Pictures in Your Mind: Using Interactive Gesture-Controlled Reliefs to Explore Art

ANDREAS REICHINGER, VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH
HELENA GARCIA CARRIZOSA, Open University
JOANNA WOOD, University of Sussex
SVENJA SCHRÖDER and CHRISTIAN LÖW, University of Vienna
LAURA ROSALIA LUIDOLT, MARIA SCHIMKOWITSCH, ANTON FUHRMANN,
STEFAN MAIERHOFER, and WERNER PURGATHOFER, VRVis Forschungs-GmbH

Tactile reliefs offer many benefits over the more classic raised line drawings or tactile diagrams, as depth, 3D shape, and surface textures are directly perceivable. Although often created for blind and visually impaired (BVI) people, a wider range of people may benefit from such multimodal material. However, some reliefs are still difficult to understand without proper guidance or accompanying verbal descriptions, hindering autonomous exploration.

In this work, we present a gesture-controlled interactive audio guide (IAG) based on recent low-cost depth cameras that can be operated directly with the hands on relief surfaces during tactile exploration. The interactively explorable, location-dependent verbal and captioned descriptions promise rapid tactile accessibility to 2.5D spatial information in a home or education setting, to online resources, or as a kiosk installation at public places.

We present a working prototype, discuss design decisions, and present the results of two evaluation studies: the first with 13 BVI test users and the second follow-up study with 14 test users across a wide range of people with differences and difficulties associated with perception, memory, cognition, and communication. The participant-led research method of this latter study prompted new, significant and innovative developments.

CCS Concepts: • **Human-centered computing** → **User studies**; **Gestural input**; **Accessibility systems and tools**; *Auditory feedback*; • **Applied computing** → Fine arts;

Additional Key Words and Phrases: Blind, low vision, learning disability, cognitive disability, auditory interface, design for all, gestures, multimodal interaction

This work was performed within the framework of the project Deep Pictures, supported by the Austrian Science Fund (FWF): P24352-N23, the Erasmus+ project AMBAVis (http://www.ambavis.eu) which has been funded with support from the European Commission under Grant No.: 2014-1-AT01-KA204-001014, and the H2020 project ARCHES (http://www.arches-project.eu), which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 693229. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Authors' addresses: A. Reichinger, L. R. Luidolt, M. Schimkowitsch, A. Fuhrmann, S. Maierhofer, and W. Purgathofer, VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Donau-City-Straße 11, 1220 Wien, Austria; H. G. Carrizosa, Open University, Level 3, Stuart Hall Building, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom; J. Wood, University of Sussex, School of Global Studies, Sussex House, Falmer, Brighton, BN1 9SJ, United Kingdom; S. Schröder and C. Löw, University of Vienna, Faculty of Computer Science, Research Group Cooperative Systems, Währinger Straße 29, 1090 Wien, Austria.

ACM 1936-7228/2018/03-ART2 $15.00
https://doi.org/10.1145/3155286

ACM Reference format:
Andreas Reichinger, Helena Garcia Carrizosa, Joanna Wood, Svenja Schröder, Christian Löw, Laura Rosalia Luidolt, Maria Schimkowitsch, Anton Fuhrmann, Stefan Maierhofer, and Werner Purgathofer. 2018. Pictures in Your Mind: Using Interactive Gesture-Controlled Reliefs to Explore Art. *ACM Trans. Access. Comput.* 11, 1, Article 2 (March 2018), 39 pages.
https://doi.org/10.1145/3155286

1 INTRODUCTION

Multi-modal learning materials—tactile, auditory and visual—are widely used among those with a range of learning, sensory, and cognitive impairments to help them understand graphic content that is otherwise difficult to convey. For instance, blind and visually impaired (BVI) visitors traditionally explore museums using audio guides, live audio described tours, large-print guides, Braille gallery guides, Braille captions, and touch objects. Tactile objects may be categorized according to the taxonomy in [44] into:

- two-dimensional (*2D*) objects [1, 13], such as tactile diagrams, line drawings or plans – e.g., on embossed paper, swell paper, and increasingly also with vibro-tactile cues [26, 38];
- fully *3D* objects [39, 44, 52, 57], such as anatomical models, 3D-printed reproductions, or everyday objects; and
- the *2.5D* realm in-between, i.e., "height fields, [or relief] surfaces that can be represented by a function $z = f(x, y)$, giving every point above a plane a single height value" [44].

The last group, tactile relief, is especially useful in increasing access to the visual arts of images, photos, and paintings, as it keeps the connection to the two-dimensional original[1], while the plasticity of the added height makes it easier to recognize by touch. Depicted shapes can be geometrically formed in bas-relief and painted textures can be made tactile as surface variations. The importance of this technique is demonstrated by the increasing number of art shows all over the world incorporating tactile reliefs[2] and technical developments [15, 16, 43] in order to create such reliefs.

Though primarily created for BVI people, kinesthetic learning has been found to be important for people who struggle with logical and analytical thinking, such as those with learning and cognitive differences and disabilities. In addition, such multisensory material can benefit everyone [29]. This becomes obvious when we observe general visitors approaching tactile pieces in a museum: even though these were primarily made for BVI people, many visitors touch them, enjoy them, and probably spend more time and connect deeper than with conventional exhibits. The more sensory modes that are offered, the more people are targeted. Consequently, we can do better than just offering additional pieces for the tactile sense, especially for those who cannot fully perceive or understand the originals.

While tactile material is good at conveying spatial cues, many aspects are difficult to mediate by touch alone. As stressed throughout the literature (e.g., [14]), verbal description is a very important

[1] By precisely recreating the painting in relief, it is pixel by pixel the same when viewed from above, which cannot be achieved in full 3D works in which backs and sides would need to be "hallucinated" and the overall impression would significantly change.

[2] Examples for art shows that incorporated tactile reliefs include The "Museo Tattile di Pittura Antica e Moderna Anteros" of the Istituto F. Cavazza in Bolognia [20] (http://www.cavazza.it/?q=node/315); Irish Museum of Modern Art: "Altered Images," 2010 (http://topografik.co.uk/altered-images-index); Berlinische Galerie: "Wien Berlin," 2014 [6]; Prado: "Touching the Prado," 2015 (https://www.museodelprado.es/en/whats-on/exhibition/touching-the-prado/29c8c453-ac66-4102-88bd-e6e1d5036ffa); Canadian Museum for Human Rights: "Sight Unseen," 2016 (https://humanrights.ca/exhibit/sight-unseen).

part, especially for artwork. A painting is typically composed of several parts, all with their own appearance, colors, and properties, and with relationships to each other. All of this is hard to encode into a single tactile image but can easily be described verbally.

On the other hand, a single monolithic audio script, e.g., as present in most museum audio guides, may not be satisfactory either. While most people appreciate a top-level introduction to orient themselves, detailed descriptions are better delivered on demand. Individuals are likely to be interested in different details and thus prefer to request information when they reach an interesting region, rather than receiving it in a predefined order and having to locate the right part of the relief.

With these current limitations in mind, we focused on developing a gesture-based interactive audio guide and potentially a multimedia guide. We envision it as an extensible system capable of enriching tactile objects with autonomously navigable information in a variety of sensory forms, thus enabling more people to experience objects that are at present inaccessible to them.

1.1 Requirements

The aim of this work is to create a system that enables people with a wide range of access needs to explore tactile materials in a more autonomous way. We are keen to engage as many of the sensory channels as possible in order to make it accessible for the greatest number of people. Though this work originally stems from design responses to the very specific needs of BVI people, our aim has evolved into a general one of universal design and access for everyone. Tactile and multisensory experiences increase the engagement and learning of everyone, not just those with specific access needs.

Motivated by project partners, our approach is mainly targeted at a museum setting with tactile reliefs of paintings, but the results may readily be used in a wider context: for instance, in schools or at home. This approach would work not only with paintings but with all kinds of flat objects, tactile diagrams, and to a certain extent three-dimensional objects, although currently only from a single side. We therefore envision a largely self-contained system in the form of a kiosk or installation that fits into a museum space. In order to keep the system maintainable, it should run on off-the-shelf, easily exchangeable, low-cost hardware. Custom software algorithms should not depend on specialized hardware to simplify adaptation to different architectures. The same setup should be usable with several tactile objects, one at a time. The content for each object should be easily adaptable, flexible enough to add and change interaction locations, descriptions, and interaction modes. The interface should be simple, easy to use, self-explanatory, and robust to a wide variety of users. Although the first prototype was targeted at BVI people, according to a design-for-all philosophy, the system has potential for children, older people, those with cognitive and learning impairments, and the general audience, possibly all interacting with different sensory modes.

Based on our original discussions with BVI people and reinforced by wider testing across the access needs spectrum, the main goal of our system is to allow users an *undisturbed* exploration, without unwanted explanations and with precise control over when to get information and what that information regards. The user should be able to explore the relief with one or both hands without triggering unwanted (audio) content and avoid Midas touch effects [22, p. 156]. This means that only very distinct gestures should trigger (especially) audio comments, gestures that normally do not occur during tactile exploration and that can reliably be detected. This is in contrast to systems with embedded sensors (see Section 2), that are triggered by any kind of touch, whether intentional or not.

1.2 Contributions Beyond the Conference Version

This work is an extended version of a paper presented at the ASSETS'16 conference [42].

The original version was mainly focused on people who are blind or visually impaired. Comments and observations during test sessions and public presentations made it clear that people with other access needs, and even those without access needs, might benefit from such a system. To this end, we carried out a second test study with participants across a wide range of access needs relating to sensory, learning, and cognitive impairments to test the wider applicability of the IAG. This study was divided into two testing sessions so that immediate changes could be made based on the feedback from the first session and tested in the second. The initial results of this second study, along with a broader view and analysis of the topic, are included here. We used a different, participant-led, research method [3, 45, 53] in the second study to better test the potentials and limits of the IAG and to gain a deeper insight. This approach led to immediate developments in the IAG and opened up new channels along which the IAG can now evolve.

In addition, we switched from the ad-hoc setup, using a tripod mounted camera, to a more professional and museum-ready setup based on the HP Sprout workstation. We discuss its possible benefits and required changes to the algorithms to cope with the different camera placement. Other additions include new audio-visual cues to help users in performing the correct gestures, first results of an automatic relief detection and calibration method, and more details on the tactile relief creation process.

2 RELATED WORK

A large body of work concentrates on augmentation of 2D graphics. The *Talking Tactile Tablet* [28] and *ViewPlus's IVEO* [18] detects touch gestures on tactile diagrams put on a high-resolution touch pad. This technology clearly cannot be directly used with relief surfaces of significant height.

This problem is circumvented—e.g., in LucentMaps [19]—by including conductive elements and by Taylor et al. [50] by using conductive filaments, both for 3D prints of maps so that touch information is transmitted down to a touch screen of mobile devices. This, however, narrows the possible materials and limits the size to available touch screens.

Several projects utilize color cameras to track the user's fingertips: *Access Lens* [23] recognizes and reads texts on documents pointed to by a finger. The *Tactile Graphics Helper* [17] plays prerecorded audio when the finger is over predefined labels and is triggered by voice commands. *Tactile Graphics with a Voice* [2] is an app for smartphones and Google Glass that reads labels indicated by QR codes. Similar in spirit is the commercial product ORCam[3], a glasses-mounted camera, that detects text touched by a single extended, upward-facing finger, and reads this text using OCR technology. THATS[4] uses a silhouette-based hand detection and gesture recognition approach to detect single-finger pointing gestures in a mobile device's camera stream in order to trigger prerecorded location-based audio on tactile prints that are augmented with tracking patterns. The current prototype only works with a single map with three widely spaced active areas, and the tracking patterns in the four corners must not be covered with the hands. Finally, *Kin'touch* [7] studies the combination of optical finger tracking and touch events from a capacitive multitouch screen. While these approaches focus on 2D documents, some could probably be extended to the third dimension. However, most require labels that we want to avoid and tracking based on color alone is error prone, as it is dependent on skin color, background color, and lighting conditions.

[3] ORCam's product website is at http://www.orcam.com and an evaluation study was recently published [32].

[4] THATS (Touch & Hear Assistive Teaching System) is a *free* app project for mobile devices currently available as a prototype for IOS at http://thats.wiki/.

Talking Pen Devices[5] detect barely visible printed patterns and take a somewhat special role: although originally intended for printed documents, they are usable on 3D objects by applying stickers with the detectable pattern. However, stickers affect the tactile quality and wear off.

Several full 3D approaches are based on devices integrated *into* the tactile object. For instance, *Tooteko* [10] integrates NFC Tags in 3D models, which are read by a wearable NFC reader. *Digital Touch replicas* [55] have touch sensors integrated at interesting locations. Most recently, *3DPhotoWorks*[6] managed to print the color images directly on the relief surfaces [37] and integrated touch-sensitive infrared (IR) sensors into its reliefs. While this is a robust solution for a museum setting, these approaches are less flexible. Once placed, trigger regions can no longer be changed and probably cannot be reused on other objects. Only discrete trigger locations are possible and interaction modes that require fine-grained touch positions are not possible. Furthermore, the sensors react to any kind of touch, which conflicts with tactile examination by BVI people (see Midas touch).

As an alternative, 3D scanners can be used as input devices. For example, the SandScape project [40] allows the user to model a relief in sand, which is scanned in short intervals, interpreted as terrain, and simulation data is then projected onto the sand. However, the input here was more the changing sand surface than the user's hand. Probably for the first time, Wilson [54] introduced the concept of using a depth camera as a *touch sensor* on nonflat surfaces. *CamIO* [47] extended the concept to touch interaction on 3D objects targeted at blind users. A proof of concept implementation was given with at least two different labels on an object, which could even be rotated. Magic Touch [48] offers a similar installation but works with a standard color camera. 3D objects can be freely rotated and are tracked by a cube with trackable markers attached to the object, and single-finger pointing events are inferred from the finger colored with tape. At least a few regions can be labeled per object. The approach that is probably the most similar is a feasibility study [8], which uses a Microsoft Kinect with the CVRL FORTH Hand Tracker [36] to trigger audio by touch events of the right index finger's tip on tactile reliefs. Little is reported about real-world experiences by the target group. Only the limited robustness of the tracking system is mentioned.

In contrast, our system is built around a custom hand-detection algorithm that is very stable, as it works independently on each frame. A carefully selected set of gestures already allows multiple actions and was evaluated in a user study. The theoretical concept of our system was first presented in [41] and includes a review of current depth sensors.

3 INTERACTIVE AUDIO GUIDE (IAG)

The gesture-controlled IAG consists of a depth camera (currently an Intel RealSense F200) as the only sensor, connected to a computer and rigidly mounted above a tactile relief, which it observes. Two different setups can be seen in Figures 1(c) and 3(a).

In contrast to conventional color cameras, which give an RGB color value for each pixel, a depth camera (or RGB-D camera) also returns a depth value, i.e., how far an object at this pixel is away from the camera. First, the system is initialized with only the relief present. The system stores the acquired depth image, the so-called *background image*. Whatever is now put on top of the relief creates depth measurements that are nearer to the camera, hence can be easily detected by comparing the current depth image and the background image. This process is called foreground segmentation (see Section 3.5.1), and creates a *foreground mask*, a set of pixels where all these added

[5] Multiple vendors offer talking pens, such as the TalkingPEN (http://www.talkingpen.co.uk), Touch Graphic's Talking Tactile Pen (http://www.touchgraphics.com), Livescribe (http://www.edlivescribe.com), or Ravensburger tiptoi (http://www.tiptoi.com). While the last is developed as a toy, there is a hacking community dedicated to opening this low-cost device for arbitrary content (http://tttool.entropia.de/).

[6] 3DPhotoWorks website is at http://www.3dphotoworks.com.

