Mobile Multimodal Interaction for Older and Younger Users: Exploring Differences and Similarities

Gianluca Schiavo
FBK Fondazione Bruno Kessler
Via Sommarive, 18
Trento (Italy)
gschiavo@fbk.eu

Ornella Mich
FBK Fondazione Bruno Kessler
Via Sommarive, 18
Trento (Italy)
mich@fbk.eu

Michela Ferron
FBK Fondazione Bruno Kessler
Via Sommarive, 18
Trento (Italy)
ferron@fbk.eu

Nadia Mana
FBK Fondazione Bruno Kessler
Via Sommarive, 18
Trento (Italy)
mana@fbk.eu

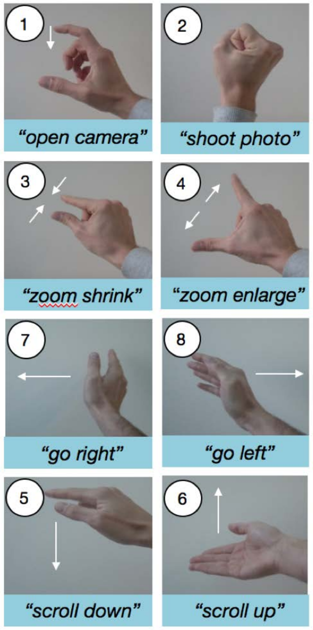
MUM 2017, November 26–29, 2017, Stuttgart, Germany
© 2017 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-5378-6/17/11.
https://doi.org/10.1145/3152832.3156615

Abstract

Since they can integrate a wide range of interactive modalities, multimodal interfaces are considered to improve accessibility for a variety of users, including older adults. However, only few works have actually explored how older adults approach multimodal interaction outside specific contexts and have done so mainly in comparison to much younger users. This study explores how older (65+ years old), middle-aged (55-65 years old) and younger adults (25-35 years old) use mobile multimodal interaction in an everyday activity (i.e. taking photos with a tablet) by using mid-air gestures and voice commands, and investigates the differences and similarities between the considered age groups. Preliminary findings from a video-analysis show that all groups easily combine the proposed modalities when interacting with a tablet device. Furthermore, compared to younger adults, older and middle-aged adults show similarities in the way they perform gesture and voice commands.

Box 1. Combining mid-air gestures and voice commands.

This work investigates multimodal interaction based on the combination of mid-air gestures and voice commands. The following image shows the set of multimodal commands tested in the user study:



The arrows indicate the movement direction. The written words are the voice commands.

Author Keywords

Multimodal interaction; mid-air gesture-based interaction; speech-based interaction; older adults

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

Multimodality and Older Users

Mobile devices are increasingly popular and widespread. However, most of them are originally designed and marketed for the young users, definitely tech savvy and accustomed to touchscreen-based interfaces. For this reason, older people are often bewildered, overwhelmed and frustrated by unnecessary difficulties when using mobile devices. Mobile multimodal interfaces may be a promising approach as they can offer more intuitive and compelling interactive experiences [10,11,16] and the possibility of choosing the interaction channel most suitable to the user capabilities [9]. According to different authors, this is particularly useful for supporting specific user groups, such as older users, who might find difficulties when interacting with touch-based interfaces [8,15]. However, multimodal interfaces have been mainly studied in relation to the younger population.

Only few examples of multimodal interfaces that simultaneously use more than one interaction modality have been specifically designed for and assessed with older users [6,14,19], sometimes comparing their performance with that of much younger people (under 35 years old) [5,9]. Even fewer studies have considered the performance and opinions of middle-aged adults, despite they represent a significant part of the population [17].

This work contributes to the discussion on multimodal interaction with mobile devices focusing on the multimodal integration of mid-air gestures and speech commands. In particular, the paper presents a study in which participants of different ages were asked to use multimodal interaction for taking photographs with a tablet device.

The User Study

The study used a Wizard-of-Oz (WoZ) approach for investigating how participants use multimodal interaction with a tablet device (Samsung Galaxy Tab S2 8.0-inch). In a WoZ experiment, participants use what they think is a fully functioning system, whose missing functions are controlled by a human operator (the “wizard”). Typically, the user is not aware that the wizard is controlling the system, but s/he believes that the system is a functioning autonomous prototype. The WoZ approach has been successfully used in the development and assessment of multimodal interaction [13], even with older adults [12]. This approach has been proved to be effective for engaging older adults in discussions about design ideas through the use of concrete examples and experiences [14].

