Accessible Point-and-Tap Interaction for Acquiring Detailed Information about Tactile Graphics and 3D Models

Andrea Narcisi[1][0009−0008−5379−0903], Huiying Shen[2][0000−0002−2195−7085], Dragan Ahmetovic[1][0000−0001−5745−1230], Sergio Mascetti[1][0000−0002−8416−4023], and James M. Coughlan[2][0000−0003−2775−4083]

[1] Department of Computer Science, Università degli studi di Milano, Milan, Italy
andrea.narcisi@studenti.unimi.it, dragan.ahmetovic@unimi.it, sergio.mascetti@unimi.it

[2] Smith-Kettlewell Eye Research Institute, San Francisco, CA USA
hshen@ski.org, coughlan@ski.org

Abstract. We have devised a novel "Point-and-Tap" interface that enables people who are blind or visually impaired (BVI) to easily acquire multiple levels of information about tactile graphics and 3D models. The interface uses an iPhone's depth and color cameras to track the user's hands while they interact with a model. When the user points to a feature of interest on the model with their index finger, the system reads aloud basic information about that feature. For additional information, the user lifts their index finger and taps the feature again. This process can be repeated multiple times to access additional levels of information. For instance, tapping once on a region in a tactile map could trigger the name of the region, with subsequent taps eliciting the population, area, climate, etc. No audio labels are triggered unless the user makes a pointing gesture, which allows the user to explore the model freely with one or both hands. Multiple taps can be used to skip through information levels quickly, with each tap interrupting the current utterance. This allows users to reach the desired level of information more quickly than listening to all levels in sequence. Experiments with six BVI participants demonstrate that the approach is practical, easy to learn and effective.

Keywords: Tactile graphics · 3D models · Blindness · Low vision · Audio labels.

1 State of the Art and Related Technology

Tactile graphics and 3D models are indispensable tools for people who are blind or visually impaired (BVI) to access information [12]. While the shapes and structures of such materials, which include tactile maps, relief maps and educational models (e.g., a molecule, skeleton or biological cell), can be accessed tactilely even without vision, other important information such as color, visual texture and printed information is often inaccessible to BVI users. Braille is often used to label important features on these materials, but there is usually only space for short braille abbreviations (which require a separate braille key defining the abbreviations), and many BVI people don't read braille [11].

A powerful solution to address the inaccessibility of many tactile materials is to make them interactive using a touch-based interface [1]. For instance, the T3 Tactile Tablet[3] from Touch Graphics allows the user to overlay a tactile graphic on a large Android tablet; the device's touchscreen senses finger contact with the graphic and issues a specific audio label for each tactile feature on the graphic that the user touches. In this way BVI users can explore tactile graphics and access semantic information about them whether or not they read braille, using a touch-based interface that is natural in part because of its similarity to the standard touchscreen interface already available on mobile devices.

Computer vision approaches have the potential to make virtually any tactile graphic or 3D model interactive, and are not restricted to 2D graphics that can be overlaid on a tablet touchscreen (and are thin enough for the tablet to sense finger touches). Past work on this approach has been implemented using special cameras with depth sensors [8, 10] or other special hardware[4], standard webcams [9] and as smartphone apps [5, 4, 13]. Some of these approaches have the user point to features by holding a special pointing tool [4], or require the user to augment their fingertip with some sort of visual feature [9]. However, the increasing availability of powerful hand tracking software has enabled the implementation of touch-based interfaces[5] [13, 10]; these natural interfaces are both convenient and facilitate natural exploration of a tactile graphic or 3D model, which is often done using many fingers on one or both hands.