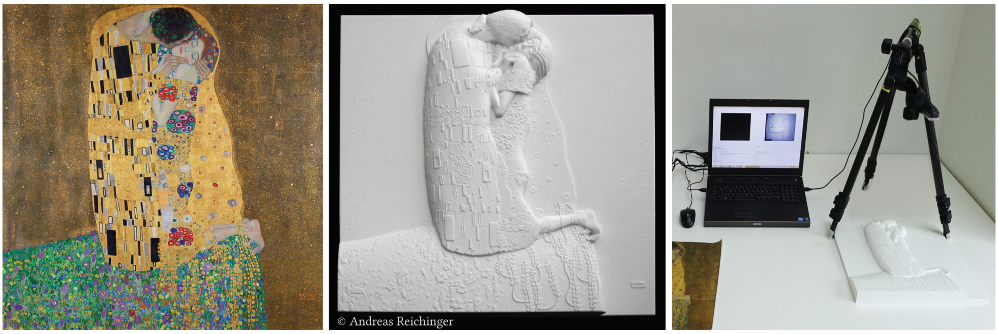


Fig. 1. Gustav Klimt's "The Kiss" (Der Kuss, 1908/09). From left to right: (a) Original image, © Belvedere, Wien. (b) Tactile relief interpretation, © Andreas Reichinger. Size: 42 × 42 cm, 10 mm base height + up to 25 mm variable relief height. Material: DuPont Corian®, Glacier White. (c) First interactive audio guide (IAG) setup using a tripod, © Andreas Reichinger.

things are located. As any objects may be added, the foreground is carefully searched for hands and whether these hands form certain input gestures (see Section 3.5.3–3.5.6). Finally, depending on the gestures, real-time audio (and potentially other) feedback is given to the user.

The use of a depth camera has multiple advantages over a conventional color camera: it is largely independent from the (visible) lighting situation, working even in complete darkness, as it has its own—for humans—invisible lighting. In contrast to color images, depth information is more reliable and independent from relief and skin color, thus allowing colored reliefs or projected images onto reliefs and hands. Even gloves may be worn as long as they reflect IR light used by the sensor. Depth further allows detection of touch-events to trigger interaction, whereas systems using color cameras have to use other triggers, e.g., voice commands, active finger gestures, or buttons pressed with the other hand or with the feet. It is more flexible than approaches with integrated sensors, works on arbitrary 3D surfaces, and allows gestures "beyond touch" [54]. Depth cameras are currently low-cost, off-the-shelf technology that is estimated to be soon integrated into laptops (already available) and mobile devices. This makes the system also attractive for home use in the future.

However, depth sensors have their own set of limitations, as they are dependent on a clear detection of the projected IR light. Although they typically use a very narrow-banded LASER illuminator and a narrow band pass filter on the camera optics, they are sensitive to high levels of ambient IR light in exactly this band, for instance, sunlight. In a controllable indoor environment—such as museums and office buildings—this can be excluded as either no windows are present or window shades can be closed, and most light sources do not have a significant IR component. Nevertheless, one museum had incandescent spotlights installed with an IR component that degraded tracking performance to some extent; thus, we had it dimmed. In addition, the scanned objects need to reflect the IR light diffusely, which makes objects with shiny or very dark (in the IR band) objects difficult to scan. Human skin and the relief materials that we tested worked well.

In the remainder of this article, we will explain the development of our prototype, detail our design decisions, and conclude with the results of the user evaluation.

3.1 Prototype

In order to test the proposed system, we developed a prototype for the interactive exploration of a tactile relief interpretation of Gustav Klimt's painting "The Kiss" (1908/09; see Figure 1(a)). This popular painting was chosen because many people have already heard about it; but, before we started, only descriptions and simple raised line diagrams were available as accessible tools.

In cooperation with experts on art history, regions of the painting have been labeled, named, and short texts (20 seconds, on average) have been recorded containing descriptions of the region, color composition, body poses, and relations between parts. The image was divided into 6 basic regions (e.g., background, meadow, male and female figure), and the two figures were further subdivided for a total of 20 different labels of varying sizes (see Figure 8(a)). In addition, five short general texts (50–60 seconds each) about the painting, its history, interpretations, and the artist have been recorded. All texts are available in German and English (switchable in the graphical user interface). The texts were voiced by a German native speaker; hence, the English version might not be optimally voiced.

3.2 Relief Design

The relief (see Figure 1(b)) was created based on our previously developed design workflow [43]. Due to the nature of the selected painting (partly plastic figural, partly abstract decorative) this workflow had to be refined. The visible body parts are painted in a very natural way. In the relief, we wanted these parts as naturalistic as possible in order to allow a good tactile recognition. The rest of the painting—especially the worn clothes—appear very flat and two-dimensional, as they are dominated by abstract patterns and no shading or other visual cues are present that would give a three-dimensional appearance.

Our first intention was that these parts need to be as flat as possible. The background and the surrounding halo can be naturally thought to be flat, as the one is very far away and the other is an optical illusion. For the clothes, this was not meaningful, however, as the visible body parts interact with each other—for instance, the male figure kissing and embracing the female figure and their hands holding each other. With flat clothes, it would not be possible to convey these interactions properly. We therefore decided to let the clothes form naturally over their bodies, but as flat and smooth as possible.

Similarly, the meadow appears flat but requires to be bent to allow a smooth transition to the sky and to provide a gradient on which the figures can be placed. The bent form of the right edge of the meadow could suggest a hill or cliff, however, which came naturally together with the rest of the relief.

In order to recreate the bulges of the clothes and the visible body parts, the body poses, hidden behind the clothes, needed to be consistently recovered. The female obviously kneels as we see her feet, but the male could have a variety of poses. Indeed, it is a controversy regarding whether he is rather short and stands or if he also kneels or otherwise stands in a bent pose.

Especially difficult for us to interpret was the large unshaded part of his neck and shoulders. It is often said that the painting could be a self-portrait of Gustav Klimt himself with his life-partner Emilie Flöge. While researching photographs of Klimt, we found good evidence (e.g., [51]) that he indeed had a very strong shoulder and neck part and that, if he bent over, it looked quite similar to the painting. In order to recover the body poses, we used digital mannequins in a 3D editing program and iteratively adjusted them until they exactly resembled the pose in the painting (see Figure 2(a)). Nevertheless, the found pose is just one of many possible poses and does not say anything about the intent of the artist. However, the consortium was satisfied with the found pose for the relief interpretation, although through lots of adjustments, the figures ended up with legs that were too short and upper bodies that were too large. But, since these parts are not visible in the final relief, it was not deemed important.

We then modeled the clothes over the naked bodies and the meadow hill under their feet using Bézier patches (see Figure 2(b)). To achieve an accurate resemblance, depth maps of the individual parts of the 3D scene were rendered as depth maps, then warped, cut, and composited to exactly align with outlines extracted from the painting.

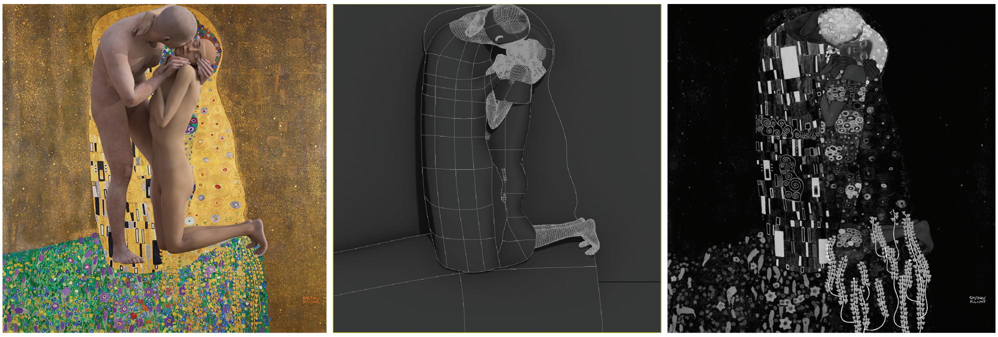


Fig. 2. Stages of relief development (images © Andreas Reichinger, original image © Belvedere, Wien). From left to right: (a) Study with digital mannequins to recreate the body poses; (b) Modeled meadow/cliff, and modeled clothes and hair over the figures using Bézier patches; (c) Final texture layer extracted from the original painting and manual enhancements.

After discussions with experts, minor revisions, and optimization of the depth composition to maximize the available depth at important parts (e.g., face, feet, hands), we started to extract the texture and add it as additional height variations over the base relief. In most parts, we could extract a meaningful texture layer directly from a gray-scale version of the original painting (see Figure 2(c)) with various filters. Several important parts needed to be improved. These were segmented either manually or color based and included as corrections on the texture layer. These parts include the spirals on the male's coat, the different kinds of flowers on the meadow and in her hair, the wreath in his hair, the signature, and the tendrils on the right part of the meadow. At the tendrils, we made the stem lines slightly higher than the triangular leaves so that the finger is easily guided along the individual stems while feeling the leaves underneath, but without being disturbed by them. The texture of the background was diminished, as it was deemed less important and should not distract too much from the more important parts.

After approval of a previsualization, the relief was milled out of a solid block of DuPont Corian® (color Glacier White). This material was chosen because it is very hard but still easy to mill. It is smooth to the touch, lets the finger glide easily over the surface, and can be cleaned and disinfected as it is normally used for countertops, bathrooms, sinks, furniture, and even façades.

We chose a size of 42 × 42 cm for convenience, large enough to feel most details, small enough to be easily reached, and not too expensive in the production. Further, a base thickness of 10 mm was chosen for stability and an additional 25 mm gives enough space for plastic variations in the relief surface. CNC-milling was performed on a Datron M8 Cube, finished with a 2 mm ball nose tool and 0.2 mm stepover in 45° diagonal paths (lower left to upper right). The final result has a very detailed but smooth surface (see Figure 1(b)).

Plaster copies were made via a silicone mold to facilitate later testing, as the Corian® original is now permanently exhibited in the gallery and milling further copies was deemed too expensive. These copies feature a similar surface but are slightly rougher, making the finger slip less easy, and on a few places some very thin details are missing, such as small parts of the stems of the tendrils, probably caused during demolding.

3.3 Hardware Setup

Two different hardware setups have been tested so far, both with the same sensor.

3.3.1 Sensor Selection. In [41], we first described the concept for the IAG, analyzed the requirements for a tracking camera, and reviewed several state-of-the-art cameras. In that article, we identified the *Intel RealSense F200* as the most suitable sensor for our application at that time. It has a sufficient resolution of the depth sensor (true 640 × 480 pixels) with up to 60 frames per second (fps) and a low noise level. Combined with its near operating range and suitable field of view, we achieve an effective resolution of up to 10.7 pixels/cm (27 dpi) on the relief.

The RealSense F200 is a time-sequential structured light scanner. For each depth measurement frame, several Gray-coded stripe patterns are projected with an IR laser projector and filmed with a high-frame-rate infrared camera. The projector consists of an on-off modulated laser, a cylindrical lens to create a laser line, and a swinging micro-mirror to scan over the whole area [11]. A set of IR camera images with different Gray-code patterns are combined to compute the final depth image. In addition to the depth image D (see Figure 7(b)), two other images are transmitted via USB 3.0: an IR image of the scene is generated that appears fully lit by the laser projector (see Figure 8(c)), and an RGB image is generated using a separate RGB camera mounted approximately 2.5 cm away from the IR camera.

This technology has only recently become available for low-cost depth cameras. Noise levels are low, with a standard deviation below 1 mm on smooth surfaces close to the camera. However, these vary in a moiré-like pattern (see Figure 7(a)), possibly caused by interferences between the projector and camera. On steep edges, where a multitude of depth measurements are equally correct, depth measurements get less reliable.

Despite low noise and high resolution, the scanner has 3 caveats that need to be dealt with:

- Like most structured light scanners, objects near the scanner cast a projection shadow on more distant objects. Therefore, foreground objects are surrounded by pixels with no or erroneous measurements.
- Since the scanner requires multiple frames per measurement, fast-moving objects or, more specifically, depth-changes at a pixel during the measurement result in unreliable measurements. This results in blurred and unusable measurements around the edges of hands and arms when they are in motion.
- We measure significant drifts in the depth measurements of a static scene over time, possibly caused by timing issues during pattern projection with the swinging mirror. These are noticeable as a tilt of the measurements in the depth image, slowly changing over time, and some abrupt changes over a few frames. The tilt is only significant in x-direction and was measured in our setup to be up to 15 mm between the left and right end of the sensor after a cold start, and still varying over 5 mm after warm-up.

3.3.2 Setup with Tripod. Our first prototype setup has the RealSense camera mounted on a tripod (see Figure 1(c)) in order to maximize the effective resolution on the relief. The sensor is centered approximately 40 to 45 cm above the 42 × 42 cm relief, overlooking the whole relief, including a few centimeters of its surrounding (see Figure 8(c)). We chose portrait orientation, as it is more important to detect the hands beyond the relief toward the user. In addition, the sensor does not give measurements in an up to 40- pixels-wide region[7] on the very right side of the view, caused by the maximum projection width of its infrared projector. We rotated this side beyond the top of the relief, away from the user (see Figure 7(b)) to have the maximum coverage near the user.

Although the sensor is placed as near as possible to still capture the whole relief, this distance is already at the performance limit of the infrared projector and depth measurements start to become unreliable. The RealSense has a *motion-range trade-off* parameter to boost the usable range. This

[7] This is visible as a black region at the top of Figure 7(b). The width of this region is dependent on the depth.

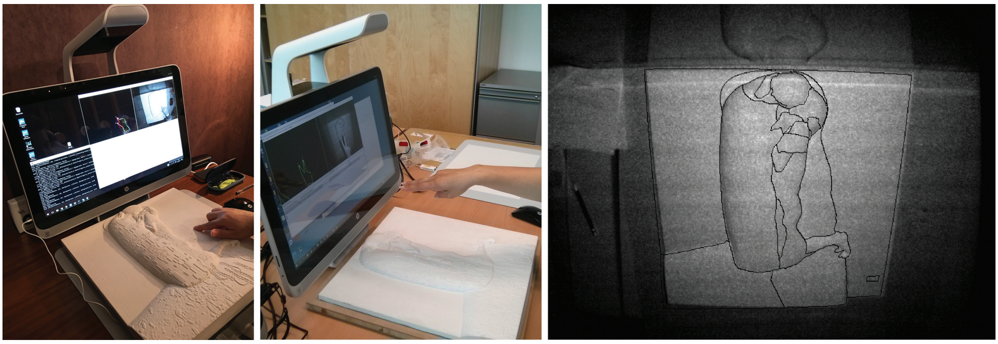


Fig. 3. Setup with HP Sprout, images © ARCHES project team. From left to right: (a) User testing the new setup, making a pointing gesture. Depth camera and projector are integrated in the top part of the device. The relief is simply placed on the table. (b) User making the off-relief gesture "two." (c) Infrared image taken with HP Sprout. Superimposed label borders are the result of our automatic calibration.

setting slows down pattern projection to allow for a longer exposure time so that with the same laser power scanning can still be performed at a larger distance. For our setup, this means a slower frame rate of only 25fps as opposed to the theoretical maximum of 60 fps for the RealSense camera. The longer exposure also makes depth measurements more susceptible to fast hand motion, which we account for in the tracking software, and slows down gesture detection and increases latency. This setup was used in the first evaluation study.

3.3.3 Setup with the HP Sprout. Our current setup uses the HP Sprout workstation, an all-in-one computer built specifically for innovative desktop 3D interaction (see Figure 3(a)). It has a RealSense F200 sensor directly integrated in a beam extending from the top of the computer screen, mounted at about 60 cm above the desk, pointing down, where we place the relief directly in front of its monitor.