Our WoZ setup comprised a tablet device, held by the participant (Figure 1), which was controlled by a computer operated by an experimenter (the “wizard”) who was present in the same room.

As most of the studies assessing mid-air gesture interaction with technology for older users focus on physical activity or gaming contexts [3], and our objective was to explore multimodal interaction in an everyday activity with the tablet device, we opted for a commonly known task, such as tablet photography [2].

Box 2. Participants

Participants were from three age groups:

- **Older adults**
  Mean age: 69 (SD= 3.2) (range 65-75)
  7 of them use a mobile phone daily, while 3 on a weekly base.
  Only 1 of them regularly use a tablet device
- **Middle-aged adults**
  Mean age: 51 (SD= 2.9) (range 45-55)
  All of them use a smartphone on a daily base.
  Only 2 of them regularly use a tablet device
- **Young Adults**
  Mean age: 30 (SD= 3.7) (range 25-35)
  All of them use a smartphone on a daily base.
  Only 2 of them regularly use a tablet device

Tablet photography is becoming increasingly widespread and represents a commonly known activity in daily tablet interaction [2]. Moreover, it can be a relatively simple task that can be performed by a variety of users, regardless of participants’ technological familiarity with tablet devices. Specifically, the task consisted of using the tablet camera application to take four pictures of the surrounding environment using a set of multimodal inputs based on mid-air gestures and vocal commands (see Box 1).

Procedure

Thirty participants, ten for each age group (see Box 2), took part in the study. All groups included 5 female and 5 male participants. The study procedure included three different sessions:

1) a training session on ***touch-based interaction***, where the camera application and the common touch were explained. In this way, we could assure that all participants (especially the older adults) were familiar with the basics of the camera application;

2) a training session on ***multimodal touchless interaction***, where participants were introduced to the specific set of multimodal commands: participants were asked to watch videos from the tablet screen showing the 8 multimodal commands (see Box 1) and replicate them;

3) a ***“task” session***, where participants were invited to use the multimodal commands, seen in the previous session, to accomplish the task of taking pictures. In this session, the wizard operated the tablet device and a facilitator guided the participant throughout the study procedure. The task, video and audio recorded, was composed of several sub-tasks suggested to the participants by the facilitator. Each step focused on a specific command (see Box 1) used to: open the camera application (command #1), shot a photo (#2), zoom in and out the scene (#3 and #4), scroll up and down the effect list (#5 and #6) and scroll the picture gallery to the right (#7) or to the left (#8).

After completing the task session, participants received full information on the study procedure and aims.

Results from the video analysis

A total of 1h 35’ video footage of task sessions was analyzed. Each of the 240 observed interactions (8 gestures x 30 participants) was manually annotated, considering:

- **interaction type** (gesture-only, speech-only or multimodal – if both gesture and speech input were performed) and **temporal occurrence** [16]. For temporal occurrence, the interaction was labeled as in parallel if the two modalities were performed with less than 2 sec. delay, otherwise it was considered in sequence;
- **time taken** to complete the task;
- **words** used for the voice commands; and
- **gesture** features (qualitative description of how physical commands were executed).

Data were analyzed using non-parametric methods (except for time of execution); the results are as follows:

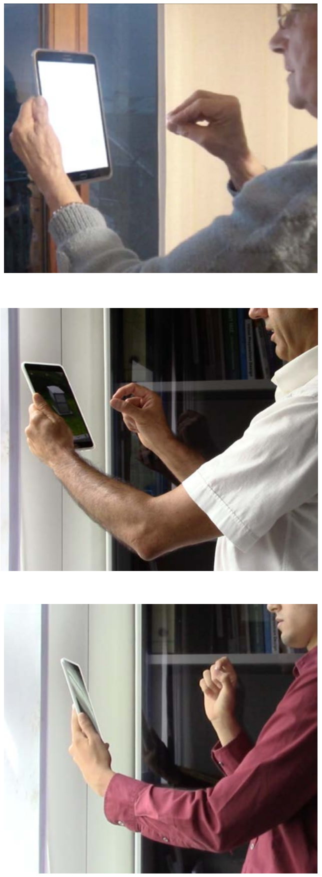


Figure 1. Participants from the three user groups (older, middle-aged and younger adults) interacting with the tablet device by using multimodal interaction.