Our research continues the development of natural interfaces, motivated by our experience with the CamIO system [4], in which BVI study participants were enthusiastic about the ability to access audio labels but indicated they would prefer not to have to hold a stylus. Leveraging the availability of high-quality depth sensors in high-end iOS devices, we combine depth and color image data to create a novel "Point-and-Tap" interface that allows BVI users to access detailed information about tactile graphics and 3D models that is organized hierarchically in multiple levels. The benefit of the depth sensor is that it allows our interface to sense when the user is touching the surface of a tactile model as opposed to hovering above it; this determination is key to the interface because it enables the ability to detect touch events. We note that this interface is a significant advance on an earlier interface devised for CamIO [3], in which the user holds the tip of a hand-held stylus to the surface of a feature on a 3D map to hear a basic audio label about the feature, and hovers the stylus tip roughly 15 cm above the feature to hear higher-level information about the feature. By contrast, the "Point-and-Tap" interface is more natural and intuitive, and also supports easy access to several levels of information.

[3] https://www.touchgraphics.com/store/t3-t3-books

[4] https://www.tactonom.com/en/

[5] https://tactileimages.org/en/reader-app/

In experiments with BVI participants we demonstrate that our approach is practical, easy to learn and effective. Finally, while we acknowledge that the advanced depth-sensing iOS devices that enable our approach are expensive, many BVI users enjoy the convenience of a multi-function smartphone and often use smartphone apps to accomplish a variety of daily activities [7], which is arguably more economical and practical than having to purchase multiple devices, each for a specific task.

2 The Point-and-Tap System

Our approach (see[6] for a video demonstration) uses a high-end iOS device with depth sensors (in our case, the iPhone 14 Pro) rigidly mounted above a tactile graphic (Fig. 2) so that the camera and depth sensors capture the entire graphic, as well as some space around the graphic. The space around the graphic is needed to facilitate hand tracking, which is performed using the MediaPipe Hands library[7], and which requires a clear view of the entire hand; note that part of the hand may rest outside the graphic when it points to a feature of interest. An iOS app was written that tracks the user's fingers, and recognizes a pointing gesture (Fig. 2) when it occurs, which is defined as a gesture in which the index finger of a specific hand (left or right, as specified by the user) is pointed straight while the other fingers are bent or closed.

The iOS app passes image frames from the iPhone's color camera stream to MediaPipe for hand tracking, and simultaneously acquires depth images from the iPhone's TrueDepth sensor[8], which specifies the depth estimated across a lower-resolution image having the same field of view as the color (RGB) image. When the app is first launched, reference RGB and depth images are acquired of the scene, which includes the tactile graphic without the user's hands being present. A simple color segmentation algorithm is performed on the reference RGB image to determine which pixels in the image belong to which features of interest (e.g., each region in the British Isles map shown in Fig. 2); for the purposes of this algorithm, we used five colors of cardstock to create the tactile graphics for our experiments. Each feature on a tactile graphic has a unique color, and it is associated with five text labels: the first is the name of the feature, and the four remaining labels specify additional information about the feature. To create a new tactile graphic model, the five colors of cardstock are cut out and pasted to a sheet of white cardstock, and the desired text labels for the entire model are entered as text strings in the app. (See Sec. 5 for a discussion of future work to eliminate our dependence on color cues.)

When a pointing gesture is recognized, the reference depth image is used to determine whether the pointing fingertip is touching the graphic or is above it. The first time the user points to a feature on the tactile graphic, the first text

[6] https://www.ski.org/projects/camio-hands

[7] https://developers.google.com/mediapipe/solutions/vision/hand_landmarker

[8] https://developer.apple.com/documentation/avfoundation/additional_data_capture/streaming_depth_data_from_the_truedepth_camera

label corresponding to that feature is read aloud using text-to-speech (TTS). Each time the fingertip is raised up above a certain height threshold and then lowered again to touch the feature, the text label at the next level is read aloud. (Repeated taps cycle the levels from one through five and then back again to one.) An utterance in progress is halted whenever the fingertip is lifted off the feature, and this behavior allows the user to cycle rapidly through multiple levels to arrive quickly at the level they desire. (This way of eliciting multiple levels of feedback from repeated finger taps was inspired by the user interface in the T3 Talking Tablet.)

