SmartRSVP: Facilitating Attentive Speed Reading on Small Screen Wearable Devices

Wei Guo, Jingtao Wang
Department of Computer Science
Learning Research and Development Center (LRDC)
University of Pittsburgh, Pittsburgh, PA 15260 USA
{weg21, jingtaow}@cs.pitt.edu

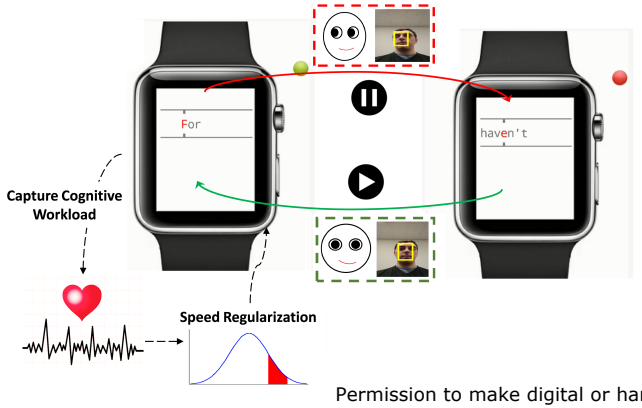

Figure 1. SmartRSVP in action.

CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
ACM 978-1-4503-4656-6/17/05.
http://dx.doi.org/10.1145/3027063.3053176

Abstract
Smart watches can enrich everyday interactions by providing both glanceable information and instant access to frequent tasks. However, reading text messages on a 1.5-inch screen is inherently challenging, especially when a user's attention is divided. We present SmartRSVP, an attentive speed-reading system to facilitate text reading on small screen wearable devices. SmartRSVP leverages camera-based *visual attention tracking* as a play/pause control channel, and uses *implicit physiological signal sensing* to adjust speed in real-time to make text reading via **R**apid **S**erial **V**isual **P**resentation (RSVP) more enjoyable and practical on smart watches. Through two pilot studies, we received positive feedback on the visual attention controlling feature and confirmed the feasibility of the speed adjusting feature of SmartRSVP. This paper reports the preliminary results of the studies.

Author Keywords
Smart watch; RSVP; Gaze tracking; PPG; Heart rate variability; Cognitive workload; Visual attention

Introduction

Small screen wearable devices are flourishing nowadays. By staying on users' wrists, smart watches provide instant access to important notifications and frequent tasks. Smart watches are also ideal for tracking users' activities and physiological signals for personal wellbeing. Although many interaction techniques [4] and *input* modalities [12, 15] have been invented for smart watches, text reading on a 1.5-inch small display, the primary *output* channel for today's smart watches, is inherently challenging.

At least three major challenges arise when a user reads textual information on a smart watch. First, the small screen of a watch only affords showing three or four words per line in small fonts, thus demanding a higher cognitive workload and more frequent lateral movements of eye gaze, i.e., *saccade*; Second, showing fewer words per screen also leads to more text scrolling actions, exacerbating the notorious "fat finger problem" in scenarios where both hands of a user are occupied; Third, text reading on a watch increases the likelihood of experiencing *divided attention* and *higher interruption* due to the envisioned usage scenarios. Paradoxically, the ever growing amount and type of information accessible via smart watches *increases* our exposure to such reading interfaces.

We present SmartRSVP (figure 1), a novel speed-reading system to facilitate text reading on small screen wearable devices. SmartRSVP leverages real-time *visual attention tracking* and *implicit physiological signal sensing* to make text reading via **R**apid **S**erial **V**isual **P**resentation (RSVP) more enjoyable and practical on smart watches. SmartRSVP determines the *visual attention* of a user via camera-based facial orientation estimation and eye gaze tracking, and leverages the information on *visual attention* to play/pause the presentation of dynamic texts. SmartRSVP also uses the *cognitive workload* inferred from Heart Rate Variability (HRV) features to regulate the speed of RSVP. Overall, SmartRSVP employs the spatial and temporal efficacy of the RSVP technique, and reduces its high workloads in both visual attention and cognitive processing via a perceptual and affect-aware interface. In this paper, we first describe the design of SmartRSVP, and then report preliminary findings from two studies.