Compared to our first setup (see Figure 1(c)), the sensor no longer looks straight down mounted over the center of the relief, but is mounted higher and further back over the far end of the relief. This requires an even higher setting of the RealSense's motion-range trade-off, dropping the frame rate further to around 16.6 fps. In addition, higher depth-map filtering settings are necessary, which smooth away some detail of the depth map, and the effective resolution at the relief drops from 10.7 down to 7.3 to 8.2 pixels/cm[8], making finger detection and localization more difficult.

The RealSense is mounted in landscape orientation and tilted so that the upper 20% of the sensor is worthless, as it films the Sprout's monitor (see Figure 3(c)). In the remaining vertical view, the 42 cm high relief just fits in, with very little space left in front of the relief. Thus, the user's hand is no longer detected when touching the lowest parts, as it is then already largely out of the camera view.

To cope with these changes, we simplified the label map a bit and enlarged small regions that were already difficult to activate in the previous setup. After some adjustments of finger and fingertip detection parameters and, most important, making them vary with camera distance and current frame rate, detection is now equally reliable.

Overall, the new setup seems to have positive effects on the ease of use. The lower resolution and stronger filtering of the depth map made single-finger gesture detection (see Section 3.4.1) more stable. The main source of error was the detection of multiple fingers of the hand when the user did not hide them perfectly underneath the palm (see Sections 4.2.2 and 5.2.2). With the new setup,

[8] Due to the perspective distortion, the lower resolution is at the bottom edge.

detection of such fingers is less likely. In addition, there is a lot more room above the relief. People do not accidentally bump into the camera and off-relief gestures are easier to perform and to detect.

This updated article reflects all algorithmic changes made for the new setup. For reference, the original values of the ASSETS'16 version [42] are given in brackets. These were the settings used during the first evaluation study.

Finally, the HP Sprout is nicely designed and fits much better in a museum space. In addition, it features a large touchscreen that could be used to display additional interactive content. It also has a built-in projector, normally used to project an additional screen on a detachable touch-sensitive mat on the desk. Although designed to project on a space measuring only 30 × 40 cm, it could be used to project onto smaller reliefs and even on the user's hands. In contrast to colored reliefs (e.g., [37]) projecting onto a white relief gives much more freedom. This enables us to not only project the original color of the painting but also alternative versions, such as high-contrast or simplified versions and interactive content. This promises a lot of potential for future research.

3.4 Interaction Design

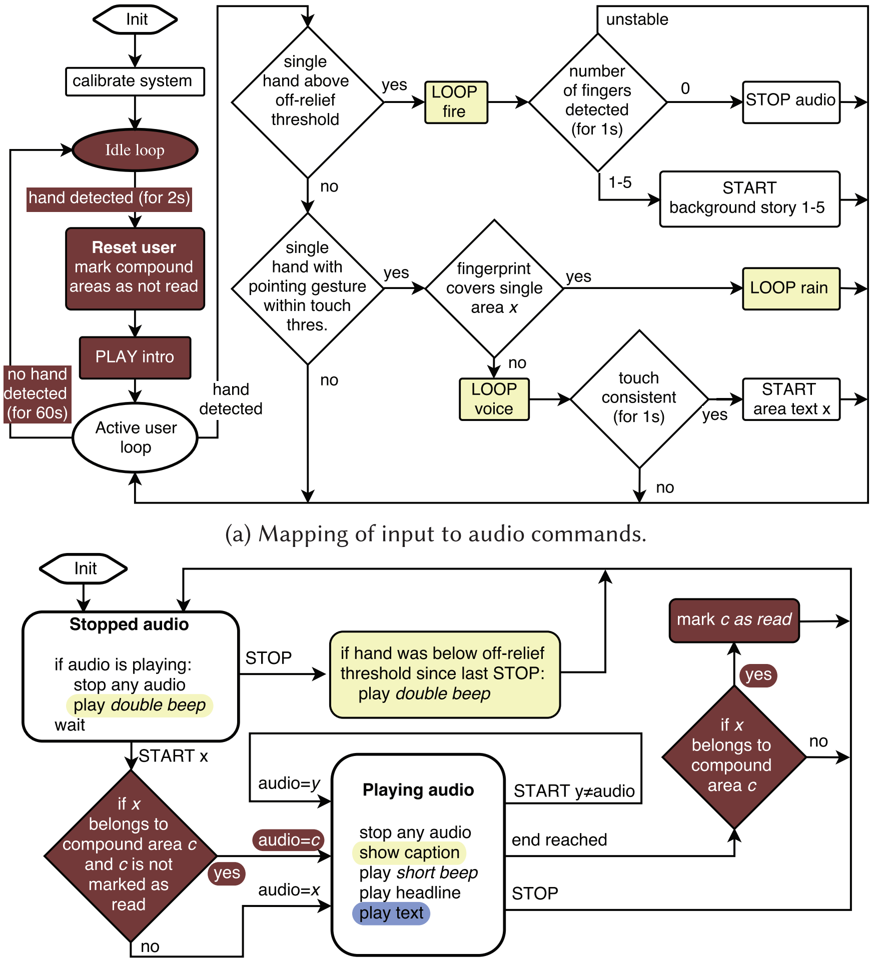
Despite ongoing research on gestures for BVI people (e.g., [24]) and user-elicited gestures (e.g., [56]), these are limited to dynamic interactions on flat screens; here, they are not directly applicable. For our specific case with static reliefs and depth sensors, careful interaction design was important. The implemented user interface will be described in the following sections and is summarized in Figure 4.

We distinguish between two kinds of information: location-specific information that describes a specific part on the explored object and general information that is unrelated to any specific location on the object. Correspondingly, we require two groups of gestures. Location-specific information should be triggered with gestures *on* the object, directly touching the part of interest. Gestures *off* the object can be used for all other interactions, e.g., to trigger the abovementioned general information but also for application commands.

These may include playback commands, of which we require at least one mandatory control, to stop an accidentally triggered comment. There could also be commands for changing interaction modes for on-object gestures, for example, "name the object," "explain the object," "historical background of an object," "describe the color of an object," "make the color under the fingertip audible," and so on.

Our design choices are based on typically used exploration strategies that we derived from informal discussions with BVI people and from observations in previous projects. Most BVI people touch the relief with both hands, often keeping one hand as a reference. This is very similar to strategies employed in reading raised line drawings. While tracing a shape with the dominant hand, the other hand stays at the starting point to have a reference for determining when tracing a shape is complete (e.g., [4]). Both hands are almost always on or close to the relief. The exploration is usually divided into two phases, although not strictly separated. In a first "overview" phase, users try to familiarize themselves with the overall composition of the painting, typically observing it with their whole hands and in larger motions. In a second "detail" phase, they are exploring selected parts in more detail, typically with the tips of individual fingers.

3.4.1 On-Object Interaction. For on-relief interaction, it feels natural to use gestures directly touching the region of interest. Using a single finger avoids ambiguous situations and matches motions occurring naturally in the detail exploration phase. We allow any finger to be used for interaction so that the users can choose whichever they feel most comfortable with. This is in contrast to Buonamici et al. [8], who require using the right index finger. We opted for the typical pointing gesture, having all fingers but one contracted into a fist (see Figure 5(g)). This gesture



(a) Mapping of input to audio commands.

(b) State machine of the audio player.

Fig. 4. Program flow and state changes of the user interface. Color indicates differences between modes. **Dark red**: *introduction* and *hierarchical exploration* only in evaluation study 1. **Yellow**: *additional sound design* and *captions* only in study 2. **Blue**: *no text* in training mode.

feels natural while at the same time it is only rarely used during normal exploration, which mostly avoids triggering unwanted audio. Coincidentally, the mobile app prototype THATS (see Footnote 4 on Page 4) uses a very similar approach.

In order to account for the two exploration phases (see Section 3.4), at first the region of the selected part is named; after a short pause, the detailed description follows. Every playback can be interrupted by triggering another region. This enables the user to quickly scan the object during exploration and to easily locate parts of interest for more detailed information. Each new trigger is accompanied by a short *confirmation click sound*, an important feedback to the user and to avoid confusion when a text was unintentionally interrupted.

3.4.2 Hierarchical Exploration. As mentioned before, two basic regions (male and female figure, see Figure 8(a)) were further subdivided into smaller parts, mainly the parts of the bodies

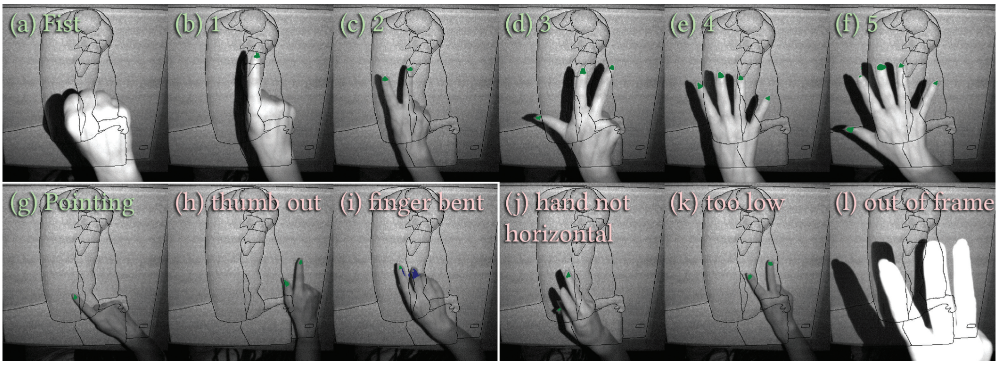


Fig. 5. Gestures as performed by a user, with detected fingertips overlaid on IR images. (a) to (g) show the full set of currently implemented gestures correctly detected by the system. Note that the 1-5 finger gestures can be performed with any fingers of the hand; thus, this list is not exhaustive. (h) to (l) depict common hand posture errors that are not correctly detected.

and clothes. The idea is that, in the beginning, only the six basic regions are used to gain a quick overview. Once the user has heard the full detailed description of a figure (which includes important information about the posture and relations to other regions), the subdivided parts of the figure will become available instead of the basic region.

3.4.3 Off-Object Interaction. As most users keep their hands on or close to the relief, we use the space above the relief to trigger off-object interactions. A *closed fist* gesture at least 20 cm (8 inches; it was 10 cm in prototype 1) above the relief will stop the current playback (see Figure 5(a)). The other hand may still remain on the relief to stay oriented. Background information can be triggered by *number gestures* (see Figures 5(b)–5(f)). As the number of our general texts (see Section 3.1) is exactly five, we chose to simply count the number of extended fingers of the lifted hand, which also generalizes to different number gestures in different cultures. Nonetheless, additional different gestures may be implemented in the future. Once the gesture is detected, the *confirmation click sound* is played, followed by the number of fingers and the title of the chapter. This allows the user to correct the hand pose until the desired number of fingers is detected and to browse through the headings of the available texts until the desired one is found. Again, a newly detected gesture interrupts the playback of the former. Once the user is satisfied with the choice, the text starts directly after its title, and the users can lower their hand and continue tactile exploration while listening.

3.4.4 Additional Sound Design. Based on the feedback of the first evaluation study (see Section 4), we implemented additional sounds to aid participants in making correct gestures:

- A *fire-crackling sound* indicates that the hand is over the minimum height required for the *off-relief* gestures.
- Two further sounds indicate that the single-finger touch gesture is correctly detected:
  - An *ethereal voice* is used on regions where descriptions are available;
  - A *rain sound* indicates that the finger is on the border between two or more regions and the algorithm cannot determine which region to select.

Before these additions, touching the border between regions would not give any feedback, and users frequently assumed that the finger gesture was not detected. Now, when hearing the *rain sound*, they can move the finger more toward the desired region until the touch is detected.

Alternatively, users can also use the sounds to scan the bounds of a region with their fingers. An interesting idea for future investigations would be to try a distinct sound for each region. It could help with orientation but could also be too confusing.

These additional sounds were used in the second evaluation study (see Section 5). We took care to select subtle ambient-like sounds so that people who need them can clearly hear them, while it does not render the spoken text incomprehensible or disturb people who do not need them. Still, these sounds double as a reminder for those who do not need them—a reminder to relax their hands into noncommand poses in order to avoid accidentally triggering another sound.

As a further cue, the sounds get louder the clearer the gestures are detected, implemented as a percentage of the frames in which the gesture was detected over the last 0.75 seconds. Incidentally, this also produces a nice fading effect. Furthermore, the sounds pan from left to right depending on where the gesture was detected in the camera frame. This might help some users to keep the hand centered under the camera.

Finally, we changed the *confirmation click sound* to a *short beep*, which can be recognized more easily, and we implemented a distinct *stop sound* as a *double beep*, which is now always played when a sound is stopped either by the *fist gesture* or at the end of each description. This might help to distinguish short text pauses from the end.

3.4.5 Screen Design. During the first evaluation study, participants with low vision unexpectedly observed the debug views on the screen, which we had for the purpose of technical support. These were showing the output of the hand detection system on the left (see Figure 8(b), without the arrows, variable names and white circles); the right debug view showed the output of the touch detection output superimposed on the infrared camera image (see Figure 8(c)). Some participants seemed to enjoy watching these views. They could see where the interactive regions were and whether the system detected the touch event, as the interaction region then becomes colored.

As the second evaluation study explicitly also targeted participants who can see, we added a visualization of the detected fingertip pixels on the touch detection output so that people could directly see which fingertips are detected and where exactly and how large the detected fingerprint actually is (see Figure 5). In addition, we added simple subtitles for all spoken texts in the lower half of the screen: one static text per description, font size 30 points, black text on white background.

As the new setup with the HP Sprout has an integrated projector, a future implementation could directly project such a visualization on the relief and fingers.

3.4.6 Making It Self-Explanatory. The current prototype was designed as an installation in a museum for people who are not familiar with the system. Therefore, the first interaction is to simply put the hands on the relief, which triggers a short introduction explaining the interface (see Figure 6). After the system is not used for a given amount of time (currently 2minutes), the system is reset and waits for the next user. This mode was used in the first evaluation study (see Section 4).

As this mode did not work well in the initial tests of the second study (see Section 5), we implemented a *training mode* as an aid for the examiner. This is basically the same as the normal mode but without description texts in order to minimize audio output and help the participants to better concentrate on the examiner's words and on practicing the gestures. In this mode, the *on-relief* gestures trigger only the *confirmation beep sound* and the name of the touched parts, and the *off-relief* interaction repeatedly tells the number of detected fingers (0–5).

3.5 Software Implementation

Based on the selected set of gestures, the requirements for the gesture detection system are rather simple:

> Welcome to the interactive touch relief of Gustav Klimt's "The Kiss".
>
> Please explore the relief with your hands, as you like. If you want to know more about a part, form a flat pointing gesture with one hand. Clench all the fingers of the hand to a fist, but keep one finger flat extended, in a way, that from above exactly one single pointing finger can be recognized. You don't need to lift the hand from the relief. Even the second hand can stay on the relief, as long as it doesn't touch the pointing hand.
>
> When the gesture is correctly recognized, at first you hear the name of the touched part, followed by a more detailed description. While the audio is playing, you can continue exploring the relief. The audio continues playing, until you touch another part with the pointing gesture.
>
> To stop the current comment, form a fist and lift the hand a few inches above the relief.
>
> In addition, there are five introductory texts for the painting. You can start them by lifting one hand about 4 inches above the relief, and spreading one to five fingers horizontally. We recommend listening at least to text number 1 as an introduction.
>
> Have fun, exploring!

Fig. 6. Introduction text of the interactive audio guide.

- a reliable detection of hands and individual spread fingers,
- measurement of the palm height for *off-relief* gestures, and
- detection of touch events of a pointing finger, together with the position of the touch.