To minimize the impact of noise in the depth sensor measurements (see Fig. 1) and in MediaPipe's finger tracking, we devised an algorithm that integrates information over a short time interval. In this algorithm, we kept track of the difference between the reference depth image and the current depth image at the location of the pointing fingertip (MediaPipe's hand landmark L8) over time. This sequence of depth differences is stored in a ring buffer (first-in-first-out) containing values for twenty consecutive camera frames (spanning under half a second of video images); a "touch" event is declared whenever three conditions are satisfied: (a) the average depth difference in the buffer is less than 2 cm; (b) the current depth difference is less than 1.5 cm; and (c) the current depth difference at the location of MediaPipe's hand landmark L5 (the knuckle of the index finger) must be less than 12 cm. The third condition rules out instances when the user's hand is moving rapidly, causing the index finger to disappear from the depth map (even though MediaPipe is still tracking it accurately), thereby making the apparent depth difference at the pointing fingertip close to zero even when the fingertip is nowhere near the surface of the tactile graphic; fortunately, under these conditions, the depth map value at L5 is likely to be accurate since the knuckle is part of the main hand structure (not just the narrow fingertip) and remains visible in the depth map.

Since TrueDepth is intended primarily for face recognition, with a recommended distance from the target of roughly 25-50 cm[9], in our setup we mount the iPhone 14 Pro about 50 cm above the tactile graphic. The limited field of view of the camera imposes a constraint on how large the tactile graphic can be and still be entirely captured in the image (Fig. 2); in the future we will explore the possibility of using iOS LiDAR to accommodate larger models. (While LiDAR has a maximum range of about 5 meters[10], empirically we have found that the minimum range of LiDAR is roughly 70 cm, and we note that it may be challenging to resolve the fingertip clearly in depth at longer distances.)

3 User Studies

We iteratively tested and refined our Point and Tap interface in a series of pilot experiments, the first with a sighted participant and the next with two BVI

[9] https://support.apple.com/en-us/102381

[10] https://www.apple.com/newsroom/2020/03/apple-unveils-new-ipad-pro-with-lidar-scanner-and-trackpad-support-in-ipados/

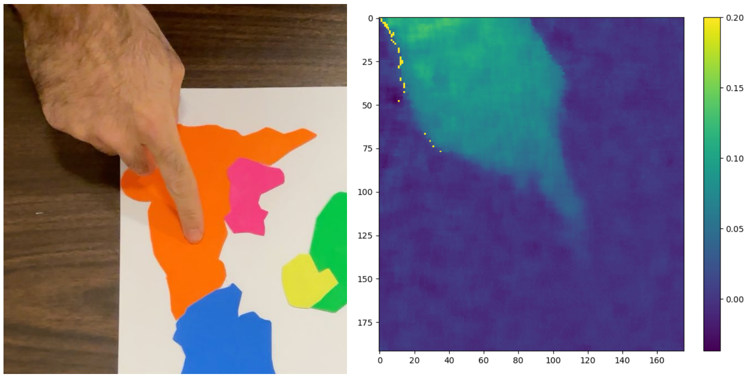

Fig. 1. Left: color image. Right: depth map difference from TrueDepth camera. Color indicates depth differences in meters. The depth difference declines from higher values on the knuckle of the pointing finger to a value at or near zero at the fingertip, but depth noise is clearly visible.

participants. We used the feedback from these pilot experiments to iteratively debug and improve our interface. Notably, we relaxed our definition of the pointing gesture so that it admitted a pointing gesture in which the non-index fingers are not curled as tightly as shown in Fig. 2. The interface and experimental protocol were fixed after these pilot experiments. We then conducted formal experiments with six more BVI participants, P1-P6, ranging in age from 36 to 74 years old (3 female/2 male); we excluded a seventh BVI participant from our analysis because of inadvertent deviations from the experimental protocol. All experiments were conducted under an approved IRB protocol.