Related Work

RSVP is a visualization technique that displays textual information one word at a time[1] in sequential order. RSVP has both spatial and temporal efficacy when compared with traditional reading. However, RSVP also raises unique challenges such as higher visual attention [3], higher recovery cost, *attentional blink* [18], and *repetition blindness* [14]. In this paper, we explore how unique sensors, such as the front camera[2] and the photoplethysmography (PPG) sensor on a smart watch, can be used to make RSVP more practical for daily use on small screen wearable devices.

Our visual attention tracking feature was inspired by the "gaze locking" technique [22] and the seeTXT technique [6]. Gaze locking refers to the robust binary sensing of eye contact in a static image via computer vision algorithms. seeTXT relies on a customized infrared eye-contact sensor (ECS) to augment media consumption on mobile devices. In comparison, the SmartRSVP technique focuses on improving the speed-reading experiences on smart watches via on-device sensing capabilities. In addition to real-time visual

[1] Or one visual item a time for stimuli such as pictures.

[2] Although cameras are only available on a small portion of smart watches today, e.g. Samsung Gear 2, more device manufacturers may include cameras once compelling usage scenarios of camera are discovered.

attention tracking, SmartRSVP can also infer cognitive workload from implicit PPG sensing on unmodified mobile devices. Hansen et al [11] demonstrated the feasibility of using a commercial gaze tracker to control RSVP playback running on a PC. Dinger and colleagues [7] demonstrated gaze controlled RSVP with a head-mounted gaze tracker and visual markers. In comparison, SmartRSVP does not rely on external eye trackers and also offers implicit cognitive state sensing for the dynamic speed regulation of RSVP.

Heart rate signals are widely explored to infer learners' cognitive and affective states in different interaction tasks, such as learning [13], operating user interfaces [21], and gaming [10]. In this paper, we propose the sensing and modeling of PPG signals to facilitate speed-reading on smart watches. We believe that with the ability to stay on a user's wrist 24/7 and collect the user's physiological signals implicitly, smart watches will become a promising test bed for the next generation of affect/emotion aware interfaces.

Design of SmartRSVP

Figure 1 shows SmartRSVP in action. SmartRSVP continuously monitors the *visual attention* of a user by analyzing the user's facial orientation and gaze contact in real-time through a front camera. RSVP will pause if there is no human face in the camera viewport, or the user's eye gaze is not in direct contact with the watch screen. SmartRSVP also infers the *cognitive workload* of the user via implicit PPG sensing through a dedicated PPG sensor or a back camera[3] [10]. The speed of RSVP will be adjusted based on the *cognitive workload*.

SmartRSVP includes three major components: 1) The RSVP module; 2) Algorithms for tracking and using the owner's visual attention; and 3) a user-independent statistical model to predict the cognitive states of the owner.

RSVP

We use a 20dp monospace font with a 2.1mm average height to render words in our RSVP module, which provides good legibility on a 1.5-inch watch screen and can display words as long as 12 characters without line breaking or resizing. SmartRSVP also visualizes the Optimal Recognition Point (ORP) [2] in a red color (Figure 1). ORP intends to make the gaze fixation point of a word, which may not be the first character of a word, stay in a fixed region to avoid unintended saccades when the gaze fixes on words of different lengths [2]. The display speed of our RSVP module can vary from 200 wpm to 500 wpm.

