Gesture-Based Interactive Audio Guide on Tactile Reliefs

Andreas Reichinger, Anton Fuhrmann, Stefan Maierhofer and Werner Purgathofer
VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH
Donau-City-Str. 1, 1220 Wien, Austria
reichinger@vrvis.at

ABSTRACT

For blind and visually impaired people, 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. However, without proper guidance some reliefs are still difficult to explore autonomously.

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

We present a working prototype, discuss design decisions and present the results of two evaluation sessions with a total of 20 visually impaired test users.

Keywords

Interactive Audio Guide; Tactile Reliefs; Finger Tracking; Gesture Detection; Evaluation; Blind Users;

1. INTRODUCTION

Tactile materials are widely used among blind and visually impaired (BVI) people that help to perceive and understand graphic content, that is otherwise difficult to convey. Such tools may be categorized according to the taxonomy in [26] into a) two-dimensional (*2D*) objects [1, 8] like tactile diagrams, line drawings or plans, e.g. on embossed paper, swell paper, and increasingly also with vibrotactile cues [17, 22], b) fully *3D* objects [23, 26, 30, 34] like anatomical models, 3D-printed reproductions or everyday objects, and c) the *2.5D* realm in-between, i.e., "height fields, surfaces that can be represented by a function $z = f(x, y)$, giving every point above a plane a single height value" [26]. The last group, tactile relief, is especially useful to ease access to the visual arts of images, photos and paintings, as it is important to keep the connection to the two-dimensional original, 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 demand and importance is demonstrated by more and more art shows (e.g. [27]) all over the world incorporating tactile reliefs, as well as technical developments [10, 11, 25] in order to ease their creation.

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. [9]), verbal description is a very important part, especially for art. 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 side, a single monolithic text may not be satisfactory as well. While most appreciate a top-level introduction to orient themselves, detailed descriptions are better given on demand. Each person might be interested in different details and might rather request them when they find an interesting region, as opposed to getting the details in a pre-defined order and having to find the matching locations.

In museums and galleries, a BVI person is typically guided by a trained person, who is prepared to answer questions to different aspects, or who can guide the hand to desired locations. However, such a guide may not always be available, or a BVI person may want to be independent and explore the relief in a more autonomous way.

Therefore, we propose a gesture-based interactive audio guide, capable of giving a user exactly this freedom.

1.1 Requirements

The aim of this work is to create a system that enables BVI people to explore tactile materials in a more autonomous way. 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. 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 adaption 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 proto-

ASSETS '16, October 23 - 26, 2016, Reno, NV, USA

ISBN 978-1-4503-4124-0/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2982142.2982176

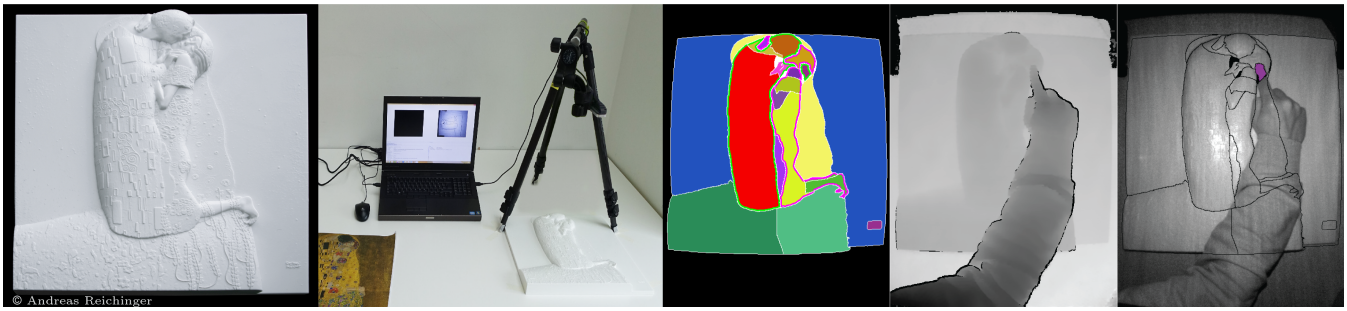

Figure 1: From left to right: a) Tactile relief interpretation of Gustav Klimt's "Der Kuss" (The Kiss, 1908/09); b) Test setup; c) Label image, warped to camera space. Outlines indicate merged base labels; d) Depth image; e) Infrared image with superimposed label borders and touched label (purple). © Andreas Reichinger

type is targeted at BVI people, according to a design-for-all philosophy, the system may be also interesting for children, elderly, people with cognitive impairments, and the general audience, possibly all with different interaction modes.

