28	**0.04**	1.00	**0.04**	0.11	0.13	0.17
Time to look toward keyboard in the type 1 transition window (f_1^{Tr})	**0.01**	**0.01**	**0.01**	**0.01**	0.45	0.52	0.39

Table 18.: Kruskal Wallis H test and Dunn's test result of significant statistics-based gaze-typing features for non-touch typists

Significant feature	P-value with correction of Dunn's test			P-value of Wallis H test	Reminiscent (Normalized mean)	Logical (Normalized mean)	Creative (Normalized mean)
	Reminiscent vs. Logical	Reminiscent vs. Creative	Logical vs. Creative				
Duration of fixations in the type R thinking window (f_5^R)	**<0.01**	0.87	**0.01**	**<0.01**	0.28	0.33	0.30
Total duration of staring and thinking in the type F thinking window (f_3^F)	0.24	**0.01**	0.06	**<0.01**	0.19	0.20	0.16
Duration of fixations in the type F thinking window (f_5^F)	0.37	**<0.01**	0.10	**<0.01**	0.24	0.23	0.20
Duration of typing in the type L typing window (f_1^L)	1,00	**<0.01**	**<0.01**	**<0.01**	0.14	0.13	0.12
Keypress interval in the type L typing window (f_2^L)	1.00	**<0.01**	**0.01**	**<0.01**	0.15	0.14	0.16
Keypress interval in the type U typing window (f_2^U)	1.00	**<0.01**	**<0.01**	**<0.01**	0.27	0.26	0.22
Duration of typing in the type U typing window (f_4^U)	1.00	**<0.01**	**<0.01**	**<0.01**	0.28	0.28	0.23
Number of keystrokes in the type N typing window (f_1^N)	**<0.01**	1.00	0.05	**<0.01**	0.18	0.16	0.18
Time to look toward keyboard in the type 1 transition window (f_1^{Tr})	**<0.01**	**<0.01**	**<0.01**	**<0.01**	0.48	0.44	0.37

fixation duration (f_5^F, f_5^R) of the Type F and R thinking windows. Significant features from the typing phase are the keypress intervals (f_2^L, f_2^N) in both Type L and U typing windows and the typing duration (f_1^L, f_4^N) in both Type L and U typing windows.

The result of the Dunn's test with Bonferroni correction on the significant features are shown in Table 18. Based on the results, we can find that composing an article in creative genre has the most distinguishable typing behaviors, which are mainly shown in two aspects: keypress interval and typing duration. For creative writing, both the keypress interval and the typing duration are the shortest in Type L and U typing windows. A similar phenomenon is also found by Wallot et al. (2013) that keypress intervals are somewhat faster when the piece of writing is more complex.

We also observe that when a subject composes an article in the logical genre, she/he tends to have longer fixations when rereading the previously-generated texts than in other genres. One of the possible explanation is that these longer fixations are indicative of more complex language processing. Henderson, Choi, Luke, and Desai (2015) have observed that texts with higher degree of logical complexity require greater *attentional focus* and more effort in language processing, as subjects attempt to connect the linkage between different parts of the text. This increased cognitive activity manifests in longer fixations.

The f_1^{Tr} feature of the Type 1 transition window shows that there is a significant difference in the writing behavior between every pair of genres for both touch and non-touch typists. When subjects are composing in the creative genre, they exhibit the smallest normalized feature value of f_1^{Tr}, which means that a subject's gaze moves downward earliest when composing a creative article, compared to other genres. This phenomenon suggests that writing a creative article is a more cognitively complex task than the other two genres, since a higher cognitive load induces people to move their gaze away from the target, scan more hastily and at higher speed (M. X. Huang et al., 2016).

7.2. Understanding Sequence-based Gaze-typing Features

Sequence-based gaze-typing features are extracted from the behavior sequence of each session to capture the occurrence patterns of *indicative* patterns across the behavior-transition sequences. *Indicative* patterns are behavior subsequences which differ across different writing genres. A weighting scheme was previously defined in Section 5.2 to determine whether a subsequence is an indicative pattern. Potential indicative patterns are all the possible subsequences with length ranging from $[2, 4]$ time window transitions. The reason we restrict the maximum length of the pattern to 4 is that based on the observation, most of the clauses are generated within 4 time windows.