As pointed out before [41], the optimum for the proposed system would be "an out-of-the-box solution for articulated finger tracking, [that works] on relief surfaces." The only publicly available implementation that we found is the CVRL FORTH Hand Tracker [36] already tested by [8] for a similar application. However, we could not use the software, because

- it currently supports sensors of the Kinect family only,
- the demonstrator tracks only a single hand and requires an initialization pose,[9] and
- according to Buonamici et al. [8], it loses tracking at fast hand movements without recovery, requiring new initialization.

Similar approaches have been published (e.g., [46]) and professionally developed by the start-up NimbleVR[10] but implementations are not available.

Since their approaches are quite demanding in terms of hardware and an implementation from scratch was beyond the scope of our work, we decided to implement a simpler, silhouette-based approach, which is basically a 2D problem, for which a lot of well-studied algorithms are available. These basically work when the hand is more or less parallel to the camera plane and the relevant fingers can be detected in the silhouette (see Figure 8(b)), which means that the fingers must not touch or overlap each other. With the selected set of gestures and the camera setup with an almost parallel view of the hands, these requirements are satisfied.

Because of the demonstrated robustness and detailed documentation, we based our implementation on [58]. We will shortly outline the original approach and detail the parts that had to be modified in order to make it work on our specific setup, directly on a relief surface.

3.5.1 Silhouette Detection. The original article [58] addresses both color-based foreground segmentation using RGB cameras and depth segmentation using a depth camera. In our prototype, we

[9] Extensions were published (e.g., [27]) but their implementations are not publicly available.

[10] NimbleVR (http://www.NimbleVR.com/) presented a working hand-tracker that even worked on desk surfaces but since its acquisition ceased to distribute its software.

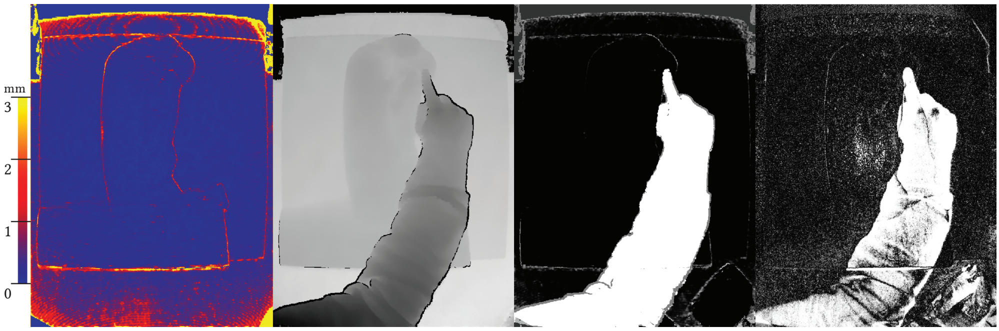


Fig. 7. Foreground segmentation, images © Andreas Reichinger. From left to right: (a) Standard deviation of background measurements over 100 frames; (b) depth image D; (c) foreground probability from depth p_D; (d) foreground probability from infrared p_{IR}. Note the higher noise and more difficult segmentation.

use the depth measurement as the main segmentation key complemented by the IR image, which proved to be adequate for our requirements. We are able to reliably segment the hands, even at a fingertip pressed against the surface, with a height difference as low as 5 mm. Color information is currently neglected in our prototype, but might further improve segmentation in a future implementation. However, the aforementioned possibility to project onto the relief may make the color information unusable for this purpose.

In contrast to hand-tracking approaches that operate in free space, we cannot use a constant depth threshold, as the hand is supposed to operate directly on the surface. However, we can exploit the fact that, in our static setup, the background does not change and can be calibrated once. This is in contrast to other setups [8, 47], in which the object and/or camera are allowed to move relative to each other and more complex tracking solutions have to be used.

During background calibration, the mean μ and standard deviation σ (see Figure 7(a)) of each pixel in the depth and IR images are computed from 100 consecutive frames, yielding a Gaussian distribution. A foreground probability based on depth, p_D (see Figure 7(c)), is computed as the one-sided p-value at the current depth, offset by a safety margin of 5 mm. p_{IR} is computed as the two-sided p-value (see Figure 7(d)). The combined foreground probability p is the weighted average $p = (w_D \cdot p_D + w_{IR} \cdot p_{IR})/(w_D + w_{IR})$, where the weights are computed as $w_D = \alpha/\sigma_D$ and $w_{IR} = 1/\sigma_{IR}$, and $\alpha = 100$ trades off depth for IR.[11]

As outlined in Section 3.3.1, rapid depth changes at a pixel caused by fast-moving objects results in unusable measurements around the edges of hands and arms when they are in motion. In order to still extract a meaningful silhouette, we track the standard deviation of the depth measurements σ_M of the last 5 frames, compute a depth-motion penalty $\beta = 0.3\,\text{mm}/\sigma_M$ clamped to $[0.2, 1]$, and replace $\alpha = 100/\beta$. If an object moves fast, the variance of an edge pixel is high, β gets low, and less weight is given to the depth probability p_D, effectively falling back to a foreground detection based on the IR channel values. Detection based on a single intensity value is of course rather error prone. Nevertheless, skin and worn clothes often have a significantly different IR reflection than the tactile relief (see Figure 8(c)), providing an additional clue as to where the correct silhouette is located.

The resulting probability image is smoothed (Gaussian blur $\sigma = 3$ pixels), thresholded to 0.5, and eroded with a 3 × 3 kernel to yield the foreground segmentation mask.

[11] If background (I_B) or current (I_C) or both (I_{BC}) measurements are invalid, special (p, w) pairs are used: $I_B = (0.5, 0)$, $I_C = (0.7, 1)$, and $I_{BC} = (0.2, 1)$.

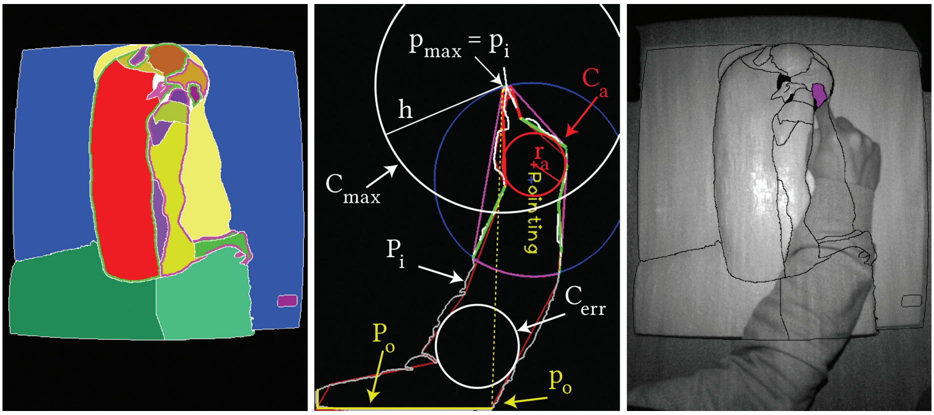


Fig. 8. Gesture recognition, images © Andreas Reichinger. From left to right: (a) Hand-drawn label image, warped to camera space. Light-green and purple outlines indicate the merged base labels of the two figures. (b) Hand detection output and palm detection diagram. (c) Infrared image with superimposed label borders and touched label (purple).

3.5.2 Continuous Sensor Calibration. As outlined in Section 3.3.1, the depth measurements show significant drift over time, which compromises the tight tolerances of the depth-based segmentation and therefore requires continuous detection and calibration with respect to the stored background image. We model the drift d at a pixel (x, y) as an additive tilt to the raw metric depth measurements d_{raw} in the form of $d(x, y) = d_{raw}(x, y) + \delta_0 + x\delta_x + y\delta_y$. This approximation proves to be sufficient for our application but is presumably not very accurate, as the measurement errors probably occur in the disparity space and are not necessarily linear in the metric depth measurements.

The tilt parameters δ_0, δ_x, and δ_y are estimated as a two-dimensional linear regression on the difference between the stored mean background $\overline{b}$ and a running average of the latest 5 depth measurements $\overline{c}$. Currently, detected foreground regions (enlarged by a safety margin of 12 pixels) are excluded, as well as unreliable measurements of b and c, which yielded at least one invalid measurement during the average computation. The differences are clamped to ±3 mm to avoid excessive outliers and are weighted by the inverse of the computed standard deviations of b and c to lower the impact of noisy regions. The changes are applied gradually to avoid abrupt changes, using an IIR filter mixing in only 10% of the new solution to the old tilt. Owing to performance reasons, the adjustments are performed only once every 4 frames.

3.5.3 Palm Detection. Following [58], the hands are detected solely based on their silhouettes. From the foreground segmentation mask of Section 3.5.1, all connected components larger than 3,000 pixels (5,000 in prototype 1) are chosen as potential hand regions, and their contours are extracted. Contained contour holes larger than 100 pixels are stored and could be used to classify some special gestures (e.g., a pinch gesture). All smaller holes are discarded as potential scanning artifacts. This approach works only if the hands and arms do not touch or overlap. This is satisfied for the selected set of gestures, since the user can be instructed to move the other hand away from the interacting hand. In a future implementation, this can be improved using multiple sensors and/or a fully articulated hand tracker that allows overlap (e.g., [27]) or by taking depth discontinuities into account.

Assuming that the region contains a single hand, the position of the palm has to be found. The original approach [58] finds the largest circle C_a inscribed in the silhouette. However, this often fails, e.g., when the user is wearing loose clothes (see Figure 8(b), C_{err}). Suggested online "solutions" include to disallow wearing such clothes, but that would not be meaningful for a public installation. Our solution is different (see Figure 8(b) for an example annotated with the following symbols).

We first intersect the contour with a rectangle, 5 pixels from the image border, and find the largest consecutive contour-part P_i that does not touch the border. If no part touches the border, we assume that the hand was segmented without the arm and continue with the largest circle search. Otherwise, we close the polygon P_i along the rectangle with one or more line segments P_o. We then find the point p_{max} on P_i that is most distant to all the points on P_o as

$$p_{max} = \underset{p_i \in P_i}{\arg\max} \min_{p_o \in P_o} \|p_i - p_o\|. \tag{1}$$

We create a bounding circle (50 pixel radius) around p_{max} and compute the average depth measurement $\overline{d}$ of all valid points inside this circle and the contour. We estimate the expected maximum hand size h at such a distance as $h = 200\, pixels \cdot 390\, mm/\overline{d}$. The maximum inscribed circle C_a with radius r_a is then only searched inside a bounding circle C_{max} around p_{max} with radius h. We compute the depth of the palm as the average depth of all valid points inside C_a.

3.5.4 Fingertip Detection. Fingertip detection is similar to [58]. The hand silhouette is clipped to a bounding circle 4 times the radius of the palm (blue circle in Figure 8(b), was 3.5 times in prototype 1), the resulting polygon is simplified, and convexity defects are computed and filtered as to whether they could represent the empty space between fingers.[12]

Between neighboring pairs of all accepted convexity defects, we test for potential fingertips. We modified the criteria as follows:

(1) No contour part between two convexity defects may go over the image boundary or the arm (new in prototype 2).
(2) The arc distance along P_i between two consecutive convexity defects must be below 120 pixels, and their angle must be below 60°.
(3) Similar to [58], we require the k-curvature[13] to be below 60°. But instead of using a constant $k = 30$, we take the curvature as the minimum k-curvature computed using a number of different k varying from 30 to 60 pixels to allow for locally flat but still elongated fingertips to be detected. Note that we limit k so that the search does not go beyond the far point of the convexity defect (the minimum point in prototype 1).

We also modified the fingertip localization. While in [58] the fingertip location is solely determined from the k curvature points, we found this to be too unreliable owing to the often rather jagged fingertip contours occurring in our setup. Instead, we take the pixels of the finger region inside a circle around p°,[14] and compute their oriented bounding box aligned with an estimate of the finger's direction[15] to get more reliable estimates for the finger's width. In order to extract the tip region, we take the valid pixels inside the top square region of this bounding box. We estimate the center of the fingertip as the centroid of these pixels, and compute the z-location of the finger as the average depth measurements of these pixels. Finally, we classify the fingertip's quality into three categories: If the tip region has too few pixels ($<15 \cdot (400\,\mathrm{mm/depth})^2$) it is not classified as

[12] We use slightly different criteria than [58]. Following their notation, instead of $r_a < l_d < r_b$, we require for both $l \in \{l_a, l_b\}$ that $l > 0.1\, r_a$ and at least for one l that $l > 0.4\, r_a$. The criterion $\theta_a < 90°$ was removed.

[13] The k-curvature at a contour point p° is defined as the angle at p° of the triangle formed by (p^-, p°, p^+) with p^- and p^+ being the points along the contour that are k pixels to the left and right of p°.

[14] The radius is set to be 2/3 of the Euclidean distance between p^- and p^+.

[15] As in [58], we estimate the finger direction as the line from the halfway point between p^- and p^+ to the point p°.

a finger (constant 50 pixels in prototype 1). If the finger is too wide for the given depth (width × depth > 25 pixels × 400 mm), it is labeled as *blob*, being probably a union of two fingers; it is labeled as a valid finger only if it satisfies both conditions.

Our hand-detection algorithm (like the original algorithm [58]) relies on a number of assumptions and hard-coded values, which have been fine-tuned to our specific setups. Despite its simplicity, the algorithm is remarkably robust, as the hard-coded values are just conservative bounds on the search space and allow a wide range of actual configurations. Observations during user tests confirmed that it works for a wide range of people, from 11 years old to the elderly, but there is no guarantee that it works for every user. For instance, the algorithm had problems with one participant with a medical condition that made the individual's index finger bend to the side, as it violated the assumption of straight fingers in our silhouette detection. Similarly, other objects placed on the relief can be detected as a hand as well, but normally do not trigger any actions.

3.5.5 Gesture Recognition. We do not perform frame-to-frame tracking, as this is not necessary for the current set of gestures and would introduce recovery problems once a hand was lost. Nevertheless, we need to make gesture recognition robust, as the detected hands and fingers may vary from frame to frame. Our solution is to require a gesture to be detected in the majority of the latest frames. For instance, off-object gestures are triggered if the average palm-to-sensor distance is below a certain threshold in 70% of the frames in the last 0.75 seconds, *with* the same amount of fingers detected (75% of the last 20 frames in prototype 1).

Similarly, the "introduction text" is triggered when, in the last 2 seconds in at least 90% of the frames, a hand with at least one *good* finger was detected. This makes it robust against nonhand objects placed on the relief and introduces a small delay before the text starts.

3.5.6 On-Object Touch Event. In order to relate the detected on-object touch events to regions on the relief, and therefore to different audio files, the regions have to be labeled by the content author (see Figure 8(a)). Since our setup is static, we simply sketch the labels on a once acquired IR image of the relief (see Figure 8(c)). While this does not adapt to a different camera placement, as in [47] and [8], it proved accurate enough for our first proof-of-concept implementation as used for the first evaluation study.

Finally, the fingertip location has to be mapped to the regions. While Buonamici et al. [8] use a complex 3D search for the nearest point of a point cloud of the relief to the fingertip, we again use a simpler 2D approach. Since the camera observes the relief almost straight on and the labels are defined in camera space, the *xy*-location of the finger is already given with maximum precision in the foreground mask. We simply take all pixels of the detected fingertip that are within some depth tolerance to the depth background, and collect the labels of these pixels. If at least 70% of these pixels are on the same label, the detection is considered unique (90% in prototype 1). Otherwise, the finger might be on a border between labels rendering it impossible to determine which label the user meant.[16] A touch event is generated when at least 70% of the frames in the last 0.75 seconds (10 frames in prototype 1) detected the same unique label (see purple area in Figure 8(c)).

Note that with a sufficient depth tolerance to robustly detect touch, actual touch is not distinguishable from a slightly hovering finger. However, no participant seemed to have noticed that, as each mostly kept the finger on the relief.