Visual Attention Tracking

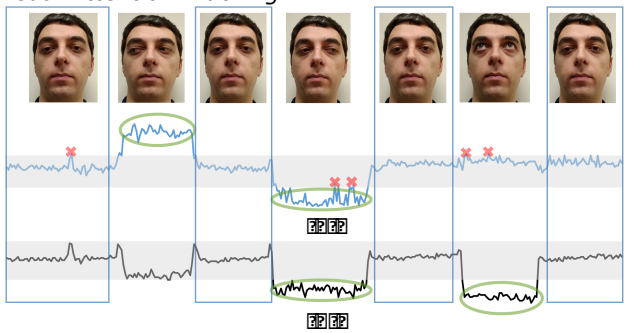

Figure 2. Visual Attention Tracking via face detection, face alignment and eye contact estimation. The x-axis is time (~17 sec). Row (a) is the predicted horizontal eye gaze locations: Row (b) is the predicted vertical eye gaze locations. Green circles highlight moments when user are not having gaze contact with his smart watch.

[3] We also explored the same front camera to extract both visual attention and PPG signals from facial images. We chose a dedicated PPG sensor or a separate camera pointing at the arm or the finger for higher signal-noise-ratio (SNR).

Figure 3. Three reading interfaces. From top to bottom: Normal Watch Reading Interface, Traditional RSVP interface, and SmartRSVP.

Due to the limited availability of front facing cameras on smart watches, we used a Google Nexus 5x smart phone running Android 6.0 to simulate a 42.0mm by 35.9mm smart watch screen. Following practices of existing research on smart watches [5, 17], we allocated the same physical region on the screen of the Nexus 5x for display and touch input.

Each image frame captured by the front camera goes through the following three steps to generate a binary prediction on visual attention. 1) *Face detection*. A Viola-Jones face detector is used to detect the existence and location of a human face; 2) *Face Alignment*. We use Cascaded Pose Regression [8] to estimate the facial orientation and landmark points on a face; and 3) *Eye contact estimation*. Similar to [16, 22], we rely on the location of the pupil relative to the rest of the eye to estimate the direction of eye gaze. Given the small display size of smart watches, it's not necessary to estimate the absolute location of eye gaze on the watch. Instead, we trained a *binary* eye contact classifier from five volunteers. We also use a low-pass filter to reduce false positives and false negatives from per-frame estimations. Figure 2 shows the continual output of the eye gaze prediction algorithm and the binary eye contact estimation.

We used the Qualcomm Snapdragon SDK to accelerate the tracking process. The per-frame image processing time is 17 ms, and we can achieve 21 frames per second on the hexacore Snapdragon 808 CPU in the Nexus 5x. We are exploring the use of alternative face alignment algorithms such as Regressing Local Binary Features [19] to further accelerate the tracking speed.

Cognitive State Inference

Since the current SmartRSVP prototype runs on a smart phone rather than a smart watch, we do not have access to the heart rate sensor on mainstream smart watches such as Apple Watch, Moto 360, and Samsung Gear 2. Instead, we used commodity camera-based PPG sensing through the back camera of a smart phone. After detecting the transparency change of fingertips via the build-in camera [1], we used the LivePulse algorithm [10] to extract the raw PPG waveforms. We expect cleaner PPG signals and higher prediction accuracies after switching to dedicated PPG sensors on smart watches.

It's possible to calculate the HRV from instant heartbeats and use a constant threshold on HRV to determine the cognitive workload of a user [21]. However, we found that we can achieve significantly more robust predictions by extracting multiple dimensions of features from the raw PPG waveforms and training a user-independent classifier to predict a user's cognitive workload.

We used a fixed-size sliding window to extract features from the temporal PPG signal. We extract 9 dimensions of heart rate and HRV features from each window, including 1) MHR (mean heart rate); 2) SDHR (standard deviation of heart rates); 3) rMSSD (the square root of the mean squared adjacent RR intervals' difference); 4) pNN12 (percentage of more than 12ms difference between adjacent RR-intervals); 5) pNN20, 6) pNN50; 7) MAD (median of absolute deviation of RR-interval); 8) AVNN (average RR-intervals); and 9) SDNN (standard deviation of the RR-intervals). After normalization, these 9 dimensions of features are used

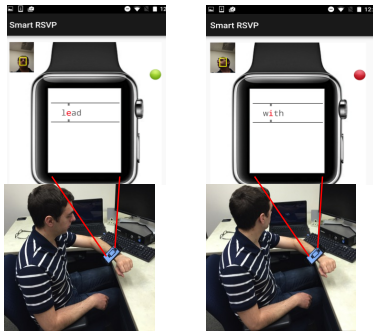

Figure 4. Distracters (random 3-digit numbers) appear on a 15-inch laptop screen. Left: reading an email message via SmartRSVP; Right: turning left to read the distracter.