Based on discussions with BVI people, the main goal of our system is to allow users an *undisturbed* exploration, without unwanted explanations, and precise control over when and about what to get information. The user should be able to explore the relief with one or both hands without triggering unwanted audio and avoid Midas touch effects [14, p. 156]. This means, that only very distinct gestures should trigger audio comments, gestures that normally don't occur during tactile exploration and that can be reliably detected. This is in contrast to systems with embedded sensors (cf. Section 2), that are triggered by any kind of touch, whether intended or not.

2. RELATED WORK

A large body of work concentrates on augmentation of 2D graphics. The *Talking Tactile Tablet* [19] and *ViewPlus' IVEO* [13] detects touch-gestures on tactile diagrams put on a high-resolution touch pad. This technology clearly cannot be extended to relief surfaces of significant height.

Several projects utilize color cameras to track the user's fingertips: *Access Lens* [15] recognizes and reads texts on documents where the finger points at. The *Tactile Graphics Helper* [12] plays pre-recorded audio when the finger is over pre-defined labels, and is triggered by voice commands. *Tactile Graphics with a Voice* [2] is an app for cell phones and Google Glass, that reads labels indicated by QR codes. And, *Kin'touch* [3] studies the combination of optical finger tracking and touch events from a capacitive multi-touch screen. While these approaches focus on 2D documents, some could probably be extended to the third dimension. However, most require labels which 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.

Talking Pen Devices[1] 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.

[1]Multiple vendors offer talking pens, like the TalkingPEN (www.talkingpen.co.uk), Talking Tactile Pen (www.touchgraphics.com), Livescribe (www.edlivescribe.com) or Ravensburger tiptoi (www.tiptoi.com).

Several full 3D approaches are based on devices integrated *into* the tactile object. For instance, *Tooteko* [5] integrates NFC Tags in 3D models which are read by a wearable NFC reader. *Digital Touch replica* [32] have touch sensors integrated at interesting locations. Most recently, *3DPhotoWorks* (www.3dphotoworks.com) managed to print the color images directly on the relief surfaces [21] and integrated infrared sensors into their reliefs. While this is a robust solution for a museum setting, these approaches are less flexible. Once placed, trigger regions cannot be changed any more, and probably not reused on other objects. Only discrete trigger locations are possible, and interaction modes requiring fine-grained touch positions are not possible. Furthermore, the sensors react to any kind of touch, which conflicts with tactile examination by BVI people (cf. Midas touch).

Probably for the first time, Wilson [31] introduced the concept of using a depth camera as a touch sensor on non-flat surfaces. *CamIO* [29] 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, that could even be rotated. The probably most similar approach is a feasibility study [4], which uses a Microsoft Kinect with the CVRL FORTH Hand Tracker [20] to trigger audio by touch events of the right index fingertip 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 [24], 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 (cf. Fig. 1b). In contrast to conventional color cameras that 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 and the hands kept away. The system stores the acquired depth image, the so-called *background image*. Whatever is now put on top of the relief creates depth mea-

surements that are nearer to the camera, and can therefore be easily detected. This process is called foreground segmentation (cf. Section 3.3.3), and creates a *foreground mask*, a set of pixels where new things are located. As any objects may be added, the foreground is carefully searched for hands, and whether these hands form certain input gestures (cf. Sections 3.3.5–3.3.8). Finally, depending on the gestures, real-time audio 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 lighting situation, working even in complete darkness, as it has its own, for humans invisible lighting. In contrast to color images, depth allows a more reliable foreground segmentation, that is independent from relief and skin color, even gloves may be worn. Depth further allows to detect touch-events to trigger interaction, whereas systems using color cameras have to use, e.g., voice commands. It is more flexible than approaches with integrated sensors, works on arbitrary 3D surfaces, and allows gestures "beyond touch" [31]. Depth cameras are nowadays 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.

In the remainder of this paper 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). This popular painting was chosen because many BVI people most certainly heard about it, but to date only descriptions, and a simple raised line diagram were available. The relief (cf. Fig. 1a) was created based on an approach by Reichinger et. al. [25], using custom software to segment and layer depicted objects and to extract surface textures. In addition, we integrated rigged 3D models for the figures, deformed to match their poses, and Beziér surfaces for the cloths.