a) Interaction type and temporal occurrence. Across all participants, the predominant interaction was multimodal commands (Friedman test: χ^2=53.4, *p*<.01; post-hoc with Bonferroni correction: both *p*<.01 - Figure 2 displays the percentages for each age group). No statistically significant differences were observed between groups. Within the multimodal interactions, hand gestures and speech were frequently performed in parallel (χ^2=46, *p*<.01; post-hoc: both *p*<.01 - Table 1). However, older adults showed fewer parallel interactions compared to the other groups (Kruskal-Wallis (K-W) test: H(2)= 11.1, *p*<.01; post-hoc with Dunn- Bonferroni (D-B) comparisons: *p*<.05 older compared to middle-aged, and *p*<.01 older compared to younger adults), and exhibited more gesture-first interactions (K-W: H(2)= 16.4, *p*<.01; D-B: both *p*<.01).

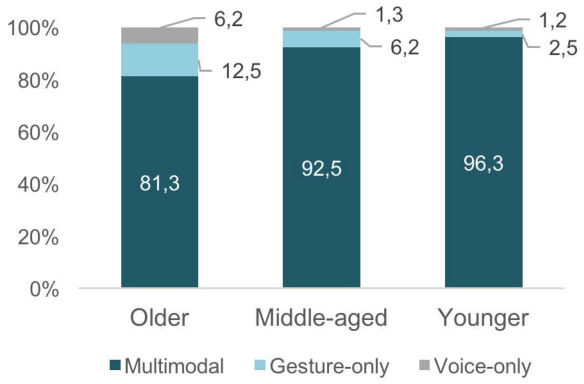


Figure 2. Types of interaction among age groups.

b) Execution time. Execution time was assessed for normality and homoscedasticity. The univariate ANOVA showed that average time for completing the whole task differed between groups (F(2,27)=9.50, *p*<.01). Older and middle-aged adults were slower compared to younger participants (post-hoc comparisons with Bonferroni correction: *p*<.05 and *p*<.01 respectively, Table 1).

c) Voice commands. Each voice command was categorized as:

- **identical**, if the command exactly corresponded to the original one;
- **similar**, in case of variations of the original command which did not change its meaning, such as the use of synonyms (e.g., “shoot/take photo/picture”), articles (e.g., “shot photo/the photo”) or single words (e.g., “zoom” instead of “zoom in”);
- **different**, in case of commands with different meaning from the original (e.g., “scroll vertical” instead of “scroll left”).

Three researchers annotated the audio and video material and a single score was calculated as the statistical mode of the three scores. Inter-rater reliability for voice command annotation was ICC(2,3)=9.75. Overall, participants gave more similar commands compared to identical and different ones (Friedman test: χ^2= 34.6, *p*<.01, post-hoc: *p*<.01). No differences between groups were observed. Qualitative analysis of the most common types of errors showed that older and middle-aged adults used many synonyms instead of using the target word (e.g., “go” instead of “open”, “photo camera” instead of “camera”, “small” instead of “shrink”). We also observed that

	Younger	Middle-aged	Older
Interaction time			
Multimodal	7.7 (0.6)	7.4 (0.8)	6.7 (1.1)
Gesture-only	0.2 (0.4)	0.5 (0.8)	0.9 (1.0)
Voice-only	0.1 (0.3)	0.1 (0.3)	0.4 (0.9)
Temporal occurrence			
In parallel	7.1 (1.2)	6.7 (1.3)	4.5 (1.4)
In sequence (gesture-first)	0.1 (0.3)	0.1 (0.3)	1.9 (1.3)
In sequence (voice-first)	0.5 (1.2)	0.6 (1.0)	0.3 (0.6)
Execution Time	150s (34)	212s (28)	200s (35)

Table 1. Classification of the 240 interactions. Average number of occurrences among interaction type and user group (Mean value, SD in parenthesis).