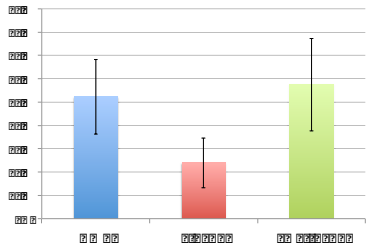

Figure 5. Comprehension rates by reading interfaces. Error bars represent one standard deviation.

to train a statistical classifier to predict users' cognitive workload within this sliding window.

We ran two studies to understand the performance and preference of SmartRSVP. During the first user study, we evaluated SmartRSVP together with today's reading interface on smart watches as well as traditional RSVP in a sitting condition. We took a closer look at the cognitive-state inference module in SmartRSVP together with a controlled color-counting task in the second user study.

User Study 1

We ran an 18-participant (3 females) within-subjects user study to evaluate the efficacy of the speed-reading technique and visual attention based play/pause channel of SmartRSVP.

The study included three interfaces, i.e. Normal watch reading (NWR), traditional RSVP (T-RSVP), and SmartRSVP (Figure 3). Each participant read 10 email messages * 3 reading interfaces = 30 unique email messages (μ = 47 words or 3.5 sentences) in total. To simulate reading activities under divided attention, we also included 3 distractions during each reading (Figure 4). Every 4 to 6 seconds, the laptop generated a beep sound, and a 3-digit random number was shown on the laptop screen lasting for 2 seconds during reading. Once the participant heard the beep sound, she was required to look at the number (Figure 4, right) and read it out loud. The participant could resume the reading task after reading the number out. After finishing each email, the participant then answered one literal question to test her text comprehension. There are three levels of text comprehension, i.e. literal, inferential, and evaluative [9]. We only used literal questions, i.e. recalling key information that was explicitly stated in the email, in our study because we focused on evaluating and comparing reading interfaces rather than testing the language and logical skills of participants. At the beginning of the user studies, users went through a warm up session to get familiar with the interfaces as well as choose a desirable speed that applied to NWR (Display duration = Number of Words / speed + 3 seconds for distractions), T-RSVP and SmartRSVP. The average user selected speed was 220 wpm in this study.

Results

Figure 5 shows the average comprehension rates of the three reading interfaces under distracted visual attention. There are significant differences of comprehension rates between NWR and T-RSVP (52.2% vs. 23.9%, $p<0.0001$) as well as between SmartRSVP and T-RSVP (57.5% vs. 23.9%, $p<0.0001$). However, there was no significant difference between NWR and SmartRSVP (52.2% vs. 57.5%, $p=0.39$). Similar for the reading efficiencies (actual reading speed * comprehension rate, wpm): there were only significant differences between NWR vs. T-RSVP (65.16 wpm vs. 43.93 wpm, $p<0.005$), as well as between SmartRSVP vs. T-RSVP(67.16 wpm vs. 43.93 wpm, $p<0.005$). The subjective ratings of perceived comfort (on scale of 1 to 5) were 3.78 ($\sigma = 0.73$), 2.06 ($\sigma = 0.96$), and 3.28 ($\sigma = 0.94$) for SmartRSVP, T-RSVP and NWR respectively (Figure 6). All the 18 participants gave positive feedback on the use of eye-gaze as an implicit control channel for RSVP.

The results indicated that by leveraging camera-based visual attention tracking, SmartRSVP was able to overcome the recovery cost of traditional RSVP when a

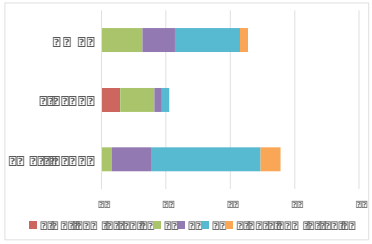

Figure 6. Subjective ratings on perceived comfort on a 5-point Likert scale (1 = not comfortable at all, 5 = very comfortable).