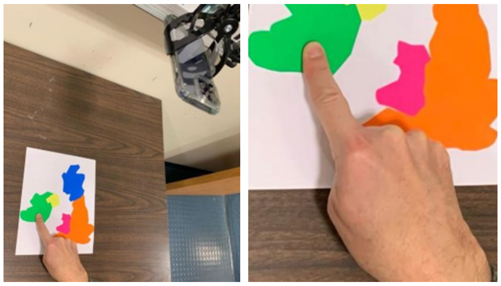

Fig. 2. Left: Experimental setup shows iPhone 14 Pro mounted above the British Isles tactile map in cellphone holder with the user making a pointing gesture on the map. Right: close-up of pointing gesture. The five levels of text labels for the region being pointed to are: “Republic of Ireland”; “2: Capital is Dublin”; “3: Population 4.9 million”, “4: Area 27133 square miles”, “5: Limerick is another city in Ireland”.

Each formal experiment included the following: (1) A demographic survey. (2) A 15-minute training session in which the experimenter demonstrated how to use the interface on a simple tactile graphic containing four features (a circle, square, triangle and diamond), and helped the participant learn to use the system. (3) Experiment E1, in which the participant was presented with a tactile map of the sun with the inner planets (Mercury, Venus, Earth and Mars) and was asked to access specific information (e.g., "Mars, level 2") in 50 trials. For each trial, the experimenter gave two scores for each of the participant's responses to evaluate the following: (a) E1a: Did the system do what the user wanted it to do? (b) E1b: Was the user able to elicit the correct information by the end of the trial? (4) Experiment E2, in which the participant was asked to access information that answers high-level questions about two additional tactile models, one containing five rockets and the other a map of the five regions of the British Isles (England, Scotland, Ireland, Northern Ireland and Wales). (5) A System Usability Survey (SUS)[2]. (6) Both specific and open-ended feedback about the system.

4 Experimental Results and Discussion

We report the results of the study in Table 1, with E1a and E1b scores reported as percentages out of 50 trials. We note an important caveat for the data for participant P6: during Experiment E2, despite good performance on the five rockets model, shortly after beginning to work with the British Isles model, she was unable to elicit any feedback from the system because she forgot to make the correct pointing gesture (Fig. 2), unsuccessfully attempting to point while holding her non-index fingers nearly straight. After witnessing this problem for several minutes, the experimenters intervened by reminding her how to make the pointing gesture. Her performance returned to normal after this reminder.

The E1 scores demonstrate that, while the system sometimes (the average E1a score was 86.3%) misinterpreted participants' tap gestures (most often because the user failed to lift their fingertip up high enough above the model before tapping it again, or because they failed to make the correct pointing gesture), in all but one trial for P5 the participants were nevertheless able to get the desired information. The SUS scores demonstrate the usability of the system, with an overall average score of 88.3.

The results of experiment E2 provide additional evidence of the effectiveness of the interface: participants were able to answer all the questions correctly. Positive feedback about the interface included its simplicity and usefulness for acquiring detailed information about a model; some negative feedback was reported about the incidence of errors when the interface failed to recognize a tap gesture. Some participants suggested that the interface should be expanded to recognize additional gestures, such as left/right swipe gestures to jump down/up several levels of audio labels.

Overall, the experiments with BVI participants demonstrate that the approach is practical, easy to learn and effective. Our experiences with the experiments highlight two main areas of improvement. First, the training session should have included a component in which the participants were asked to practice more Point-and-Tap tasks, with the experimenters making specific recommendations based on their performance. Second, the system could have provided periodic (e.g., no more than once every several seconds) audio feedback whenever the pointing fingertip was stationary but no pointing gesture was recognized, which would remind the user to make the pointing gesture if they want feedback.

Table 1. Data for all participants in formal experiments. (See the text for an important caveat about participant P6.) The choices for self-reported Experience scales included None, Low, Medium, High.