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 (like background, meadow, male and female figure), and the two figures were further subdivided for a total of 20 different labels of varying size (cf. Fig. 1c). In addition five short general texts (50–60 seconds each) about the painting, its history, interpretations and the artist have been recorded.

3.2 Interaction Design

Despite ongoing research on gestures for BVI people (e.g. [16]) and user-elicited gestures (e.g. [33]) these are only valid for dynamic interaction on flat screens, and are not directly applicable. For our specific case with static reliefs and depth sensors, careful interaction design was important.

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 above-mentioned general information, but also for application commands, like audio controls.

Our design choices are based on typically used exploration strategies 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. 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.

3.2.1 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 also matches motions occurring naturally in the detail exploration phase. We allow any finger to be used for interaction, so BVI people can choose whichever they feel most comfortable with. This is in contrast to Buonamici et al. [4] who require using the right index finger. We decided for the typical pointing gesture, having all fingers but one contracted into a fist. This gesture feels very natural, while at the same time it is only rarely used during normal exploration, which mostly avoids triggering unwanted audio.

In order to account for the two exploration phases, we at first play back the region name of the selected part, and only after a longer touch gesture (until the name was played), the detail 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 click sound, as an important feedback to the user and also to avoid confusion when a text was unintentionally interrupted.

3.2.2 Hierarchical Exploration

As mentioned before, two basic regions (male and female figure, cf. Fig. 1c) were further subdivided into smaller parts, mainly the parts of the bodies and cloths. 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 detail description of a figure (which includes important information about the posture and relation to other regions) the subdivided parts of the respective figure will become available instead of the basic region.

3.2.3 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 10 cm above the relief will stop the current playback. The other hand may still remain on the relief to stay oriented. According to our main goal of an undisturbed exploration, this command is the most important, as it allows to cancel unwanted or unintentional audio. Background information can be triggered by *number-gestures*. As our number of general texts (cf. Section 3.1) is exactly 5, we chose to simply count the number of extended fingers of the lifted hand, which also generalizes to differ-

ent number gestures in different cultures. More chapters are unlikely for paintings in a museum context, but different gestures may be implemented in a future work. Once the gesture is detected, a click sound followed by the number of fingers and the title of the chapter is played. Again, a newly detected gesture interrupts the playback of the former. 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. Once the user is satisfied with the choice, the text starts directly after its title, and the users can lower their hand and continue the tactile exploration while listening.

3.2.4 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. After the system is not used for a given amount of time, the system is reset and waits for the next user.

3.3 Implementation

Based on the selected set of gestures, the requirements for the gesture detection system are as follows: 1) a reliable detection of hands and individual fingers, 2) measurement of the palm height for off-relief gestures, 3) detection of touch events of a pointing finger and the position of the touch.

3.3.1 Sensor Selection

In [24] the concept for our setup is described, requirements for a tracking camera are analyzed, and several state-of-the-art cameras are reviewed. We follow the suggestion to use the *Intel RealSense F200* as the currently most suitable sensor for our application. It has a sufficient resolution of the depth sensor (true 640×480 pixels) with up to 60 fps and a low noise level. Combined with its near operating range and suitable field-of-view we achieve an effective resolution of more than 10 pixels per cm (25 dpi) on the relief. In our setup the sensor is centered approximately 40–45 cm above our 42×42 cm relief, overlooking the whole relief including a few centimeters of its surrounding (cf. Fig. 1). We chose portrait orientation, as it is more important to detect the hands beyond the relief towards the user. At the depth in our setup, the sensor does not give measurements at an up to 40 pixels wide region on one side, which we rotated beyond the top of the relief, away from the user (cf. Fig. 1d).

The RealSense F200 is a time-sequential structured light scanner. For each depth measurement frame, several Gray-coded stripe patterns are projected with an infrared (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 [6]. 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, cf. Fig. 1d), 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 (cf. Fig. 1e), 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. It has a low noise level in the depth values, with a standard deviation below 1 mm on smooth surfaces. However, noise levels vary in a moiré-like pattern (cf. Fig. 2b), 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:

1) 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.

2) Since the scanner requires multiple frames per measurement, fast moving objects, or more specific, 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 in motion.