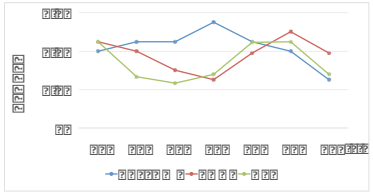

Figure 7. The classifiers' optimal kappa by local window size in user independent models.

user switched visual attention. Although there was no significant difference in comprehension rate or reading efficiency between SmartRSVP and NWR, SmartRSVP could serve as an effective complement to NWR when a user needs to read a message on her watch when both hands are occupied.

User Study 2

We conducted an 18-participant (same participants as user study 1) within-subjects study to investigate the feasibility and performance of predicting cognitive workloads in SmartRSVP through implicit PPG sensing.

The study includes two SmartRSVP conditions: low (single reading task) and high cognitive workload (multi-task includes reading and counting). Each participant read one article under each condition. We adopted the color counting task [20] to induce a high cognitive workload. During reading, a computer placed on the side spoke the names of nine different colors randomly at the speed of one-second per word. Participants were told to focus on the reading but also count the total times that the target colors were spoken. After finishing each article, participants answered 5 questions to test their comprehension.

Results

The users' comprehension test scores were 56.67% and 26.67% for low and high cognitive workload respectively. We were able to use PPG waveforms collected to train a user-independent model to predict a participant's cognitive workload when using SmartRSVP. We used the leave-one-participant-out technique to train the user-independent models of RBFSVM classifier, which led to promising results (window size = 25s, accuracy = 77.5%, precision = 78.3%, recall = 85.0%, kappa=0.55) (Figure 7).

We also ran the cognitive workload prediction (CWP) module on Android devices, which took 63.2ms per estimate on a Nexus 5x using a single core via Java based runtime.

Conclusions and Future Work

We present SmartRSVP, a novel speed-reading system to facilitate text reading on small screen wearable devices. We investigated the performance of the camera-based visual attention control channel of SmartRSVP as well as the feasibility of real-time speed regulation based on implicit PPG sensing.

We plan to further improve SmartRSVP from the following perspectives in the near future: 1) Evaluating the usability and efficacy of SmartRSVP in more realistic environments (e.g. walking and different illumination conditions). 2) Investigating the real-time speed adaptation feature via a formal study. In addition to the current binary speed adaptation mechanism commonly used in today's adaptive interfaces [23], we are also interested in exploring the feasibility of continuous speed adaption via physiological feedback. 3) Exploring the use of alternative output modalities (e.g. vibration, sound) to provide complementary feedback when a user is not paying visual attention to the display; 4) Exploring the use of wrist gestures to assist user's fine-grained control of display speed.

We thank Xiang Xiao, Xiangmin Fan, Phuong Pham, and Shumin Zhai for the constructive feedback. We also thank Byte Dance Telecommunications Co. Ltd. and Lenovo Corp. for the generous support for this project.