In this section, we address two important questions: 1) whether the weighting scheme can help us to select patterns with discriminating power, and 2) what the selected indicative patterns are.

Section 5.2 previously defined the means by which the indicativeness of a pattern can be quantified by the weighting (w), which can be computed through pf, rf and td. pf helps to avoid selecting a rare subsequence as a pattern and rf and td determine the discriminating power of a pattern from different perspectives: rf measures the differences of the pattern's occurrences between the positive and negative groups and td quantifies the difference in the number of occurrences within different writing stages between the positive and negative groups.

Figure 11 shows the top 100 largest rf weightings and top 100 largest $rf \cdot td$ weight-

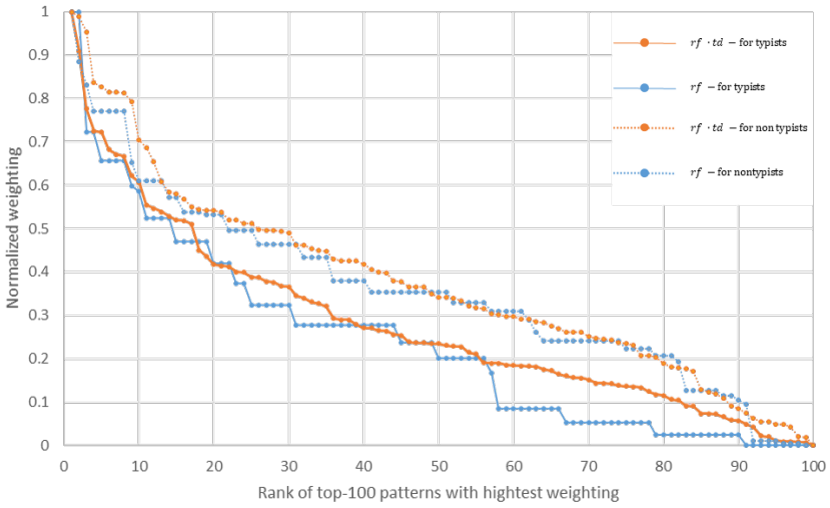

Figure 11.: Top-100 normalized rf weights and top-100 normalized rf weights for touch typists and non-touch typists

ings in descending order for both touch and non-touch typists. For easy comparison, the weightings are normalized into $[0, 1]$ range using Min-max normalization. The value of the 1^{st} largest weighting is mapped to 1 and the 100^{th} largest weighting is mapped to 0. It is obvious that many patterns share the same rf weighting. Even when the value of the weight is at a high level, this phenomenon still occurs quite often. The reason for this is that our sequence classification problem gives us 6 different states, where the transition is strictly between one of O, R, F states and one of L, U, N states. This gives us a total of $3^3 = 27$ different transitions, which may not be complex enough to cover the different behaviors evidenced in our dataset. Figure 11 shows that many patterns, which appear to be quite dissimilar, do share the same rf weight. This suggests that the rf term may not be sufficient enough on its own to quantify the discriminating power of the pattern. We therefore involve the td term for additional information.

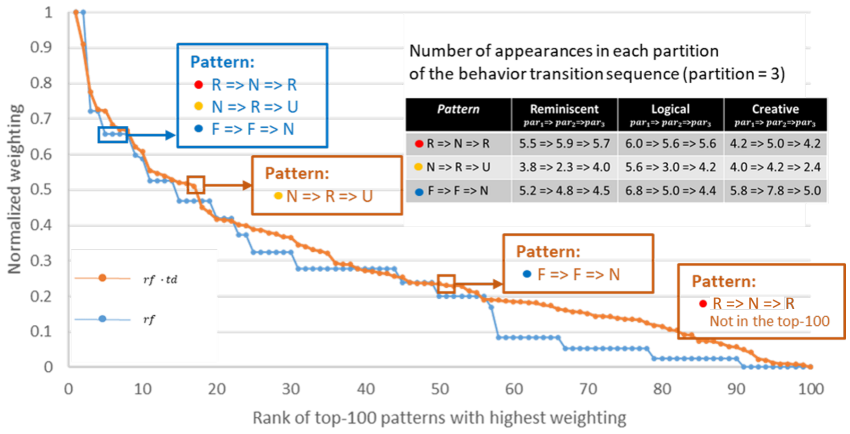

Figure 12.: Examples of how the td term further distinguish the discriminating power of patterns with same rf weight