3) 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 depth values, slowly changing over time, and some abrupt changes. The tilt exists mainly 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 Software Implementation

As pointed out by Reichinger et al. [24], 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 we found is the CVRL FORTH Hand Tracker [20] already tested by [4] for a similar application. However, we could not use the software, because a) it currently only supports sensors of the Kinect family, b) the demonstrator only tracks a single hand and requires an initialization pose,[2] and c) according to [4] it loses tracking for fast movements. Similar approaches have been published (e.g. [28]) or created by www.NimbleVR.com but implementations are not or no longer available.

As such approaches are very hardware demanding, 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 (cf. Fig. 2a), i.e., do not touch. 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 the detailed documentation we based our implementation on [35]. 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.

3.3.3 Silhouette Detection

The original paper addresses both color-based foreground segmentation using RGB cameras, and depth segmentation using a depth camera. In our prototype we use the depth measurement as main segmentation key, complemented by the infrared image, which proved adequate for our requirements. We do not currently use color information in our prototype, although it could be interesting as an extension

[2]Extensions where published (e.g. [18]) but their implementations are not publicly available.

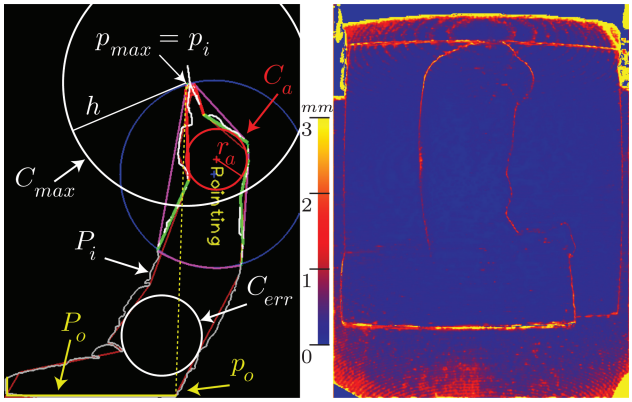

Figure 2: From left to right: a) Hand detection output and palm detection diagram; b) Standard deviation of background measurements over 100 frames.

in the future. Due to the sensor's low noise in depth measurements (with a variance below 1 mm) 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.

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 [4, 29], where 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 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 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. The combined foreground probability p is the weighted average $p = (w_D p_D + w_{IR} p_{IR})/(w_D + w_{IR})$, 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.[3]

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 infrared channel values. Detection based on a single intensity value is of course rather error prone. Nevertheless, skin and worn cloths often have a significantly different infrared reflection than the tactile relief (cf. Fig. 1e), giving an additional clue as to where the correct silhouette is located.

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

[3]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 $I_{BC} = (0.2, 1)$.

3.3.4 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.

The tilt-parameters δ_0, δ_x and δ_y are estimated as a 2D linear regression on the difference between the stored mean background $\overline{b}$ and a running average of the latest 5 depth values $\overline{c}$. Currently detected foreground regions (enlarged by a safety margin of 12 px) are excluded, as well as unreliable measurements of b and c, which yielded invalid measurements during the average computation. The differences are clamped to ±3 mm to avoid excessive outliers, and 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. Due to performance, the adjustments are only performed once every 4 frames.

3.3.5 Palm Detection

Following [35], the hands are detected solely based on their silhouettes. From the foreground mask of Section 3.3.3, all connected components larger than 5000 pixels are chosen as potential hand regions, and their contours are extracted. Of course, this approach only works if the hands and arms do not touch or overlap. This is satisfied for the selected 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. [18]).

Assuming the region contains a single hand, the position of the palm has to be found. The original approach [35] finds the largest circle C_a inscribed in the silhouette. However, this often fails, e.g., when the user is wearing loose cloths (cf. Fig. 2a, C_{err}). Our solution is as follows:

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

$$p_{max} = \arg\max_{p_i \in P_i} \min_{p_o \in P_o} \|p_i - p_o\|. \quad (1)$$

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 $h = 200px \times 390mm/\overline{d}$. 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.

3.3.6 Fingertip Detection

Fingertip detection is similar to [35]. The hand silhouette is clipped to a bounding circle 3.5 times the radius of the

palm, the resulting polygon is simplified, convexity defects are computed and filtered whether they could represent the empty space between fingers.[4] Between neighboring pairs of all accepted convexity defects we test for potential fingertips. We modified the criteria as follows: a) The arc-distance along P_i between two consecutive convexity defects and their angle must be below certain thresholds. b) Similar to [35], we require the k-curvature 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 px, to allow for locally flat but still elongated fingertips to be detected. c) We limit the width of such a potential fingertip, to eliminate cases like two fingers pressed together that still may pass the k-curvature test.