left/right and up/down were often mixed up, especially among younger participants (only 2 younger participants over 10 correctly used the up/down commands).

d) Gesture features. Mid-air gestures were annotated following the same categories defined for voice commands (where *similar* gestures include small variations in amplitude or in the parts of the hand used – e.g., only one finger or the whole hand). The same three researchers performed the gesture annotation, achieving an inter-rater agreement of ICC(2,3)=9.78. Results are reported in Table 2. Analyses show that more gestures were categorized as identical, compared to the other two categories (Friedman test: χ^2=8.4, *p*<.05; post-hoc: *p*<.05). No differences were found between groups. We observed that the assigned metaphorical gestures (#1 and #2) were rarely remembered and often improvised, especially by older and middle-aged. Conversely, the scrolling and zooming gestures were correctly performed by most participants. Moreover, we observed that a considerable number of gestures was performed moving the index finger instead of the whole hand (resulting in *similar* gestures), mostly among younger participants.

	Younger		Middle-aged		Older	
	Mean (SD)	*%*	*Mean (SD)*	*%*	*Mean (SD)*	*%*
Voice commands						
Identical	2.6 (1.7)	**32.5**	1.4 (1.2)	**17.5**	1.6 (1.7)	**20**
Similar	4.6 (1.6)	**57.5**	5.2 (1.0)	**65**	5.5 (1.5)	**69**
Different	0.8 (0.7)	**10**	1.4 (1.2)	**17.5**	0.9 (0.9)	**11**
Mid-air gesture commands						
Identical	4.8 (2.1)	**60**	4.0 (1.9)	**50**	3.6 (1.2)	**45**
Similar	1.9 (1.7)	**24**	1.8 (1.2)	**22**	1.9 (0.9)	**24**
Different	1.3 (1.1)	**16**	2.2 (1.8)	**28**	2.5 (1.4)	**31**

Table 2. Classification of each voice and gesture commands compared to the reference command.

Discussion

Our findings from the user study point out both differences and similarities between age groups. The results from the video analysis showed that all groups, including older participants, could combine modalities when interacting with the tablet device. However, we observed a tendency in older participants toward performing gestures instead of voice commands. Older adults tended to perform more gesture-only interactions (even though they were instructed to use multimodal commands), and, when using multimodal commands, they gave more gesture-first commands compared to the other groups. This observation suggests a potential advantage of gestures over voice commands: when older participants were unsure about the word to use, they firstly performed the gesture and secondly gave the voice input. This was not observed for middle-aged and younger adults. Furthermore, we also observed that multimodal commands performed by older adults were less synchronous than those performed by the other two groups. In other words, older adults tended to perform more commands in sequence with respect to the other groups. This might be due to mild cognitive impairment related to advancing age, but further investigation would be needed.

Ageing might also be a possible explanation for longer execution times: older and middle-aged adults were slower than younger participants in performing the interaction commands, whereas not significant differences were observed between middle-aged and younger adults. This is in line with previous research on gestural interaction [15].

Regarding voice interaction, younger participants provided a higher number of correct voice commands (using the same commands from the tutorial) compared to the other groups. However, even if older and middle-aged adults often forgot the assigned command, they used meaningful synonyms instead. We noted that one of the most confused voice command was left/right and up/down, especially among younger participants. This might be explained by differences in prior experience with touch-based technology and by familiarity with scrolling interaction [4]. However, further investigations are needed given the limit size of our sample.

Conclusions

Following these findings, our study provides a basis for supporting design choices for multimodal systems based on a combination of mid-air one-hand gestures and voice commands, both for older and middle-aged adults. Older adults perform multimodal interactions as ably as younger users, but they tend to give speech commands after gestures. Moreover, older and middle-aged adults might require more time when giving the multimodal command. Designers should also expect variations in the way arbitrary gestures are performed and in the use of synonyms for voice commands. These variations are to be expected not only from older users but also from middle-aged adults.

Considering further steps in this research line, future work will address a more extensive evaluation of multimodal command recall and learnability, by taking into consideration the interdependencies between the complexity of the interaction set, the task difficulty, and the learnability of mid-air gesture [1], voice [18] and multimodal [19] commands. Moreover, approaches such as the use of image schemas and conceptual metaphors [7] can be evaluated as they could potentially lead to a more intuitive multimodal interaction.