As an example, Figure 12 compares three patterns: $R => N => R$, $N => R => U$ and $F => F => N$, based on their rf and $rf \cdot td$. The patterns have the same rf weighting value of 0.65. However, the $rf \cdot td$ weightings of these 3 patterns are completely different:

- $N => R => U$ has the highest $rf \cdot td$ of 0.51. In reminiscent and logical writing, it decreases in frequency as the writer approaches the middle part of the writing, and then increases again as the writer approaches the conclusion of the writing period. In creative writing, however, this pattern slightly increases as the writer approaches the midpoint of the writing activity, but then decreases dramatically as the conclusion approaches.
- $F => F => N$ has a lower $rf \cdot td$ of 0.23. From the figure, it can be seen that even though there is some difference in the behavior of the pattern across different genres, the difference is less dramatic than for $N => R => U$. For $R => N => R$, its $rf \cdot td$ does not make the top-100 list.

Table 19.: Top-5 selected patterns for both touch typists and non-touch typists

	Touch typist		Non-touch typist	
	$pf \cdot rf$	$pf \cdot rf \cdot td$	$pf \cdot rf$	$pf \cdot rf \cdot td$
1	F => F => F	O => O => O	L => F => F	L => F => F
2	F => F => N	L => N => R	N => N => R	N => N => R
3	N => F => F	N => O => O	N => R => N	N => F => N
4	F => N => F	N => N => R	F => N => R	F => N => R
5	F => F => U	O => O => N	F => U => N	F => U => N

Table 19 lists top-5 selected indicative patterns for both touch and non-touch typists based on the $pf \cdot rf$ and $pf \cdot rf \cdot tf$ weightings. We note that in most of the selected patterns, at least two of the three states are the same (e.g. $F => F => N$ has two F states). This suggests that the indicative patterns describe a period of time during which the subject's state is relatively stable. For example, the pattern $F => F => F$ describes the behavior in which a subject stares at the screen for a while (presumably thinking) before typing.

The top-ranked indicative patterns differ depending on the weighting terms used. In particular, for touch typists, the top-5 indicative patterns selected based on the $pf \cdot rf$ weighting contains more F states, whereas the $pf \cdot rf \cdot td$ weighting more highly weighs the O states. Compared to touch typists, the top 5 indicative patterns selected based on the $pf \cdot rf$ weighting and the $pf \cdot rf \cdot td$ for non-touch typists are more similar to each other. One possible reason is that non-touch typists are less efficient when typing, and the process of hunting for the correct key on the keyboard dominates the behavior across the entire process of writing the article.

7.3. Evaluating the Performance of Writing Genre Detection

Our writing genre detection method is evaluated on the datasets constructed in Section 6.3. In real-life applications, a method should be able to work for a never-seen-before new user. Therefore, we employ a *leave-one-subject-out* cross-validation mechanism for evaluation. Specifically, a supervised learning model will be built based on the statistics-based gaze-typing features and the sequence-based features. The model will be trained on all but one of the subjects, and evaluated on the remaining subject. The process will be iterated for N_s times, where N_s equals to the total number of

subjects. Since we build separate models for touch and non-touch typists, the overall performance of our approach is calculated as the weighted average of the performance achieved over the touch and the non-touch groups.