We also modified the fingertip localization. While in [35] the fingertip location is determined from the k-curvature points, we found this too unreliable, due to the often rather jagged contours occurring in our setup. Instead we take the part of the finger contour between the two k-convexity end points, and compute the oriented bounding box to get more reliable estimates for the finger's direction and width. In order to extract the tip region, we take the valid pixels inside the top square region of the 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 values of these pixels. Finally, we classify the fingertip's quality into three categories: If the tip-region has too few pixels (<50) it is not classified as finger. If the finger is too wide for the given depth (width × depth > 25 px × 400 mm) it is labeled as *blob*, being probably a union of two fingers, and only if it satisfies both, it is labeled as a single finger.

3.3.7 Gesture Recognition

We do not perform frame to frame tracking, as this is not necessary for the current gestures, and would introduce recovery problems once a hand was lost. Nevertheless, we need to make the gesture recognition robust, as the detected hands and fingers may vary each 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 palm-to-sensor distance is below a certain value in 75% of the last 20 frames, *with* the same amount of fingers detected.

3.3.8 On-Object Touch Event

In order to relate the detected touch events to regions on the relief, and therefore to different audio files, the regions have to be labeled by the content author (cf. Fig. 1c). Since our setup is static, we simply sketch the labels on a once acquired IR image of the relief (cf. Fig. 1e). While this of course does not adapt to a different camera placement as in [29] and [4], it proved accurate enough for our rigid setup. For added flexibility, a future implementation might incorporate an automatic and dynamic calibration. Manual initialization [4] and fiducial markers [29] might be eliminated by detecting the base plane and corners of rectangular reliefs, or by using novel 3D feature based algorithms [7].

Finally, the fingertip location has to be mapped to the regions. While Buonamici et al. [4] 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 90% of these pixels are on the same label, the detection is considered unique. Otherwise, the finger might be on a border between labels and it is not decidable, which label the user meant. A touch event is generated, when at least 70% of the last 10 frames detected the same unique label (cf. purple area in Fig. 1e). 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 they mostly kept the finger on the relief.

4. EVALUATION

The implemented system was evaluated in two sessions in two different European countries. The first session was an informal 2.5 hours long evaluation with 7 mostly elderly BVI people, with the majority having some rest of sight. Based on this first feedback we implemented a structured evaluation which took place in the course of 2 full days with 13 people (5 female, aged 11–72, avg. 50).

Of the 13 volunteers, 6 were fully blind with no sense of sight, 4 had a minimum rest of sight below 1% that did not help them perceive images and 3 had some rest of sight. 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 4–5 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 between 6–8, three <3).

The presented prototype was part of a larger evaluation with four different devices. However, 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. Each participant 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 allowed. These are summarized in Figure 3.

The testers where seated in front of the relief, so they could comfortably reach it. The introduction was kept minimal, stating the general idea of the IAG, and showed them 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 was sufficient. However, one examiner was always present, prepared to answer questions or help with the interface.

4.1 General Impression

We got very good feedback for the system in general. On the question, whether the IAG helped gaining a better understanding of the painting, all gave a rating above 8 (average 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

[4]We use slightly different criteria than [35]. 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.

1	2	3	4	5	6	7	8	9	10	avg.
How important are touch tools for you?										
1	1	1			1	2	1	2	4	7.2
How did you get along with the system?										
				1		4	1	2	5	8.4
Did IAG help to better understand painting?										
							1	5	7	9.5
How understandable is the introduction?										
			1			2	1	6	3	8.5
How easy is it to perform the gestures?										
				1	1		2	3	5	8.8
How easy is it to trigger a desired comment?										
			1		1		6	2	3	8.3
How satisfied with number of described parts?										
							3	5	3	9.1
How satisfied with description texts?										
						1	2	3	7	9.3
How import that audio only played when wanted?										
						1	2		10	9.5
Is this technology meaningful in museums?										
							3	1	9	9.5
Would you rather go to a museum offering IAG?										
1	1				1		1	1	8	8.2
Would you use this technology at home?										
1		1		3			1	3	4	7.2
Would you buy such system (approx. 200 EUR)?										
		1		1			1	3	6	8.6
Your general impression of the relief?										
				1			1	2	7	9.2
How good did you get the overall composition?										
				1	1			4	7	9.0
How good did you get the details of painting?										
		1		1			2	4	4	8.4

Figure 3: Results of ranking questions on a Likert scale from 1 to 10, 10 being the "best". Translated abbreviated questions, histograms and averages.