3.5.7 Relief Calibration. During the second evaluation study, the assumption of a static setup of the previous section proved to be impractical in practice. For each subsequent setup, it is required to exactly reproduce the same relief to camera pose for which the label map was drawn. All six

[16] In prototype 2, we now give distinct audio feedback in this case; see Section 3.4.4.

degrees of freedom of camera position and orientation have to match. This is achievable for an experienced operator when the exact same tripod is used without moving its joints between setups but is otherwise highly impractical.

Our second setup facilitates this task, as the camera on the HP Sprout always has the same height and tilt relative to a flat table it is placed on, leaving only three degrees of freedom of placing and rotating the relief on that table for manual alignment. Still, this task is tedious and needs to be repeated each time the HP Sprout or relief is moved.

Therefore, we implemented an automatic calibration process. The content author draws the label regions L_M now no longer over the IR image but rather over the depth map D_M that was used to manufacture the relief. With the known size (s_x, s_y) and height h, a 3D model M_M can be created with vertex coordinates $P_M(x, y) = (x \cdot s_x, y \cdot s_y, h \cdot D_M(x, y))$, textured with the label region map L_M. During startup, the calibration automatically detects the relief in the depth sensor's point cloud P_S and recovers the relative Euclidean transformation T that transforms the 3D model M_M to the detected location in the point cloud P_S. As the point cloud P_S is given in the coordinate system of the IR camera, the recovered transformation T can be used to render the textured relief model M_M as seen from the IR camera, which produces exactly the same label image as required for our system.

The question is how to reliably detect the relief and find the transformation T. CamIO [47] and MagicTouch [48] use fiducial markers, which we do not want to apply on our reliefs but would be an option for home use or flat documents. Buonamici et al. [8] directly align the depth camera point cloud but require user interaction for a prealignment step (touching the four corners of the relief), which would be an option for home use but is not practical in a museum environment, where a fully automatic approach is more desirable. An automatic approach could constantly calibrate the relief during inactive periods or detect the change of a relief and quickly calibrate on that one. A fast implementation could even enable real-time tracking and allow the user to move the object during exploration.

Image-based object recognition methods (e.g., [31]) are fully automatic but difficult to use in our case, as these require distinctive patterns on objects and our relief is plain white. Such an approach would be directly usable for printed documents, for example. However, this might get difficult with the planned projections onto the reliefs, as these would mask any available textures on the object. Based on these considerations, we believe that a purely geometric, surface-based method [21] is more suitable in our case.

We currently use our own implementation of a point-pair feature matcher based on [12] followed by an ICP optimization based on [9]. Since the method is quite sensitive to surface variations of the models, we had to down-sample and blur the relief model depth map D_M significantly to resemble the low quality received from the depth camera. Further, the camera's point-cloud P_S is created from an average of the five most recent depth maps to reduce noise, and normals are computed using a plane-fitting algorithm over a 3 mm radius to make them more robust to noise.

Object recognition takes below 10 seconds, which is acceptable for a static setup. Although we just take the best recovered pose hypothesis and do not verify the result, an acceptable pose is recovered in at least 4 out of 5 cases and almost always recovered after a second try. The rendered label map still has a few pixels deviation (see Figure 3(c)), but only the border seems to be shifted to the left, while the contours around the figures fit tightly. We attribute this either to distortions in the plaster relief mold, resulting in an actual physical deviation, or it could be caused by nonlinear distortions in the depth values received from the depth camera. Although we carefully calibrated the IR camera using OpenCV's calibration methods [5] based on calibration images taken with a checkerboard pattern, a more sophisticated calibration method specifically for RGB-D cameras (e.g., [25, 49]) could alleviate this problem.

Table 1. Results of Study 1: Demographic Data (Study 1, Second Session in Vienna)

Sex	Male		Female								avg.
	8		5								
Age (years)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80			50
		1		1	1	7	2	1			
Vision	totally blind		legally blind		low vision						
	6		4		3						
Years blind	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80			46
	1	1	3		2	4	2				
% of life VI	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%	78%
		1		1	2			1	2	6	
Can read Braille?	No		Yes								
	6		7								
How often do you visit a museum per year?											
	0×		1×		2×		≥3×				
					1		12				
1	2	3	4	5	6	7	8	9	10	sum	
How interested are you in museums in general? (1 = not, 10 = very)											
							3	3	6	12	9.3
How important are tactile materials (in a museum) for you? (1 = not, 10 = very)											
1	1	1			1	2	1	2	4	13	7.1

Although not perfect on a global scale, all important features are reliably aligned, enabling fully functional interaction and label mapping. This already proves that automatic calibration is possible. A future implementation may be accurate and fast enough for real-time detection and tracking. Then, it would be possible to exchange the relief under the sensor at runtime to move and rotate the relief freely during exploration and even to use the system with full 3D objects.

4 EVALUATION: STUDY 1

The implemented system was evaluated in two separate studies, each consisting of multiple sessions. The first study, presented in this section and already present in the conference version of this article [42], was targeted at BVI people. The second study is presented in Section 5 and was conducted with a broader range of participants.

Study 1 consisted of two sessions. The first session was an informal evaluation 2.5 hours long performed in Manchester, Great Britain, with 7 mostly elderly BVI people, with the majority having low vision. Based on this first feedback, we implemented a structured evaluation that took place in the course of 2 full days in Vienna, Austria, with 13 people (5 female, aged 11–72 years, average 50 years; see Table 1).

Of the 13 volunteers, 6 were *fully blind*, with no sense of sight; 4 had a minimum rest of sight equivalent to the American classification *legally blind*, that did not help them perceive images (one wearing a hearing aid in addition); and 3 had *low vision*. Nine participants have been visually impaired for the majority of their lives, three at least 20 years and one for 9 years. Seven are able to read Braille, and all are very interested in museums, going at least twice a year, four at least 4 or 5 times, two even over 20 times. Most participants reported that touch tools are important for them (on a Likert scale from 1 to 10, six reported 9–10, four reported 6–8, three <3).

The basic tripod setup was used (see Section 3.3.2), and none of the abovementioned extensions were implemented. The presented prototype was part of a larger evaluation with four different

Table 2. Results of Study 1: Questions Concerning the Technology of the Interactive Audio Guide

1	2	3	4	5	6	7	8	9	10	avg.	sum
1. Did the IAG help to better understand the painting? (1 = no, 10 = absolutely)											
							1	5	7	9.5	13
2. How did you get along with the system? (1 = not well 10 = very well)											
				1		4	1	2	5	8.4	13
3. How understandable is the introduction for you? (1 = not, 10 = very)											
			1			2	1	6	3	8.5	13
4. How easy is it to perform the gestures? (1 = not, 10 = very)											
				1	1		2	3	5	8.8	12
5. How important is it for you that audio is only played when wanted? (1 = not, 10 = very)											
						1	2		10	9.5	13
6. How easy is it to trigger a desired comment? (1 = not, 10 = very)											
			1		1		6	2	3	8.3	13

devices. We concentrate here on the questions regarding the present system. The results of the full evaluation will be presented elsewhere. Only one relief was tested to keep the load of the evaluation tractable, but during development a number of reliefs were used. The participants spent at least 30 minutes evaluating this device and could test it as long as they wanted. Afterwards, the examiner asked 24 questions in a structured interview. Most questions asked for a ranking on a 10-point Likert scale, 1 being the most negative, 10 the most positive ranking; giving no answers was allowed. These are summarized in Tables 2–4.

The participants were seated in front of the relief so that they could comfortably reach the whole relief. The introduction was kept minimal, stating the general idea of the IAG and pointed out that an affordable low-cost depth camera was used. Participants were shown where the relief and camera were located so that nobody accidentally crashed into it. No interface was initially described as we wanted to test whether the introductory text (see Figure 6) was sufficient to use the device. One examiner was always present, observing the interaction, and prepared to answer questions or help with the interface if there were obvious problems.

4.1 General Impression

We got very good feedback for the system in general (see Table 2). On the question regarding whether the IAG helped in gaining a better understanding of the painting, all gave a rating above 8, with an average of 9.5. Several people spontaneously praised the system, calling it “super”, “perfect,”, “cool,”, “I am in love with it,”, “It has to go into the museum, for eternity,” and “finally I have a mental picture of ‘The Kiss.’” They liked the direct interaction with the finger, the intuitive interface, its simplicity, and the combination of 3D touch and simultaneous audio, the in-depth descriptions, and that the texts are “pleasantly short.” Some felt that the independence from a human guide gives them the freedom to explore it without pressure, as long and as detailed as they wanted. One person “felt guided” during the exploration, probably caused by descriptions that cross-reference nearby regions, guiding the person from one region to the next. Another put a thought into the future and liked the fact that the object to be observed could be exchanged below the camera, and began to sketch scenarios in which he could choose between different reliefs and put them under the camera in a kiosk in the museum or at home.

Negative feedback was rare. One person with low vision questioned the necessity of such a system, concluding that it probably depends on the complexity of the relief. In general, it seemed that totally and legally blind people appreciated the system most, as people with low vision are not that dependent on touch and audio. One person wished to have a description about the painting first, but did not follow the suggestion in the introduction to first listen to the general text about the painting. When asked how good they were getting along with the system, all but one ranked it above 7, two gave it a 9, and five a full 10. One person ranking 7 noted: “the functions are clear but it did not always work.”

4.2 Interface

Since the system is designed as a kiosk in a museum, an introductory text (see Figure 6) should be sufficient to use the interface. Indeed, nine people rated the understandability of the introduction above 9. Participants giving lower ranks stated that they did not pay full attention, that the text was too fast or too long, or that they simply did not memorize everything. Some wished for a possibility to repeat the introduction or to include an interactive tutorial session.

Four participants immediately mastered the interface and could reproduce all gestures without any intervention from the examiner. Others needed tips (e.g., “Please also hide your thumb under the hand.”) or slight manual corrections of their hands: If they allowed it, the examiner carefully moved the hand and fingers until the gesture was correct and asked the participant to repeat the gesture holding the hand like that.

After a short familiarization phase, nearly all could perform the gestures on their own. When asking how easy it was to perform the gestures, eight participants rated 9 or higher. Comments included that it is “as simple as possible,” “as good as it gets,” and “even funny to silence it with the fist.”

Especially significant was the confirmation of our design goal, to only have the system play audio when it is explicitly requested by the user. Ten participants gave a ranking of 10, stating that it is very important to concentrate on the tactile exploration every now and then without being disturbed by constant audio information.

4.2.1 Off-Object Gestures. Off-object gestures worked for most people as they got audio feedback about the number of detected fingers when reaching the desired height, allowing instant corrections. Problems were mostly caused by the hand not positioned at the required minimum height (see Figure 5(k)) or the camera not detecting all fingers for the chapter selection. Either the hand was partly outside the camera (see Figure 5(l)), or was not held fully frontal to the camera (see Figure 5(j)). Some people frequently lifted their hands up from the relief and accidentally triggered off-object commands. This mainly occurred with people with low vision and while talking to the examiner.

Participants encountering such problems suggested using hardware buttons, voice-commands, or knocking signals instead of the gestures. Some also expressed the desire for additional playback commands, such as pause, back/repeat, or the change of reading speed. Others disliked the idea of browsing the text headlines with the finger gestures and requested a table of content function.

4.2.2 On-Object Gestures. The pointing gesture worked for most people, at least after some training. A common problem was that sometimes more than one finger was detected when the fist was not fully closed. Mostly, the thumb was still extended as the testers did not think of it as part of the fist (see Figure 5(h)). Some participants mentioned that the gesture is uncomfortable or feels unnatural and expressed their wish to relax the gesture and to allow more fingers being extended, at least the thumb. It remains unclear regarding how to best select the interacting fingers in such alternative gestures. We theorize that performing a kind of double-tap with the specific finger might be better.

Another source of error and probably also the main source of the discomfort is the current requirement to perform it in a very flat way, required by the silhouette-based hand detector. Especially elderly people from the first evaluation session had problems performing the required flat pointing gesture (see Figure 5(i)), as their hands were already less flexible or they had medical conditions, such as arthritis. Some participants thought that the pointing gesture was more like pressing a button and held the finger steep down, making it difficult to detect. This gets more severe at the top of the relief: as the camera is mounted over the center of the relief, the observation angle gets steeper to the top edge and even with a flat hand position, the fingertip detection gets less reliable. A possible solution would be a different camera placement, observing the hands from a lower perspective. However, this might have negative implications on the localization. A combination of multiple cameras might be able to solve this in the future.

Another limitation of the current setup occurs near the left, right, and lower edges. We placed the scanner as low as possible to maximize the effective resolution, with only a few centimeters around the relief still captured by the scanner. When the finger touches a feature near an edge, the hand typically protrudes beyond the relief and outside the scanning region, hindering proper hand detection. A future setup with possible higher-resolution scanners should keep ample space around the relief.

The hierarchical exploration was not specifically tested but seemed to work for most users. People seeking detail listened to the top-level description and explored further without noticing the transition. Others were either satisfied with the general description or did not fully listen to it. In the future, an explicit level control may be investigated.

Last, localization accuracy has some room for improvement. The majority of testers rated the question "How easy is it to trigger a desired comment?" with 8. There were no problems selecting larger areas. However, most participants had problems selecting the smallest regions, such as the hands of the figures, which are not much larger than the fingertip itself. This is probably caused by the current algorithm, which requires 90% of the fingertip pixels to be over a single area. Although it is possible to select all regions, especially for sighted users with visual feedback from the tracking system, it is currently unknown how to make it easier at small regions as well as at borders between two regions. The single point interaction of [8] may be advantageous here.

4.3 Content

Nearly all participants were satisfied with the presented content (see Table 3). High rankings confirm a good readability of the created relief. The average rating for the general impression was 9.2, 9.0 for getting the overall composition and 8.4 for getting the details of the painting. All but one stated that the amount of detail was chosen right; one said it was too much. They liked the high elevation, the three-dimensional plastic appearance, the size, the detailed textures, the smooth, rounded parts, the recognizable body parts, and the faithfulness to the original painting. Some wanted it slightly larger and higher or suggested detachable parts for easier recognition.

The material (DuPont Corian®) was comfortable for most; only two people did not like it at all. Four people mentioned that it would be nice to have a colored relief for people with low vision, while others found it irrelevant as long as the original can be seen next to it.

People were highly satisfied by the texts (average rank, 9.3) and by the number of described parts (average rank, 9.1). One very eager participant would have liked to know the number of descriptions in advance in order to check to have not missed anything. Indeed, there is currently no mechanism that reports the completeness of the exploration, as our system was designed as an interactive experience that should give descriptions to the parts the user is most interested in. Conversely, such a mechanism could put unwanted pressure on the users regarding whether they are capable of finding all hidden spots. The positive self-reported ranking shows that they found

Table 3. Results of Study 1: Questions Concerning the Content, i.e., Relief and Audio Description of "The Kiss"

1	2	3	4	5	6	7	8	9	10	avg.	sum
1. Your general impression of the relief? (1 = not good, 10 = very good)											
				1			1	2	7	9.2	11
2. How good did you get the overall composition? (1 = not, 10 = very)											
				1	1			4	7	9.0	13
3. How good did you get the details of painting? (1 = not, 10 = very)											
		1		1			2	4	4	8.4	12
4. How satisfied are you with the number of described parts? (1 = not, 10 = very)											
							3	5	3	9.1	11
5. How satisfied are you with the description texts? (1 = not, 10 = very)											
						1	2	3	7	9.3	13
6. How would you rate the amount of detail in the relief?											
too little		OK		too much							
		10		1							11

Table 4. Results of Study 1: Questions Concerning the Application of this Technology

1	2	3	4	5	6	7	8	9	10	avg.	sum
1. Do you find it meaningful to use this technology in museums? (1 = no, 10 = absolutely)											
							3	1	9	9.5	13
2. Would you rather go to a museum when it offers an IAG? (1 = no, 10 = absolutely)											
1	1				1		1	1	8	8.2	13
3. Would you use this technology at home? (1 = no, 10 = absolutely)											
1		1		3			1	3	4	7.2	13
4. Would you buy such system (approx. 200 EUR)? (1 = no, 10 = absolutely)											
		1		1			1	3	6	8.6	12

enough information for whatever amount of parts they were looking for. Still, we will think about how we could include an optional hint system that guides explorers to parts they have not yet discovered.