We first investigate the proper parameter values for our approach. The parameter (n_{par}) determines the number of partitions that a behavior sequence will be segmented into, which will be used to compute the td term. Physically, it also represents the number of writing stages, so it is not reasonable to have an overlarge or oversmall n_{par}. The parameter n_{select} denotes the number of indicative patterns that will be considered, sorted by weight. A too-small n_{select} may omit some useful patterns, but an over-large n_{select} is too large will select some non-indicative patterns, thus diluting the impact of the truly indicative features.

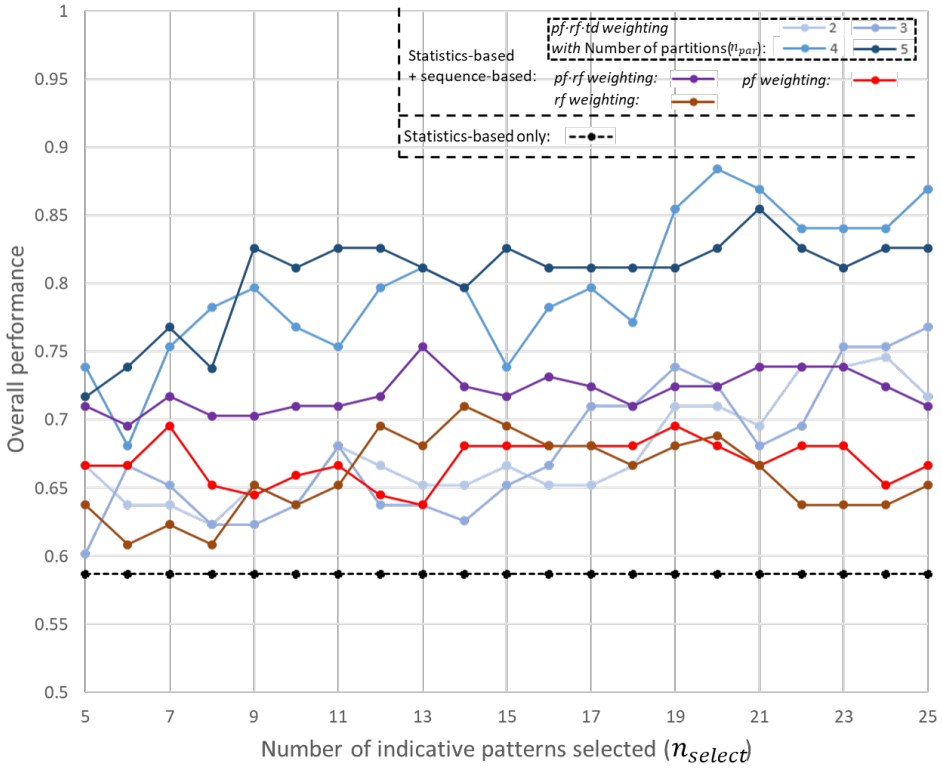

Figure 13.: Overall performance trends of writing genre detection approach with different number of partitions n_{par} and number of indicative patterns selected n_{select}

In this experiment, we explore the impact of different value combinations of n_{par} and n_{select} on the performance. Linear support-vector machine (SVM) models are built based on the concatenation of the statistic-based features with the sequence-based features, which is generated by different values of n_{par} and n_{select}. Figure 13 summarizes the results.

Compared with the overall baseline of 36.2%, which is achieved by predicting every instance as the majority class, the best performance of our approach ($n_{par} = 4, n_{select} = 20$) can achieve 88.4% accuracy, which is quite promising. Table 20 and 21 show the detailed confusion matrices.

We notice from the figure that when $n_{par} = 4$ or 5, our approach always yields the best performance. It makes sense, since normally most articles can be divided into 3

Table 20.: Confusion matrix of the article-category detection for touch typists

Ground truth \ predicted as	Reminiscent	Logical	Creative
Reminiscent	**18**	1	2
Logical	0	**24**	1
Creative	1	2	**22**

Table 21.: Confusion matrix of the article-category detection for non-touch typists

Ground truth \ predicted as	Reminiscent	Logical	Creative
Reminiscent	**22**	1	2
Logical	4	**16**	1
Creative	1	0	**20**

parts: introduction, body and conclusion and the body part has around $2-3$ times the length as the length of the introduction and the conclusion parts. We note that when $n_{par} = 4$ and $n_{par} = 5$, the max performance is achieved when n_{select} is around 20, which also meets our intuition that selecting too many patterns will worsen the overall performance since non-indicative patterns may be included.