P#	Age	Sex	Perception of		Experience with		E1a	E1b	SUS
			light	form	tactile graphics	touchscreens			
P1	46	F	N	N	High	High	82%	100%	75
P2	43	M	N	N	Medium	High	92%	100%	97.5
P3	74	F	Y	Y	Low	Medium	86%	98%	92.5
P4	74	F	N	N	Low	High	76%	100%	77.5
P5	36	M	Y	N	Medium	High	98%	100%	90
P6	62	F	N	N	Low	Medium	84%	100%	97.5

5 Conclusion

We have devised a novel "Point-and-Tap" interface that allows BVI users to acquire detailed information about tactile graphics and other models. User studies with six BVI participants demonstrate that the interface is practical, easy to learn and effective.

We have recently improved the system so that it recognizes the pointing gesture with either the left or right index finger, so the user doesn't need to specify which hand they are pointing with; an audio alert is sounded if the system detects the pointing gesture in both hands. We plan to release the system as a free app on the Apple App Store which will allow non-developers to create and label their own models using colored cardstock. Soon we will add 3D geometric reasoning to our system as used in CamIO, which will eliminate the need for color segmentation, and will allow the app to support both 2D and 3D models, including existing tactile graphics and 3D models. Finally, in the future we will explore methods such as those used in [6] to allow our system to more precisely distinguish between fingertip touch and hover events.

Acknowledgments HS and JMC were supported by NEI/NIH grant number 2R01EY025332 and NIDILRR RERC grant number 90REGE0018.

Disclosure of Interests The authors have no competing interests.

Bibliography

[1] Brock, A.M., Truillet, P., Oriola, B., Picard, D., Jouffrais, C.: Interactivity improves usability of geographic maps for visually impaired people. Human–Computer Interaction (2015)

[2] Brooke, J., et al.: Sus-a quick and dirty usability scale. Usability evaluation in industry (1996)

[3] Coughlan, J.M., Biggs, B., Shen, H.: Non-visual access to an interactive 3d map. In: International Conference on Computers Helping People with Special Needs. Springer (2022)

[4] Coughlan, J.M., Shen, H., Biggs, B.: Towards accessible audio labeling of 3d objects. In: Journal on technology and persons with disabilities: Annual International Technology and Persons with Disabilities Conference. CSUN (2020)

[5] Fusco, G., Morash, V.S.: The tactile graphics helper: providing audio clarification for tactile graphics using machine vision. In: International ACM SIGACCESS Conference on Computers & Accessibility. ACM (2015)

[6] Goussies, N.A., Hata, K., Prabhakara, S., Amit, A., Aube, T., Cepress, C., Chang, D., Cheng, L.T., Ciurdar, H.S., Cleron, M., et al.: Learning to detect touches on cluttered tables. arXiv preprint arXiv:2304.04687 (2023)

[7] Griffin-Shirley, N., Banda, D.R., Ajuwon, P.M., Cheon, J., Lee, J., Park, H.R., Lyngdoh, S.N.: A survey on the use of mobile applications for people who are visually impaired. Journal of Visual Impairment & Blindness

[8] Shen, H., Edwards, O., Miele, J., Coughlan, J.M.: Camio: a 3d computer vision system enabling audio/haptic interaction with physical objects by blind users. In: International ACM SIGACCESS Conference on Computers & Accessibility. ACM (2013)

[9] Shi, L., Zhao, Y., Azenkot, S.: Markit and talkit: a low-barrier toolkit to augment 3d printed models with audio annotations. In: Annual ACM symposium on User Interface Software and Technology. ACM (2017)

[10] Wang, X., Kayukawa, S., Takagi, H., Asakawa, C.: Touchpilot: Designing a guidance system that assists blind people in learning complex 3d structures. In: International ACM SIGACCESS Conference on Computers and Accessibility. ACM (2023)

[11] Wiazowski, J.: Can braille be revived? a possible impact of high-end braille and mainstream technology on the revival of tactile literacy medium. Assistive Technology (2014)

[12] Zebehazy, K.T., Wilton, A.P.: Straight from the source: Perceptions of students with visual impairments about graphic use. Journal of Visual Impairment & Blindness (2014)

[13] Zeinullin, M., Hersh, M.: Tactile audio responsive intelligent system. IEEE Access (2022)