On the question of whether they were missing descriptions, four mentioned a better description of color and texture, possibly not only for the area but more specifically at the location of the fingertip. This was especially apparent at the comparatively large area of the male figure's coat, for which several people expected more descriptions than just a single text covering the whole area.

4.4 Acceptance and Field of Application

A final set of questions concerned the acceptance and possible fields of applications. All test users found the presented technology to be meaningful in a museum setting, with an average ranking of 9.5 (see Table 4). However, not all would *rather* go to a museum if it was offering an IAG (average ranking, 8.2), as they would go to the museums in any case. Even less would consider it for home use (average ranking, 7.2) as they would not have space and time to use it. However, after telling them that the technology is very low-cost, possibly included in many future devices and that it could be extended to any objects, not just reliefs, six would buy it without hesitation and another four ranked it 8 to 9. They would like to use it for the annotation of plans; for object detection ("which bottle was the good wine?"); for photo exploration, geography, and education; and would like to see it also in schools or other educative institutions (e.g., at the zoo).

5 EVALUATION: STUDY 2

In order to expand on the findings presented at the ASSETS'16 conference [42], we instigated a second study with a broader range of participants across the access needs spectrum so that the wider application of the IAG could be tested. This study was conducted in London, UK, with a participant-led research method[17] over two testing sessions so that feedback from the first session could be immediately acted upon and the changes tested in the second. The participant-led method provided a greater depth of feedback as well as a more critical approach, which benefited our analysis. We also asked participants to identify their access needs rather than their type of disability, which provided the opportunity for a complimentary overlapping but differently orientated dataset to compare against the first study. The different group, spread of participants, research method, and approach resulted in significant progress both in the technical innovation of the IAG and analysis of its wider applicability.

The second study was conducted within the London Exploration Group of the ARCHES project.[18] The aim of the project is to develop online resources, software applications, and multisensory technologies to enable access to Cultural Heritage Sites within and beyond the project. The key aspect of the research is the participatory method, which was also used in this study. We started with an informal 5-hour session with 25 people across a wide range of sensory, learning, and cognitive impairments. After this first feedback, we implemented a structured evaluation that took place over the course of 2 days with 14 people (10 female, 4 male, aged 18–75 years, average age 45 years; see Table 5).[19]

The approach taken in this study was based on the concept that people cannot be neatly allocated into disability categories: they instead have access needs that may relate to one or more categories of disability or impairment. For example, a participant who would be typically classified as visually impaired may prefer the sort of one-to-one support typically associated with those who have learning difficulties. Likewise, visual impairment often accompanies learning difficulties and vice versa [34]. Asking participants about their preferred access needs and not their disability not only enables catering to those needs but also furthers the creation of a universal tool that can be enjoyed by everyone regardless of category or need.

The key needs that emerged from the participants in this second study were one-to-one support (7); audio description (4); captioning (3); simplified information (3); tactile books (3); Braille (1); and British Sign Language (1). See Figure 9 for further details. Though only one participant in each case required Braille and BSL, these were the main way in which they processed information and their preferred medium. Most participants were interested in museums, 7 going at least twice a year and 3 at least once a year. However, 4 rarely went to museums, if at all (see Table 5).

The same basic testing and evaluation method was used as in the first study, with one relief being tested. However, an important difference was the use of a participant-led research method. Based on the findings of the first study and taking into account the diversified nature of the second testing group, the questionnaire was expanded and improved. A total of 34 questions were asked, again most being on a 10-point Likert scale, 1 being the most negative, 10 the most positive ranking. These are summarized in Tables 5–7.

[17] Participatory research allows volunteers to be an active part in the process and have ownership of it [3, 53]. This is a vital aspect of the ARCHES (Accessible Resources for Cultural Heritage EcoSystems) project and we have adapted it for this study as well.

[18] ARCHES is a 3-year Horizon 2020–funded project that involves partners in heritage and technology across Europe. See http://arches-project.eu.

[19] Some of the responses include only 13 answers rather than 14. This was owing to the constraints within the participant-led mixed group (e.g., the need of a BSL interpreter having to leave earlier).

Table 5. Results of Study 2: Demographic Data

Sex	Male	Female										
	4	10										
Age (years)	0-9	10-17	18-29	30-39	40-49	50-59	60-69	70-79				
			3	5		2	3	1				
Why do you visit museums? (You may tick more than one)	Don't visit any	For educational purposes	For enjoyment purposes									
	1	8	11									
How often do you visit a museum per year?	Never	< 1x year	≥ 1x a year	≥ 1x a month	≥ 1x a week							
	2	2	3	5	2							
How do you visit the museum? (You may tick more than one)	Friends	Project	Alone	Family	Support	Other	no visit	School				
	7	6	4	3	2	1	1					
What technology do you use? (You may tick more than one)	iPad	iPhone	W. Laptop	A. Phone	A. Tablet	Kindle	MacBook	W. Tablet	W. Phone	Dictaphone	Other	None
	8	7	7	4	3	2	1	1	1	1	2	2
Have you used tactile reliefs before?	No	Yes	I don't remember									
	9	3	1									

A ... Android W ... Windows

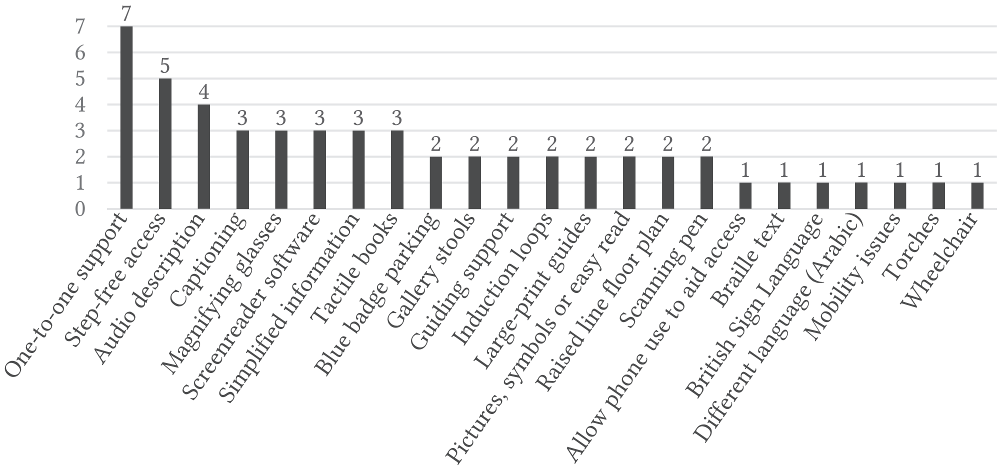


Fig. 9. Results of Study 2: "Please indicate your access needs (You may tick more than one)." 13 responses.

In response to participant feedback in the first testing sessions, several additions and improvements had been made to the IAG: a training mode (see Section 3.4.6), additional sound cues (see Section 3.4.4), additional visual cues and text captioning of audio descriptive content (see Section 3.4.5). These were then tested in the second session. The hierarchical exploration mode (see Section 3.4.2) was not used, and users had direct access to the detail areas of the figures. The introduction text was also not used, as this was replaced by one-on-one training in the training mode. Each participant spent about 2 to 5 minutes getting accustomed to the IAG in training mode, then about 6 to 12 minutes exploring.

5.1 General Impression

In general, the response was positive: it was "fun," the multimodal approach "made the information easier to process," and participants felt they "connected" to the painting and got to "explore it more deeply." Several people said that they liked having the background information about the painting, as this helped put it in context. The tactile element helped them pick out details that they would otherwise have missed. Participants particularly responded to the detail in the base and clothing where the shapes were very distinct and intricate. However, it took time and focused support for the participants to understand the process of navigating the IAG. After the first session, one participant requested a clear training mode. The training mode that was created was well received but it only allowed participants to explore the gestures—it did not give them the step-by-step guide that was needed. Also, it became clear that audio or descriptive instructions do not work for everyone: "With learning disabled people you can't just tell them what to do, you need to show them what to do." This comment from a support assistant shows that there is a need to create a more social training mode, with videos and images. In particular, people with learning difficulties took more time understanding the process. They had not used tactile elements previously and were therefore not used to it.

Overall, it seemed that those with the most severe level of visual impairment appreciated the system most, as people with sight do not have equivalent dependence on touch and audio. Of course, it is important to note that those with visual impairments are often more used to using touch as a means of learning and orientation; thus, this is not necessarily a direct or fair comparison. Some participants found it easier than others to get used to navigating by touch and audio.

5.2 Interface

As before, the system was designed to more closely resemble a kiosk in a museum; therefore, an introductory text should be sufficient for use. Our original intention was to use kiosk mode, with initial mandatory instructions as used in the first study. However, after the first session, it was clear that participants needed (and were given) one-to-one support. They consistently indicated that a training mode, taking them through the gestures, would be helpful. As a result, a training mode was created for the second session (see Section 3.4.6). This training mode allowed the instructor to layer the information necessary to use the IAG. The addition of this mode made a significant difference, especially to the accuracy of the participants' gestures and how quickly they were able to pick them up. However, though they appreciated this addition, most indicated that instructions accompanied by a video would give them the best chance of navigating the system in true kiosk mode. Further developments to this end and testing in a museum environment are required before the IAG can truly work in kiosk mode.

Having received one-to-one support and testing the setup, nine people rated the question "How did you find using the IAG?" above 8, with all but three giving it a ranking of 7 and above (Table 6, Question 1). Participants were giving feedback such as that the IAG "allowed me to precisely connect what I was feeling to the description of the painting" and that "the description provides extra detail that is engaging and interesting." Some liked it because they "could actually feel the painting" (2×), they liked "the voice and liked exploring the portrait by touch whilst the voice told you what you were feeling," it "gives you an all round experience regardless of your access needs" and it "Helps to explore the painting more deeply at an interactive level." For others, it "took time to adapt to it and getting used to the directions" but had a "very positive experience once up and running." One user found that there are "Not very good tactile markers to find your place," and

Table 6. Results of Study 2: Questions Concerning the Technology of the Interactive Audio Guide

1	2	3	4	5	6	7	8	9	10	avg.	sum
1. How did you find using the IAG? (1 = not good, 10 = very good)											
				3		2	2	3	4	8.0	14
2. How understandable did you find the instructions? (1 = not, 10 = very)											
		3			3	2	2	2	2	6.8	14
3. How simple were the gestures for you? (1 = not, 10 = very)											
	1	1		2	1	2	2	3	2	7.0	14
4. How important is it, that the audio guide is triggered only by certain gestures and does not talk all the time? (1 = not, 10 = very)											
1					1	1	3	2	5	7.6	14
5. How easy was it to select certain parts? (1 = not, 10 = very)											
	1	1		2		3	3	3	1	6.9	14
6. How important is the tactile element for you? (1 = not, 10 = very)											
1				3		2	2	1	5	7.5	14

another would like to see "more visual content especially an image of the picture to make a comparison." Over half found the instructions easy to understand (ranking above 7) and all but three gave a ranking of 6 and above for this (Table 6, Q2). Some found that the starting instructions are "not clear enough" and that they needed "step-by-step instructions on how to use it." Among those that gave the lower rankings on both questions were people who had not listened to the instructions as well as (and partly in combination with) having learning or cognitive impairments that make retention of information more difficult. These participants needed and received one-to-one support from the instructor and, in a museum setting, would frequently be accompanied by a support assistant.

Half the participants mastered the interface through the training mode and could reproduce all gestures without any intervention from the researcher. As with the previous study, others needed guidance or slight manual corrections of their hands. Two participants had difficulties using the gestures and indicated that they would prefer a "keyboard" or "button" alternative. A support assistant also pointed out that some disabled people might have additional conditions, such as Parkinson's or arthritis. However, generally, when asked how easy it was to perform the gestures, nine participants rated 7 or higher (Table 6, Q3). Comments included, "it's really instinctive," "cool!" and "I like the fist!" The design goal, to only have the system play audio when it is explicitly requested by the user was again supported by the responses: 7 participants gave a ranking of 9 or 10, and all but two gave a ranking of 7 and above (Table 6, Q4). The repeated feedback was that people wanted to explore the relief and the information at their own pace and not be bombarded by constant audio information.

5.2.1 Off-Object Gestures. In the first session, the off-object gestures were the ones that participants struggled with most. Getting the correct height, keeping the hand flat and being able to spread the fingers without distorting the hand shape were all common problems. Some conditions affect the physiology of the hands – e.g., shorter, thicker fingers – which meant that the gestures were harder to make for those participants and that it was harder for the algorithms to detect the gestures. Those with the most visual impairment sometimes found it hard to know when they were at the correct height, making the correct shape, and if their hand was flat, as they had no visual or tactile cue for checking and correcting this.

Taking these difficulties into account, we created a training mode for the second session (see Section 3.4.6), and added a *fire-crackling sound* to guide participants to the correct height to trigger the audio (see Section 3.4.4). This allowed participants to practice making the gestures and gave audio feedback when these were performed correctly. This gave the participants a standard to refer back to when they were navigating the IAG. Both additions significantly improved the overall experience and meant that less guidance and physical adjustments were needed from the instructor.

Participants with joint difficulties and smaller, thicker hand shapes pointed out how they would prefer a button mechanism at the bottom of the relief and/or an alternative set of gestures that worked on the movement of the fist rather than specific hand shapes (i.e., one shake for item 1, two for item 2, etc.). One participant even had to support the gesture-making hand with her other arm as she got easily tired of waiting until the camera read her hand gesture. These alternatives would also be useful for those with learning or cognitive impairments, as there is not as much to learn and does not require remembering as long a sequence of instructions.

5.2.2 On-Object Gestures. Using the pointing gesture with the index finger became easier for the participants after clear instructions and corrections. The main issue was that participants did not close the fist completely and left the thumb extended. The participants also struggled with the correct angle. The tendency was to position the finger vertically, which was uncomfortable for them. In addition, the participants tended to press on the relief harder rather than gliding over it. At this current stage, the silhouette-based hand detector is too sensitive and will need improvement. Participants relaxed after being reassured that they were doing the right gesture. A more social element would give the participant the confidence needed to explore the relief.

Another limitation detected after the first session was that BVI participants were unsure where individual sections or the relief itself stopped. As a result, we implemented a “rain sound” that is triggered as the hand approaches a new section or boundary of the relief. This addition helped BVI participants significantly but also some of those with cognitive and learning impairments.

One of the aspects that needs to be thought through further is the automatic necessity of the additional sound design (see Section 3.4.4). Participants with learning difficulties were calmly guided through the process of testing and largely ignored the sounds, but for independent testing, these sounds might become distressing. Being able to switch the function on and off should therefore become an option for the user.

5.3 Content

When asked whether the IAG helped in gaining a sense of the whole painting, nine gave a rating above 6 (Table 7, Q2). This increased to twelve participants when asking about the details of the painting (Table 7, Q3). All but one of the participants felt that the level of detail was just right; one participant wanted more detail as the individual was particularly interested in exploring the painting at a deep level and was also highly competent at navigating the system (Table 7, Q7). The physical replication of the details such as “the way that the different shapes represented the pattern of the clothes” were particularly popular and received high praise from most of the participants, with an average ranking of 6.7 (Table 7, Q1). There was an overall desire for color on the relief, regardless of whether there was a copy adjacent to the station.