The figure also presents the performance trend of linear SVM models built on the statistics-based features and sequence-based features by using the $pf \cdot rf$ weighting scheme with different n_{select} values, pf weighting scheme, rf weighting scheme and the performance of only using statistics-based features. It is clear that with reasonable values of n_{par}, the overall performance of the $pf \cdot rf \cdot td$ weighting scheme is always better than the others.

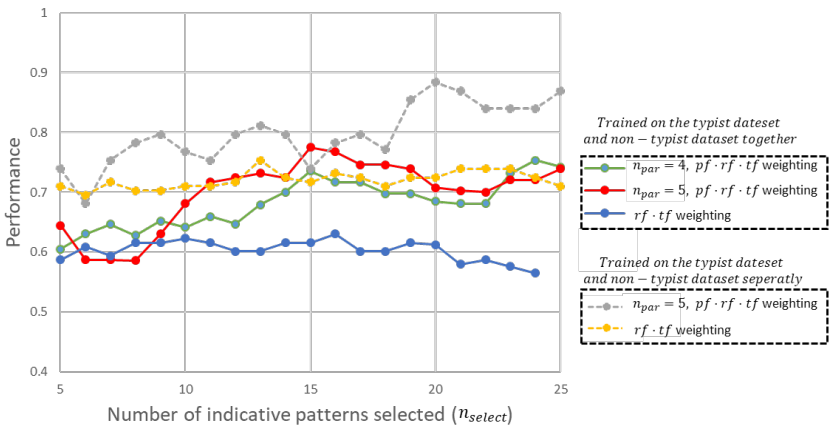

Figure 14.: Performance trends of writing genre detection approach trained on the touch typists dataset and non-touch typists dataset together

We also evaluate the performance of our approach without differentiating between touch typists and non-touch typists. We construct a new dataset by combining data from all subjects and train linear SVM models on the new dataset with different values of n_{par} and n_{select}. Based on the previous results, potential values of n_{par} are 4 and 5. Figure 14 shows the performance trends. The best performance is around 77%, which is attained when $n_{par} = 5$ and $n_{select} = 15$. According to the figure, we see that the

Table 22.: Article-category detection for different age groups

	Article-category detection for				
	Children	College students	Elders	Touch typists in elderly-age group	Non-touch typists in elderly-age group
CCR	87.0%	83.3%	77.8%	91.2%	88.9%

performance of the $pf \cdot rf \cdot td$ weighting scheme is better than that of the $pf \cdot rf$ scheme, which is consistent with the results of training separate models for touch typists and non-touch typists. It can also be seen that the combined model performs worse than training separate models for different levels of typing ability. A possible reason is that gaze-typing behaviors differ so much between touch typists and non-touch typists, and these inconsistent behaviors may confuse the model.

Finally, we evaluate the performance of our approach across different age groups to ascertain the effect of the age factor. As shown in Table 11, the college student group are all touch typists and all but one subject in the child group are non-touch typists, as are around 33% of subjects in the elderly-age group. We therefore further divide the elderly-age group into the touch typist elderly-age group and non-touch typist elderly-age group. For each specific age group, we then construct a linear SVM model on the concatenation of statistics-based and sequence-based features with $pf \cdot rf \cdot td$ weighting scheme, where $n_{par} = 4$ and $n_{select} = 20$, which achieves the best performance in the previous evaluation.

Table 22 presents the results of the evaluation by age group. It can be seen that the performances for the child-age group and college student group are close to the best performance achieved by differentiating the touch typists and the non-touch typists (Figure 13). However, for the elders, the performance drops more significantly compared to the two other groups, approaching the performance we achieved in Figure 14 when there was no differentiation between the touch typists and the non-touch typists. However, when the elders are broken down into touch typists and non-typists, the performance improves significantly, even outperforming the best performance previously achieved. These observations suggest that (1) the typing skill has a bigger effect on writing genre detection than the age factor, and (2) the age factor may provide additional information that can contribute additionally to the performance of writing genre detection after the dominant factor (typing skill) is accounted for.