References

1. Banitsas, K., Pelegris, P., Orbach, T., Cavouras, D., Sidiropoulos, K. and Kostopoulos, S., 2009, November. A simple algorithm to monitor hr for real time treatment applications. In *Information Technology and Applications in Biomedicine, 2009. ITAB 2009. 9th International Conference on* (pp. 1-5). IEEE.
2. Brysbaert, M. and Nazir, T., 2005. Visual constraints in written word recognition: evidence from the optimal viewing-position effect. *Journal of Research in Reading*, *28*(3), pp.216-228.
3. Castelhano, M.S. and Muter, P., 2001. Optimizing the reading of electronic text using rapid serial visual presentation. *Behaviour & Information Technology*, *20*(4), pp.237-247.
4. Chen, X.A., Grossman, T., Wigdor, D.J. and Fitzmaurice, G., 2014, April. Duet: exploring joint interactions on a smart phone and a smart watch. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 159-168). ACM.
5. Chen, X.A., Grossman, T. and Fitzmaurice, G., 2014, October. Swipeboard: a text entry technique for ultra-small interfaces that supports novice to expert transitions. In *Proceedings of the 27th annual ACM symposium on User interface software and technology* (pp. 615-620). ACM.
6. Dickie, C., Vertegaal, R., Sohn, C. and Cheng, D., 2005, October. eyeLook: using attention to facilitate mobile media consumption. In *Proceedings of the 18th annual ACM symposium on User interface software and technology* (pp. 103-106). ACM.
7. Dingler, T., Rzayev, R., Schwind, V. and Henze, N., 2016, September. RSVP on the go: implicit reading support on smart watches through eye tracking. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers* (pp. 116-119). ACM.
8. Dollár, P., Welinder, P. and Perona, P., 2010, June. Cascaded pose regression. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (pp. 1078-1085). IEEE.
9. Fialding, L.G. and Pearson, P.D., 1994. Synthesis of research reading comprehension: What works. *Educational Leadership*, *51*, pp.62-62.
10. Han, T., Xiao, X., Shi, L., Canny, J. and Wang, J., 2015, April. Balancing accuracy and fun: designing camera based mobile games for implicit heart rate monitoring. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 847-856). ACM.
11. Hansen, J.P., Biermann, F., Madsen, J.A., Jonassen, M., Lund, H., Agustin, J.S. and Sztuk, S., 2015, September. A gaze interactive textual smartwatch interface. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (pp. 839-847). ACM.
12. Harrison, C. and Hudson, S.E., 2009, October. Abracadabra: wireless, high-precision, and unpowered finger input for very small mobile devices. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology* (pp. 121-124). ACM.
13. Jraidi, I., Chaouachi, M. and Frasson, C., 2013, December. A dynamic multimodal approach for assessing learners' interaction experience. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 271-278). ACM.
14. Kanwisher, N.G., 1987. Repetition blindness: Type recognition without token individuation. *Cognition*, *27*(2), pp.117-143.

15. Kim, J., He, J., Lyons, K. and Starner, T., 2007, October. The gesture watch: A wireless contact-free gesture based wrist interface. In *Wearable Computers, 2007 11th IEEE International Symposium on* (pp. 15-22). IEEE.

16. Mariakakis, A., Goel, M., Aumi, M.T.I., Patel, S.N. and Wobbrock, J.O., 2015, April. SwitchBack: Using Focus and Saccade Tracking to Guide Users' Attention for Mobile Task Resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 2953-2962). ACM.

17. Oney, S., Harrison, C., Ogan, A. and Wiese, J., 2013, April. ZoomBoard: a diminutive qwerty soft keyboard using iterative zooming for ultra-small devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2799-2802). ACM.

18. Raymond, J.E., Shapiro, K.L. and Arnell, K.M., 1992. Temporary suppression of visual processing in an RSVP task: An attentional blink?. *Journal of experimental psychology: Human perception and performance*, *18*(3), p.849.

19. Ren, S., Cao, X., Wei, Y. and Sun, J., 2014. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1685-1692).

20. Rodrigue, M., Son, J., Giesbrecht, B., Turk, M. and Höllerer, T., 2015, March. Spatio-temporal detection of divided attention in reading applications using EEG and eye tracking. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 121-125). ACM.

21. Rowe, D.W., Sibert, J. and Irwin, D., 1998, January. Heart rate variability: Indicator of user state as an aid to human-computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 480-487). ACM Press/Addison-Wesley Publishing Co..

22. Smith, B.A., Yin, Q., Feiner, S.K. and Nayar, S.K., 2013, October. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology* (pp. 271-280). ACM.

23. Yuksel, B.F., Oleson, K.B., Harrison, L., Peck, E.M., Afergan, D., Chang, R. and Jacob, R.J., 2016, May. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5372-5384). ACM.