Overall, 11 of the 14 were happy with the amount of description (Table 7, Q4). One participant went through the parts but did not listen to the descriptions, as the individual was short of time and wanted to have tried each of the aspects before leaving. One of the challenges with the content was (a) testing it on people with hearing loss and (b) testing it with participants with learning difficulties. During the first session, we presented subtitles in a text editor in an ad-hoc attempt to make it accessible for one participant with hearing loss, although in a very small font. For the second session, proper subtitles were implemented (see Section 3.4.5). This enabled us to test the IAG with D/deaf participants. People with learning difficulties used this as well as it helped them

Table 7. Results of Study 2: Questions Concerning the Content, i.e., Relief and Audio Description of "The Kiss" and the Application of This Technology

1	2	3	4	5	6	7	8	9	10	avg.	sum
1. What was your impression of the relief? (1 = not good, 10 = very good)											
1				4		4	2	1	2	6.7	14
2. How well could you get a sense of the whole painting? (1 = not, 10 = very)											
1			3	1	3	3	1	1	1	6.0	14
3. How well could you get a feeling for the details of the painting? (1 = not, 10 = very)											
1			1		4		2	4	2	7.2	14
4. How happy were you with the number of parts described? (1 = not, 10 = very)											
		1	2		1	2	3	3	2	7.3	14
5. Would you find it useful to use such technologies in a museum? (1 = not, 10 = very)											
1						2	1	1	8	8.0	13
6. Would you rather go to a museum that offers such audio guides? (1 = does not matter, 10 = much rather)											
1				1	2		3		6	7.3	13
7. How would you rate the amount of detail in the relief?											
too little		OK		too much							
1		12									13
8. Would you prefer different types of description? (e.g. music; poetry; storytelling)											
No		Maybe		Yes		I don't know		I don't mind			
3				10				1			14

"focus" more on the content. Further comments from participants suggest that inverted colors would ease the reading and "make the text pop out more." This was a suggestion from a support assistant with dyslexia and is also typically found with people who have a visual impairment.

Nevertheless, several participants still found the language difficult and too "academic" and "formal." This further supports the abovementioned statistic that 23.1% of our participants (3 out of 14, Figure 9) have an access need of simplified information. Here, the differences in the level of education attained was noticeable in how accessible the participants found the information: broadly speaking, the lower the level of education, the more difficult the participant found the information. Overall, the language needs to be simplified but there may also be a case for providing a setting in which you can choose your level of description with an option consciously aimed at those who require simplified content. This could also include symbols and pictures to support the content, which would be a valuable addition for those with cognitive and learning impairments. In contrast to the first study, this study was conducted in the United Kingdom and the language used was English. One participant had severe difficulties understanding the speaker and translated text "maybe it's his heavy accent...I don't know."

Therefore, if this application is to be used beyond Austria, text creation and audio recordings by native speakers would need to be considered.

The majority (71%, 10 out of 14; Table 7, Q8) said that they would enjoy an alternative, more creative form of description, such as music, poetry, or storytelling. Only 3 participants did not like or see the need for this. The most consistent suggestion was a storytelling approach that included a first-person narrative. Such a narrative would also help those who need simplified language, as they can find it easier to relate to narrative content rather than the purely factual. As one participant said, it needs to be "more informal and creative rather than just giving the facts in quite a scientific way. At the moment a bit robotic, make it a bit more human." Having music as part of the descriptive process was also popular "as it creates an atmosphere." Finally, a strong case was made for having audio content related by BSL with a signer in a pop-up video: this is especially important for those who have BSL as their first language. Ideally, content would be created in BSL itself, in collaboration with a signer, rather than necessarily being a literal translation of the text. This would increase the quality of the language for users.

5.4 Acceptance and Field of Application

All but 3 participants had never tried a similar technology before. This number is not necessarily surprising, as a participant said "I would have never tried it. I am used to the visual learning and that's it." Compared to similar experiences, participants said: "Good because it is live but not enough explanations compared to other reliefs, e.g., Living paintings[20]"; "Great improvement on others"; "Very different experience (never done paintings)." A total of 62% stated that this technology would be very useful within the museum environment (8 out of 13 ranked this 10; Table 7, Q5). Only one participant said that it would not be useful. It is interesting to note how these responses map onto the users' experience of technology in general: all used a smartphone, all but one used a tablet of some form, and 8 out of 14 had a laptop or MacBook (Table 5). Therefore, the feedback on the IAG specifically and this sort of technology in museums generally was coming from participants familiar with using technology in everyday life. Thus, it is important not to assign unfamiliarity with technology in general as a significant factor, especially in analyzing the more critical comments or noted problems in navigating the IAG. In the future, more time and a clearer, layered and social tutorial should be offered to them and would support a confident and competent use of the IAG.

Six participants said that they would strongly (10 points on the Likert scale) rather go to a museum if this would be available to them, three who rated 8 and three who remained rather neutral between 6 and 5 (Table 7, Q6). Though half the participants visit museums at least once a month, we had 2 who visited museums less than once a year and 2 who indicated they would never visit a museum. It is also important to note the conditions under which participants visit museums: the half who visit infrequently are much less likely to go outside of a project or organized trip and require support in physically getting to the museum as well as one-to-one support once on site (Table 5). This highlights that even with high-tech solutions to accessibility, such as the IAG, there remains a social element that is vital, especially for those who have the greatest access needs. However, it still remains the case that innovations such as the IAG open up new potential for these participants in seeing museums as relevant for and accessible to them. We just need to make sure the social support is also in place.

Several participants particularly raised how good this system would be for children and in a school or education setting – they were keen for this aspect to be developed. Given the bulky and heavy nature of the IAG, especially with the HP Sprout, a more portable version may need to be developed if this is to be practical, unless such devices are already available there.

5.5 Important Qualifications to the Data

People with learning and cognitive impairments are often keen to please and to give the "right" answer in a question situation as well as not wanting to show that they do not know or understand. For example, one participant had answered the age category in the questionnaire but this was corrected by the person's support worker as, in reality, the participant did not know or could not remember the age but did not want to admit this. Therefore, without careful analysis, the data can be unreliable and misleading. To offset this and avoid any potential skewing, we have complemented the answers given by these participants with the data and observations of the

[20] Living paintings is a charity that provides touch to see books for BVI people; see http://www.livingpaintings.org.

examiners and the participants' support assistants. We can therefore have greater confidence that the data used is accurate and a fair representation of participants' experiences of the IAG.

It is common for disabled people in general, and among our participants specifically, that people do not fit solely in one category, such as BVI or Hard of Hearing (HoH) – they have intersectional impairments and needs, often the combination of a sensory impairment and learning/cognitive impairment. Therefore, it is important that this is taken into account when reading the raw data and drawing conclusions on the use of the IAG for specific groups. We have taken care to build this intersectionality into our analysis of the data and our conclusions.

6 FUTURE WORK

As the access needs indicate, if we wish to expand on the usability of the device, changes will have to be considered: we will critically review the selected gestures and the interface design with the different groups. Currently, the HP Sprout provides an all-in-one solution that is readily available and can be used directly. The algorithms are lightweight and may even run on embedded systems, such as the ORBBEC Persee[21], the first depth camera with an integrated computer. Although the system remained stable during several consecutive days without restart, accumulated sensor drift made it less reliable. A future implementation might overcome this limitation by performing a background calibration right before a new user is detected. The new setup, using the HP Sprout together with changes in the algorithms, already has improved gesture detection, and the addition of audio cues give additional feedback to help in forming the correct gestures.

For further improvements, we will investigate other finger-tracking solutions, the inclusion of the RGB sensor, alternative sensor placement, or multiple sensor setups to better observe the fingertip, especially near the top edge of the working space, and to relax the need for flat finger gestures that still cause a problem for some, even through the improvements of the new setup. Likewise, we will consider alternative input possibilities: different gestures, touchscreen input, HP Sprout's included touch mat, voice commands, or more traditional usage of buttons or a keyboard at the bottom. Thanks to the participative research methodology, the final user interface can be designed in a user-elicited process, possibly with the help of a Wizard of Oz study. Based on the feedback on the content, we will have to edit the text and consider more creative outputs.

Apart from these considerations, a comprehensible tutorial mode seems to be missed by most participants in the second study. This has to be investigated further, as well as the possibility to quickly set individual options and preferences for people with different access needs, maybe in the form of a QR coded badge that is shown to the camera first.

Further, we plan to investigate new interaction possibilities: these include multiple information layers, multifinger gestures, educative games, sonification of color, and an extension to exchangeable 3D objects with arbitrary and dynamic placement.

Since the inclusion of the much broader target audience of people with differences and difficulties associated with perception, memory, cognition, and communication, we are just beginning to understand the additional and sometimes contradictory demands on the system. Together with the extended capabilities of the new setup, especially the projector and additional touchscreen, this opens up lots of possibilities for future research and development.

6.1 Multisensory Experience

The IAG has the potential to further enrich the experience for non-BVI users. Therefore, we included another set of questions in Evaluation Study 2 regarding different ideas to further enrich multisensory experiences. Five participants (36%) think that a projection would very much add to

[21] Orbbec 3D's website is at https://orbbec3d.com/.

Table 8. Results of Study 2: Questions Concerning Multisensory Experience

1	2	3	4	5	6	7	8	9	10	avg.	sum
1. How did you find the material? (1 = not good, 10 = very good)											
1		1				1	3	2	6	8.1	14
2. How would you rate the added value of projections on the relief? (1 = not good, 10 = very good)											
			1	1		2	5		5	8.1	14
3. What would you like to project on the relief? (You may tick more than one option)											
Original Color		Simplification		High Contrast		Animation		Nothing			
13		4		3		3					
4. Would you like to have a haptic response? (i.e. vibration)											
No		Maybe		Yes		I don't know		I don't mind			
1		5		6		1		1			14
5. Would you like color/texture represented by sound as you move your fingers over the relief?											
No		Maybe		Yes		I don't know		I don't mind			
2		5		6				1			14
6. Would you prefer a relief or a 3D model?											
I prefer a relief				I prefer a 3D		I don't know		I don't mind			
4				8		1		1			14

the value of the relief (i.e., ranked 10; Table 8, Q2). When asking participants about possible projections onto the relief, 13 participants indicated that they wanted the original painting; 4 wanted a more simplified version; 3 wanted high contrast; and 3 asked for an animated version of the painting (Table 8, Q3). None of the participants said that they would rather have it blank. For future applications, these results will have to be taken into consideration and implemented, especially if aiming to widen the application beyond BVI users.

When asked about a haptic response (such as vibration), 6 (43%) indicated that they would like to have it (Table 8, Q4). Vibration could help participants with orientation but also help those who have both hearing and vision loss: such intersectional needs are currently rarely addressed and providing this would set the IAG apart. Participants who currently use technology such as smartphones (particularly those who use the accessibility features for low vision and hearing loss) were already familiar with this technology and keen to see it used in this context.

Regarding how suitable the users from the participatory research group found the material on a scale from 1 to 10 (Table 8, Q1) and if they would prefer a different one, some participants indicated that they wanted something rougher and matte textured, such as clay. This was more common to those with visual impairment but also was mentioned by other participants. Another indicated wanting the material to be chosen according to the artwork, for example, rougher material replicating brushstrokes for Van Gogh and smooth, shiny materials in the case of "The Kiss" to represent the gold and metallic elements. This feedback indicates a preference for the creation of a "collage" tactile effect and almost a move to the relief itself becoming an artistic object in its own right. This is an interesting avenue to consider.

The majority of participants (11, 79%, yes and maybe) were interested in sonifying the relief – representing colors and materials through sound (Table 8, Q5). Six (43%) definitely wanted this feature added and were excited by the idea – "it would be the final piece of the puzzle, making it a complete experience." Such a development has the potential to enable participants to relate to paintings based on personal experience. Even participants with complete sight loss are likely to have had some sight at some stage in their life and therefore have some memory and knowledge of color [35]. For those who have been blind since birth, colors can be translated into temperatures (blue equaling cold, red as hot, etc.); thus, this feature has a wide applicability. Again, this might be introduced as an on/off switch option to give the user more control of the experience.

Eight participants indicated that they would like to test a 3D model compared to a 2.5D model (Table 8, Q6). Three participants with learning disabilities had difficulties exploring the faces of the depicted figures and identifying their gender. By exploring it from every angle, this might solve the issue. Even though the production of such a model is rarer and more costly, depending on the nature of the artwork, this might be preferable. However, the use of the IAG would have to be reconsidered and the system redesigned to full 3D model tracking so that these can be freely turned and have accessible touch-sensitive parts all around.

6.2 Generalization to Other Works of Art

The formal evaluation of this work was based on a single tactile relief, but we are confident that it generalizes to all kinds of tactile, and even nontactile, objects. From a technical point, gesture detection works on any surface as long as it can be scanned by the depth sensor. During development, we tested it with several reliefs. It also works with plain surfaces, such as a piece of paper. However, for plain objects easier touch input devices exist, for example, the touch mat already integrated in the HP Sprout workstation.

While this work focused on 2.5D tactile reliefs of paintings, the techniques should generalize to any 2D, 2.5D, and 3D object. For flat objects only, our relief calibration algorithm has to be exchanged with a state-of-the-art planar object tracker (e.g., [30]). For 3D objects, a faster tracking algorithm needs to be used, as the user might want to rotate the object during exploration, and possibly a better hand detection algorithm is needed.

Contentwise, it should generalize as well. Throughout the world, a lot of paintings have already been converted to tactile relief, including our own work;[22] this demonstrates that tactile reliefs can be made from a wide variety of images. Clearly, some paintings are more difficult to convey and to understand than others, and complicated or very large scenes need to be simplified or concentrated on a few close-ups. But with the additional aid of the IAG, we believe that this technology is suitable for a wide variety of artwork. The limits are only in the size of the areas that still can be reliably touched, which can be handled by carefully authoring the interactive content.

7 CONCLUSION

We presented a gesture-controlled interactive audio and text guide that allows access to location-specific content, triggered directly with the fingertips on relief surfaces, and demonstrated its real-world usability. In contrast to our first setup, the HP Sprout platform offers a commercially available and museum-ready option that minimizes setup time and has a nice look and feel, although sensor placement is not optimal and the effective workspace and accuracy might suffer a bit. It is interesting that the purely silhouette-based hand-detection algorithm performs really well in a real-world situation, at least when the set of gestures is carefully selected. After a thorough refinement process, we arrived at a good selection of gestures that are both convenient for most users and technically feasible.

Furthermore, we showed that through a more diverse user group and the use of a participant-led method of testing, the development of new technologies aimed at better accessibility is not only more inclusive, but also more efficient.

The majority of the 27 test users of the two formal evaluation sessions found it useful and worth further development. It seemed to be especially interesting for fully blind people, who want to access the paintings in detail and to do this autonomously. The quantitative rankings are consistently

[22] "The Kiss" was our seventh relief of paintings; four others have been documented in [43]. We worked on reliefs from stereoscopic photographs [33] and from 3D objects [44].

lower in the second study and are spread more widely between high and low. This is no surprise, as the IAG was primarily created for BVI people and we only just began to include features for a wider audience. We also have the impression that the system is less suitable for people who rate quantity over quality, i.e., who want to experience as many art pieces as possible, for whom a short description of the content is already sufficient. These people seemed too impatient during the sessions. With over 11 minutes of audio and 20 different locations to discover, our interactive installation goes clearly in the other direction: “less pieces, more detail,” a quote we heard from participants who have been blind for a long time.

The main findings from our two studies are the key importance of the multisensory nature of the task; the ability to personalize systems for specific needs, e.g., sign language, captions, color, simplified language; and the limitations of needing precise gestures to control the IAG, especially for those with mobility issues. Especially the second user study highlighted the need for a simple and socially engaging tutorial to teach users how to use the system. While this finding may perhaps be obvious in hindsight, it is a very useful reminder of the importance of compelling training techniques for new access technologies.