8. Discussion

In our experiment, we successfully utilize statistics-based gaze-typing features and sequence-based features to determine the genre of an article during the process of writing. Both statistics-based and sequence-based features are based on the gaze-typing behaviors. Our results suggest the cognitive process of generating articles in different genres can be inferred by the gaze-typing behaviors. Based on the result presented in the Table 17, when a subject is composing a complex article, which involves more idea-generating phases and text-organizing phases, he/she will reread previous generated texts more frequently. The purpose of rereading previous generated texts could be providing hints of what to write next or helping organize the current generating sentence. They can be differentiated by the length of the rereading texts since organizing the sentence needs reread longer length to ensure the correctness both

logically and semantically. (different task)

Unlike the copy-type tasks used in previous work, in which keypress intervals were consistent across the task, we notice that pauses exist throughout our task when the subject composes their own texts. This is most likely because the process of composition requires subjects to convert their ideas into text in addition to generating that text via the keyboard. This hypothesis is supported by the observation that longer pauses are observed in logical and creative writing, which require the subject to imagine and visualize a scenario, and also to express it coherently in textual language with logical and semantic correctness. These requirements presumably require more cognitive effort than reminiscent writing, in which subjects are simply asked to recall an event. We also observe that the f_1^{Tr} feature, which captures the action in which non-touch typing subjects shift their gaze away from the screen and towards the keyboard before they start typing, appears earlier for logical and creative writing. This is consistent with previous work (M. X. Huang et al., 2016) on a different domain (mathematics calculations), which shows that when a subject is in a high cognitive load state, they are more likely to move their gaze away from the target earlier, and at higher speed.

The result in Figure 14 shows that the best correct classification rate (CCR) is achieved by combining statistics-based and sequence-based gaze-typing features. The transition between different kinds of behaviors also appear to capture the information of the cognitive process of writing, especially with certain behaviors that appear frequently during a particular process during the activity. Moreover, we find that using the sequence-based features alone achieves much better performance than the statistics-based features achieve on their own. Writing original articles is a *dynamic* process, in which writers' gaze and typing behaviors may change as they move through different writing phases. Statistics-based features focus more on the overall behaviors from the whole writing activity, which does not account for the change in the writer's state of mind as the writing progresses. Our sequence-based features were developed to extract writers' gaze-typing behaviors as they progress across a writing activity. The success of the sequence-based features illustrates that variation in the gaze-typing behaviors is a powerful indicator of users' cognitive and mental state in writing tasks. We believe that variation-based behaviour features (such as the sequence-based features) can be extended to other applications, such as stress detection (Hernandez et al., 2014; M. X. Huang et al., 2016; Wang et al., 2019) and behavior-based continuous authentication (Bours & Mondal, 2015; Kumar, Phoha, & Serwadda, 2016; Locklear et al., 2014).

We also observe that the same behavior may have different causes that are affected by the typing proficiency. Since they need to look at the keyboard while typing, non-touch typists' eye gaze movements exhibit many saccades with greater variation along the y-axis, and the eye gaze cannot be captured for large amounts of time. For touch typists, saccades with greater variation along the y-axis are generally related to rereading the previously generated texts, and periods of time when the subject's gaze is off-screen are often associated with deep thought and planning what to write next. These behaviors, though superficially the same, have very different causes, which argues for the need to train separate models based on the typing proficiency of the subject. Our experimental results demonstrate that training separate models indeed achieves better performance than training a single model to cover both touch and non-touch typists, and the impact of typing skill is far greater than the impact of age, at least on our task of writing genre detection. A model trained with data from both touch and non-touch typists is easily confused by behaviors that have the same patterns but different causes. This implies that combining data from different groups to

increase the size of training data is not always helpful, as it risks conflating data with different root causes. Human-computer interaction studies often involve data from different groups, particularly different subject populations. Our results suggest that one should be very careful with managing the training data based on the understanding of user behaviors, a point which is seldom mentioned in previous work. We hope that our finding can benefit the human-computer interaction community and lead to better behavior-based models.