Acknowledgements

This work is supported by the EU HORIZON project ECOMODE - Event-Driven Compressive Vision for Multimodal Interaction with Mobile Devices (http://www.ecomode-project.eu/), under Grant Agreement 644096.

References

1. Fraser Anderson and Walter F. Bischof. 2013. Learning and Performance with Gesture Guides. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1109–1118.

2. Cati Boulanger, Saeideh Bakhshi, Joseph “Jofish” Kaye, and David A. Shamma. 2016. The Design, Perception, and Practice of Tablet Photography. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, ACM, 84–95.

3. Micael Carreira, Karine Lan Hing Ting, Petra Csobanka, and Daniel Gonçalves. 2016. Evaluation of in-air hand gestures interaction for older people. *Universal Access in the Information Society*: 1–20.

4. Jing Chen and Robert W. Proctor. 2013. Response-effect compatibility defines the natural scrolling direction. *Human Factors* 55, 6: 1112–1129.

5. Kathrin M. Gerling, Kristen K. Dergousoff, and Regan L. Mandryk. 2013. Is Movement Better?: Comparing Sedentary and Motion-based Game Controls for Older Adults. *Proceedings of Graphics Interface 2013*, Canadian Information Processing Society, 133–140.

6. Julia Himmelsbach, Markus Garschall, Sebastian Egger, Susanne Steffek, and Manfred Tscheligi. 2015. Enabling Accessibility Through Multimodality?: Interaction Modality Choices of Older Adults. *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, ACM, 195–199.

7. Jörn Hurtienne, Christian Stößel, Christine Sturm, et al. 2010. Physical gestures for abstract concepts: Inclusive design with primary metaphors. *Interacting with Computers* 22, 6: 475–484.

8. Masatomo Kobayashi, Atsushi Hiyama, Takahiro Miura, Chieko Asakawa, Michitaka Hirose, and Tohru Ifukube. 2011. Elderly User Evaluation of Mobile Touchscreen Interactions. *Human-Computer Interaction – INTERACT 2011*, Springer, Berlin, Heidelberg, 83–99.

9. Anja B. Naumann, Ina Wechsung, and Jörn Hurtienne. 2010. Multimodal Interaction: A Suitable Strategy for Including Older Users? *Interact. Comput.* 22, 6: 465–474.

10. Zeljko Obrenovic, Julio Abascal, and Dusan Starcevic. 2007. Universal Accessibility As a Multimodal Design Issue. *Commun. ACM* 50, 5: 83–88.

11. S. Oviatt. 2003. User-centered modeling and evaluation of multimodal interfaces. *Proceedings of the IEEE* 91, 9: 1457–1468.

12. Julie Rico and Stephen Brewster. 2010. Gesture and Voice Prototyping for Early Evaluations of Social Acceptability in Multimodal Interfaces. *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ACM, 16:1–16:9.

13. Daniel Salber and Joëlle Coutaz. 1993. Applying the Wizard of Oz technique to the study of

multimodal systems. In L.J. Bass, J. Gornostaev, and C. Unger, eds., *Human-Computer Interaction*. Springer Berlin Heidelberg, 219–230.

14. Gianluca Schiavo, Michela Ferron, Ornella Mich, and Nadia Mana. 2016. Wizard of Oz Studies with Older Adults: A Methodological Note. *Proceedings of the COOP 2016 - Symposium on challenges and experiences in designing for an ageing society*, 93–100.

15. Christian Stößel and Lucienne Blessing. 2010. Mobile Device Interaction Gestures for Older Users. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, ACM, 793–796.

16. Matthew Turk. 2014. Multimodal interaction: A review. *Pattern Recognition Letters* 36: 189–195.

17. U.S. Census Bureau. *Mid-year Population by Five Year Age Groups and Sex - Europe*. .

18. Linda Wulf, Markus Garschall, Julia Himmelsbach, and Manfred Tscheligi. 2014. Hands Free - Care Free: Elderly People Taking Advantage of Speech-only Interaction. *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, ACM, 203–206.

19. Mary Zajicek and Wesley Morrissey. 2003. Multimodality and interactional differences in older adults. *Universal Access in the Information Society* 2, 2: 125–133.