Beyond the visually impaired participants, participants with different access needs said that the tactile element “explains the painting by getting people involved and helps them understand more. We all want to touch things!,” “It helps with low attention span – taps into curiosity,” and “It gives me a deeper understanding of the piece.”

The developed prototypes are targeted at a museum setting, but the low-cost sensor hardware and the fact that these sensors will soon be integrated in laptop and mobile devices makes it very attractive for home use or in educational institutions. Overall, the work provides an approach that may not only reduce barriers to the accessibility of visual art for people with many different disabilities but may provide an entirely new modality that helps all museum visitors appreciate art in an exciting new way.

ACKNOWLEDGMENTS

Special thanks to all volunteers testing the application, the London Exploration Group, V&A Museum (Suzana Skrbic in particular), the Wallace Collection, Florian Rist and the Institut für Kunst und Gestaltung of the Vienna University of Technology, HP Austria GmbH for initial tests with the HP Sprout, Ruth List and Felicia Kerschbaum of Economica, Brigitte Hauptner and Susa Wögerbauer of Belvedere, and the reviewers for their great contributions.

REFERENCES

[1] Elisabeth Salzhauer Axel and Nina Sobol Levent (Eds.). 2003. *Art Beyond Sight: A Resource Guide to Art, Creativity, and Visual Impairment.* AFB Press, New York, NY.

[2] Catherine M. Baker, Lauren R. Milne, Jeffrey Scofield, Cynthia L. Bennett, and Richard E. Ladner. 2014. Tactile graphics with a voice: Using QR codes to access text in tactile graphics. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS'14).* ACM, New York, NY, USA, 75–82. https://doi.org/10.1145/2661334.2661366.

[3] Colin Barnes. 2003. What a difference a decade makes: Reflections on doing ‘emancipatory’ disability research. *Disability & Society* 18, 1, 3–17. https://doi.org/10.1080/713662197.

[4] Edward P. Berlá and Lawrence H. Butterfield Jr. 1977. Tactual distinctive features analysis: Training blind students in shape recognition and in locating shapes on a map. *The Journal of Special Education* 11, 3, 335–346. http://journals.sagepub.com/doi/10.1177/002246697701100309.

[5] Gary Bradski and Adrian Kaehler. 2008. *Learning OpenCV: Computer Vision with the OpenCV Library.* O'Reilly Media, Inc., Sebastopol, CA.

[6] Diana Brinkmeyer. 2014. Museum ohne Grenzen – Multimediale Anwendungen und Barrierefreiheit in der Berlinischen Galerie. In *Kunstvermittlung 2.0: Neue Medien und ihre Potenziale*, Andrea Hausmann and Linda Frenzel (Eds.). Springer Fachmedien Wiesbaden, Wiesbaden, 105–121. https://doi.org/10.1007/978-3-658-02869-5_7.

[7] Anke Brock, Samuel Lebaz, Bernard Oriola, Delphine Picard, Christophe Jouffrais, and Philippe Truillet. 2012. Kin'touch: Understanding how visually impaired people explore tactile maps. In *Extended Abstracts on Human Factors in Computing Systems (CHI'12)*. ACM, 2471–2476.

[8] Francesco Buonamici, Rocco Furferi, Lapo Governi, and Yary Volpe. 2015. Making Blind People Autonomous in the Exploration of Tactile Models: A Feasibility Study. In *Proceedings of the 9th International Conference on Universal Access in Human-Computer Interaction. Access to Interaction (UAHCI'15), Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Part II*. Springer International Publishing, 82–93. https://doi.org/10.1007/978-3-319-20681-3_8.

[9] Yang Chen and Gérard Medioni. 1992. Object modelling by registration of multiple range images. *Image and Vision Computing* 10, 3, 145–155. https://doi.org/10.1016/0262-8856(92)90066-C.

[10] F. D'Agnano, C. Balletti, F. Guerra, and P. Vernier. 2015. Tooteko: A case study of augmented reality for an accessible cultural heritage. Digitization, 3D printing and sensors for an audio-tactile experience. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XL-5/W4. 207–213.

[11] Sinjin Dixon-Warren. n.d. Inside the Intel RealSense Gesture Camera. (n.d.). http://www.chipworks.com/about-chipworks/overview/blog/inside-the-intel-realsense-gesture-camera.

[12] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. 2010. Model globally, match locally: Efficient and robust 3D object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*. IEEE, 998–1005.

[13] Polly K. Edman. 1992. *Tactile Graphics*. American Foundation for the Blind, New York, NY.

[14] Yvonne Eriksson. 1999. How to make tactile pictures understandable to the blind reader. In *65th IFLA Council and General Conference*. Bangkok, Thailand.

[15] Rocco Furferi, Lapo Governi, Yary Volpe, et al. 2014. Tactile 3D bas-relief from single-point perspective paintings: A computer based method. *Journal of Information & Computational Science* 11, 16, 5667–5680.

[16] Rocco Furferi, Lapo Governi, Yary Volpe, Luca Puggelli, Niccolóš Vanni, and Monica Carfagni. 2014. From 2D to 2.5D i.e. from painting to tactile model. *Graph. Models* 76, 6, 706–723. https://doi.org/10.1016/j.gmod.2014.10.001.

[17] Giovanni Fusco and Valerie S. Morash. 2015. The tactile graphics helper: Providing audio clarification for tactile graphics using machine vision. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS'15)*. ACM, 97–106. https://doi.org/10.1145/2700648.2809868.

[18] John A. Gardner and Vladimir Bulatov. 2006. Scientific diagrams made easy with IVEO^{TM}. In *ICCHP 2006, Lecture Notes in Computer Science*, Klaus Miesenberger, Joachim Klaus, Wolfgang L. Zagler, and Arthur I. Karshmer (Eds.), Vol. 4061. Springer, Berlin, 1243–1250. https://doi.org/10.1007/11788713_179.

[19] Timo Götzelmann. 2016. LucentMaps: 3D printed audiovisual tactile maps for blind and visually impaired people. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS'16)*. ACM, New York, NY, USA, 81–90. https://doi.org/10.1145/2982142.2982163.

[20] Paolo Gualandi and Loretta Secchi. 2000. Tecniche di rappresentazione plastica della realtà visiva. In *Toccare L'arte: L'educazione Estetica Di Ipovedenti e Non Vedenti*, Andreas Bellini (Ed.). Armando Editore, 49–98.

[21] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, and Jianwei Wan. 2014. 3D object recognition in cluttered scenes with local surface features: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 11, 2270–2287.

[22] Robert J. K. Jacob. 1991. The use of eye movements in human-computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems* 9, 2, 152–169. https://doi.org/10.1145/123078.128728.

[23] Shaun K. Kane, Brian Frey, and Jacob O. Wobbrock. 2013. Access lens: A gesture-based screen reader for real-world documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 347–350.

[24] Shaun K. Kane, Jacob O. Wobbrock, and Richard E. Ladner. 2011. Usable gestures for blind people: Understanding preference and performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, New York, NY, USA, 413–422. https://doi.org/10.1145/1978942.1979001.

[25] Branko Karan. 2015. Calibration of kinect-type RGB-D sensors for robotic applications. *FME Transactions* 43, 1, 47–54.

[26] Roberta L. Klatzky, Nicholas A. Giudice, Christopher R. Bennett, and Jack M. Loomis. 2014. Touch-screen technology for the dynamic display of 2D spatial information without vision: Promise and progress. *Multisensory Research* 27, 5–6, 359–378.

[27] Nikolaos Kyriazis and Antonis A. Argyros. 2014. Scalable 3D tracking of multiple interacting objects. In *IEEE Computer Vision and Pattern Recognition (CVPR'14)*. IEEE, Columbus, OH, USA, 3430–3437. https://doi.org/10.1109/CVPR.2014.438.

[28] Steven Landau and Karen Gourgey. 2001. Development of a talking tactile tablet. *Information Technology and Disabilities* 7, 2 (2001).

[29] Nina Levent and Alvaro Pascual-Leone (Eds.). 2014. *The Multisensory Museum: Cross-Disciplinary Perspectives on Touch, Sound, Smell, Memory, and Space*. Rowman & Littlefield Publishers, Lanham, MD.

[30] Pengpeng Liang, Yifan Wu, and Haibin Ling. 2017. Planar object tracking in the wild: A benchmark. arXiv: 1703.07938v1

[31] D. G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision*, Vol. 2. 1150–1157. https://doi.org/10.1109/ICCV.1999.790410.

[32] Elad Moisseiev and Mark J. Mannis. 2016. Evaluation of a portable artificial vision device among patients with low vision. *JAMA Ophthalmology* 134, 7, 748. https://doi.org/10.1001/jamaophthalmol.2016.1000.

[33] Moritz Neumüller and Andreas Reichinger. 2013. From stereoscopy to tactile photography. *PhotoResearcher* 19, 59–63.

[34] Royal National Institute of Blind People. n.d. Across the spectrum: learning disability and and sight loss. Retrieved February 3, 2018 from http://www.rnib.org.uk/services-we-offer-advice-professionals-nb-magazine-health-professionals-nb-features/across-spectrum.

[35] Royal National Institute of Blind People. n.d. Key information and statistics. Retrieved February 3, 2018 from http://www.rnib.org.uk/knowledge-and-research-hub/key-information-and-statistics.

[36] I. Oikonomidis, N. Kyriazis, and A. Argyros. 2011. Efficient model-based 3D tracking of hand articulations using kinect. In *BMVC'11*. BMVA.

[37] John Stewart Olson and Angelo Raymond Quattrociocchi. 2015. Method and apparatus for three-dimensional digital printing. (June 23 2015). US Patent 9,061,521.

[38] S. O'Modhrain, N. A. Giudice, J. A. Gardner, and G. E. Legge. 2015. Designing media for visually-impaired users of refreshable touch displays: Possibilities and pitfalls. *IEEE Transactions on Haptics* 8, 3, 248–257. https://doi.org/10.1109/TOH.2015.2466231.

[39] Susumu Oouchi, Kenji Yamazawa, and Lorreta Secchi. 2010. Reproduction of tactile paintings for visual impairments utilized three-dimensional modeling system and the effect of difference in the painting size on tactile perception. In *ICCHP'10, Part II, Lecture Notes on Computer Science*, Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer (Eds.), Vol. 6180. Springer, Berlin, 527–533. https://doi.org/10.1007/978-3-642-14100-3_79.

[40] Ben Piper, Carlo Ratti, and Hiroshi Ishii. 2002. Illuminating clay: A 3-D tangible interface for landscape analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'02)*. ACM, New York, NY, USA, 355–362. https://doi.org/10.1145/503376.503439.

[41] Andreas Reichinger, Anton Fuhrmann, Stefan Maierhofer, and Werner Purgathofer. 2016. A concept for re-usable interactive tactile reliefs. In *ICCHP'16, Part II, Lecture Notes in Computer Science*, Klaus Miesenberger, C. Bühler, and Petr Penaz (Eds.), Vol. 9759. Springer, Berlin, 108–115. https://doi.org/10.1007/978-3-319-41267-2_15.

[42] Andreas Reichinger, Stefan Maierhofer, Anton Fuhrmann, and Werner Purgathofer. 2016. Gesture-based interactive audio guide on tactile reliefs. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS'16)*. ACM, New York, NY, USA, 91–100. https://doi.org/10.1145/2982142.2982176.

[43] Andreas Reichinger, Stefan Maierhofer, and Werner Purgathofer. 2011. High-quality tactile paintings. *Journal of Computing and Cultural Heritage* 4, 2, Article 5 (2011), 13 pages. https://doi.org/10.1145/2037820.2037822.

[44] Andreas Reichinger, Moritz Neumüller, Florian Rist, Stefan Maierhofer, and Werner Purgathofer. 2012. Computer-aided design of tactile models. In *ICCHP'12, Part II, Lecture Notes in Computer Science*, Klaus Miesenberger, Arthur Karshmer, Petr Penaz, and Wolfgang Zagler (Eds.), Vol. 7383. Springer, Berlin, 497–504. https://doi.org/10.1007/978-3-642-31534-3_73.

[45] Jonathan Rix and Ticky Lowe. 2010. Including people with learning difficulties in cultural and heritage sites. *International Journal of Heritage Studies* 16, 3, 207–224.

[46] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, et al. 2015. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3633–3642.

[47] Huiying Shen, Owen Edwards, Joshua Miele, and James M. Coughlan. 2013. Camio: A 3D computer vision system enabling audio/haptic interaction with physical objects by blind users. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS'13)*. ACM, 41.

[48] Lei Shi, Ross McLachlan, Yuhang Zhao, and Shiri Azenkot. 2016. Magic touch: Interacting with 3D printed graphics. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS'16)*. ACM, New York, NY, USA, 329–330. https://doi.org/10.1145/2982142.2982153.

[49] Aaron Staranowicz, Garrett R. Brown, Fabio Morbidi, and Gian Luca Mariottini. 2013. Easy-to-use and accurate calibration of RGB-D cameras from spheres. In *Pacific-Rim Symposium on Image and Video Technology*. Springer, 265–278.

[50] Brandon Taylor, Anind Dey, Dan Siewiorek, and Asim Smailagic. 2016. Customizable 3D printed tactile maps as interactive overlays. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS'16)*. ACM, New York, NY, USA, 71–79. https://doi.org/10.1145/2982142.2982167.

[51] Emma Teschner. 1909. Gustav Klimt im Ruderboot. Retrieved February 3, 2018 from http://www.bildarchivaustria.at/Pages/ImageDetail.aspx?p_iBildID=14203224.

[52] Yoshinori Teshima, Atsushi Matsuoka, Mamoru Fujiyoshi, Yuji Ikegami, Takeshi Kaneko, Susumu Oouchi, Yasunari Watanabe, and Kenji Yamazawa. 2010. Enlarged skeleton models of plankton for tactile teaching. In *ICCHP'10, Part II, Lecture Notes in Computer Science*, Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer (Eds.), Vol. 6180. Springer, Berlin, 523–526. https://doi.org/10.1007/978-3-642-14100-3_78.

[53] Jan Walmsley and Kelley Johnson. 2003. *Inclusive Research with People with Learning Disabilities: Past, Present and Futures*. Jessica Kingsley Publishers, London.

[54] Andrew D. Wilson. 2010. Using a depth camera as a touch sensor. In *ACM International Conference on Interactive Tabletops and Surfaces (ITS'10)*. ACM, New York, NY, USA, 69–72. https://doi.org/10.1145/1936652.1936665.

[55] Judy Wing. 2012. Ancient hieroglyphics meet cutting-edge technology at Loughborough University. (November 2012). Retrieved February 3, 2018 from http://www.lboro.ac.uk/service/publicity/news-releases/2012/197_Manchester-Museum.html.

[56] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*. ACM, New York, NY, USA, 1083–1092. https://doi.org/10.1145/1518701.1518866.

[57] Kenji Yamazawa, Yoshinori Teshima, Yasunari Watanabe, Yuji Ikegami, Mamoru Fujiyoshi, Susumu Oouchi, and Takeshi Kaneko. 2012. Three-dimensional model fabricated by layered manufacturing for visually handicapped persons to trace heart shape. In *ICCHP'12, Part II, Lecture Notes in Computer Science*, Klaus Miesenberger, Arthur Karshmer, Petr Penaz, and Wolfgang Zagler (Eds.), Vol. 7383. Springer, Berlin, 505–508. https://doi.org/10.1007/978-3-642-31534-3_74.

[58] Hui-Shyong Yeo, Byung-Gook Lee, and Hyotaek Lim. 2015. Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware. *Multimedia Tools and Applications* 74, 8, 2687–2715. https://doi.org/10.1007/s11042-013-1501-1.

Received April 2017; revised July 2017; accepted August 2017