9. Conclusion

In this paper, we explore the gaze-typing behaviors of subjects who are producing original texts across different genres in Chinese by using the Pinyin input method. 46 subjects are involved in the experiment and they are all native Chinese speakers. Experiments are conducted via the writing tasks, with which subjects are required to compose three articles in genres of reminiscent, logical, and creative. Statistics-based gaze-typing features and sequence-based gaze-typing features are extracted to capture the different writing cognitive processes while writing in different genres. To evaluate the performance of our approach, we construct touch typists and non-touch typists dataset by differentiating different levels of typing skill, where each dataset contains 23 subjects. An user-independent classifier is then constructed based on the extracted features to detect the writing genres for each dataset and achieves overall 88.4% accuracy.

Our results indicate that when people composing articles in different genres, their writing cognitive processes are different, which can be inferred by the gaze-typing behaviors, especially, rereading behavior, pauses during typing and transitions between different gaze-typing behaviors. We also illustrate that the gaze-typing features can efficiently infer the writing cognitive process if the subject's typing proficiency is taken into account when training the machine learning model.

In a nutshell, our results are promising and provide a more in-depth understanding of human gaze-typing behaviors of writing.

Funding

This work was partially supported by the Hong Kong Research Grant Council and the Hong Kong Polytechnic University under Grants PolyU 5222/13E and 156002/19H.

References

Alamargot, D., Chesnet, D., Dansac, C., & Ros, C. (2006). Eye and pen: A new device for studying reading during writing. *Behavior Research Methods*, *38*(2), 287–299.

Beauvisage, T. (2009). Computer usage in daily life. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 575–584).

Bieg, H.-J., Chuang, L. L., Fleming, R. W., Reiterer, H., & Bülthoff, H. H. (2010). Eye and pointer coordination in search and selection tasks. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 89–92).

Bours, P., & Mondal, S. (2015). Continuous authentication with keystroke dynamics. *Norwegian Information Security Laboratory NISlab*, 41–58.

Butsch, R. L. (1932). Eye movements and the eye-hand span in typewriting. *Journal of Educational Psychology*, *23*(2), 104.

Chukharev-Khudilaynen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research*, *6*(1), 61.

Debole, F., & Sebastiani, F. (2004). Supervised term weighting for automated text categorization. In *Text Mining and Its Applications* (pp. 81–97). Springer.

Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using dunn's test. *The Stata Journal*, *15*(1), 292–300.

Feit, A. M., Weir, D., & Oulasvirta, A. (2016). How we type: Movement strategies and performance in everyday typing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4262–4273).

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387.

Gladis, S. D. (1993). *Writetype: Personality types and writing styles.* Human Resource Development.

Henderson, J. M., Choi, W., Luke, S. G., & Desai, R. H. (2015). Neural correlates of fixation duration in natural reading: Evidence from fixation-related fmri. *NeuroImage*, *119*, 390–397.

Hernandez, J., Paredes, P., Roseway, A., & Czerwinski, M. (2014). Under pressure: sensing stress of computer users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 51–60).

Howell, D. C. (2012). *Statistical methods for psychology.* Cengage Learning.

Huang, J., White, R., & Buscher, G. (2012). User see, user point: gaze and cursor alignment in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1341–1350).

Huang, M. X., Li, J., Ngai, G., & Leong, H. V. (2016). Stressclick: Sensing stress from gaze-click patterns. In *Proceedings of the 2016 ACM on Multimedia Conference* (pp. 1395–1404).

Inhoff, A. W., & Gordon, A. M. (1997). Eye movements and eye-hand coordination during typing. *Current Directions in Psychological Science*, *6*(6), 153–157.

Johansson, R., Wengelin, Å., Johansson, V., & Holmqvist, K. (2010). Looking at the keyboard or the monitor: relationship with text production processes. *Reading and Writing*, *23*(7), 835–851.

Joshi, A., Ganu, A., Chand, A., Parmar, V., & Mathur, G. (2004). Keylekh: a keyboard for text entry in indic scripts. In *Chi'04 extended abstracts on human factors in computing systems* (pp. 928–942).

Klein, P. D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review*, *11*(3), 203–270.

Kumar, R., Phoha, V. V., & Serwadda, A. (2016). Continuous authentication of smartphone users by fusing typing, swiping, and phone movement patterns. In *2016 ieee 8th international conference on biometrics theory, applications and systems (btas)* (pp. 1–8).

Lan, M., Tan, C. L., & Low, H.-B. (2006). Proposing a new term weighting scheme for text categorization. In *Aaai* (Vol. 6, pp. 763–768).

Likens, A. D., Allen, L. K., & McNamara, D. S. (2017). Keystroke dynamics predict essay quality. In *Cogsci.*

Locklear, H., Govindarajan, S., Sitová, Z., Goodkind, A., Brizan, D. G., Rosenberg, A., ... Balagani, K. S. (2014). Continuous authentication with cognition-centric text production and revision features. In *Ieee international joint conference on biometrics* (pp. 1–8).

Logan, G. D. (1983). Time, information, and the various spans in typewriting. In *Cognitive aspects of skilled typewriting* (pp. 197–224). Springer.

Meena, Y. K., Cecotti, H., Wong-Lin, K., & Prasad, G. (2016). A novel multimodal gaze-controlled hindi virtual keyboard for disabled users. In *2016 ieee international conference on systems, man, and cybernetics (smc)* (pp. 003688–003693).

Nachar, N., et al. (2008). The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*,

4(1), 13–20.

Papoutsaki, A., Gokaslan, A., Tompkin, J., He, Y., & Huang, J. (2018). The eye of the typer: a benchmark and analysis of gaze behavior during typing. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (p. 16).

Ramos, J., et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning* (Vol. 242, pp. 133–142).

Rayner, K., Smith, T. J., Malcolm, G. L., & Henderson, J. M. (2009). Eye movements and visual encoding during scene perception. *Psychological Science*, *20*(1), 6–10.

Rodden, K., Fu, X., Aula, A., & Spiro, I. (2008). Eye-mouse coordination patterns on web search results pages. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems* (pp. 2997–3002).

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (pp. 71–78).

Samura, T., & Nishimura, H. (2009). Keystroke timing analysis for individual identification in japanese free text typing. In *2009 iccas-sice* (pp. 3166–3170).

Schiffman, H. R. (1990). *Sensation and perception: An integrated approach.* John Wiley & Sons.

Torrance, M., Johansson, R., Johansson, V., & Wengelin, Å. (2016). Reading during the composition of multi-sentence texts: an eye-movement study. *Psychological Research*, *80*(5), 729–743.

Van Waes, L., Leijten, M., & Quinlan, T. (2010). Reading during sentence composing and error correction: A multilevel analysis of the influences of task complexity. *Reading and Writing*, *23*(7), 803–834.

Vargha, A., & Delaney, H. D. (1998). The kruskal-wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics*, *23*(2), 170–192.

Wallot, S., & Grabowski, J. (2013). Typewriting dynamics: What distinguishes simple from complex writing tasks? *Ecological Psychology*, *25*(3), 267–280.

Wang, J., Fu, E. Y., Ngai, G., Leong, H. V., & Huang, M. X. (2019). Detecting stress from mouse-gaze attraction. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (pp. 692–700).

Wang, J., Liu, P., She, M. F., Nahavandi, S., & Kouzani, A. (2013). Bag-of-words representation for biomedical time series classification. *Biomedical Signal Processing and Control*, *8*(6), 634–644.

Zheng, Y., Xie, L., Liu, Z., Sun, M., Zhang, Y., & Ru, L. (2011). Why press backspace? understanding user input behaviors in chinese pinyin input method. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 485–490).