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 of a human guide gives them the freedom to explore it without pressure, as long and as detailed as they wanted. One person expressed that they "felt guided", probably caused by descriptions that cross-reference nearby regions, guiding 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, where 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 rest of sight questioned the necessity of such a system, concluding that it probably depends on the complexity of the relief. In general, it seemed that completely blind people appreciated the system most, as people with rest of sight 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 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 or slight manual corrections of their hands. After a short training phase, nearly all could perform the gestures on their own. When asking for 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, or the camera not detecting all fingers for the chapter selection. Either the hand was partly outside the camera, or was not held fully frontal to the camera. Some people frequently lifted their hands up from the relief, and accidentally triggered off-object commands. This mainly occurred with people with a rest of sight, and while talking to the examiner. Participants encountering such problems suggested to use hardware buttons, voice-commands or knocking-signals instead of the gestures. Some also expressed the desire for additional playback-commands, like 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.

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. 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 is yet unclear, how to best select the interacting fingers in such alternative gestures. Maybe by performing a kind of double-tap with the specific finger?

Another source of error and probably also the main course of the discomfort, is the current requirement to perform it in a very flat way, required by the the silhouette-based hand detector. Especially elderly people from the first evaluation session had problems performing the required flat pointing gesture, as their hands were already less flexible, or had medical conditions like arthritis. Some participants thought, 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. Maybe a combination of multiple cameras can 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 going into 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 even fully listen to it. In the future, an explicit level control may be investigated.

Lastly, 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 like the hands of the figures, which are not much larger than the fingertip itself. This is probably caused by the current algorithm, that 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. Maybe the single point interaction of [4] is of advantage here.

4.3 Content

Nearly all participants were satisfied with the presented content. 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 (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 rest of sight, while others found it irrelevant as long as the original can be seen next to it.

People were highly satisfied by the texts (average 9.3), and by the number of described parts (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. On the question 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, where several people expected more descriptions than just a single text covering the whole area.

4.4 Acceptance and Field of Application

All test users found the presented technology to be meaningful in a museum setting with an average ranking of 9.5. However, not all would *rather* go to a museum if it was offering an IAG (average 8.2), as they would go to the museums in any case. Even less would consider it for home use (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 8–9. They would like to use it for the annotation of plans, for object detection ("which bottle was the good wine?"), for photo exploration, geography, education, and would like to see it also in schools or other educative institutions (e.g. at the zoo).

5. CONCLUSION AND FUTURE WORK

We presented a gesture-controlled interactive audio guide that allows access to location-specific content, triggered directly with the fingertips on relief surfaces, and demonstrated its real-world usability. The prototype is targeted at a museum setting, but the low-cost sensor hardware and the perspective that these sensors will soon be integrated in laptop and mobile devices, makes it very attractive also for home use, or in educational institutions. The algorithms are lightweight and may run on embedded systems, like the ORBBEC Persee, the first depth camera with integrated computer. Although this work focused on 2.5D tactile reliefs of paintings, the techniques should generalize to any 2D, 2.5D and 3D object.

The majority of the 20 test users found it useful and worth further developments. It seemed to be especially interesting for fully blind people, who like to go into detail, and to do this autonomously. Based on the feedback, we will critically review the selected gestures and the interface design with the target group. We will investigate fully articulated finger trackers, alternative sensor placement, or multiple sensor setups, to better observe the fingertip, especially near the top edge, and to relax the need for flat finger gestures.

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. Finally, we would like to investigate new interaction possibilities. Planned features include multiple knowledge layers, multi-finger gestures, educative games, sonification of color, and an extension to exchangeable 3D objects with arbitrary and dynamic placement.

6. ACKNOWLEDGMENTS

This work was performed within the framework of the project Deep Pictures, supported by the Austrian Science Fund (FWF): P24352-N23, and within the Erasmus+ project AMBAVis (http://www.ambavis.eu) and has been funded with support from the European Commission. 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.

7. REFERENCES

[1] E. S. Axel and N. S. Levent, editors. *Art Beyond Sight: A Resource Guide to Art, Creativity, and Visual Impairment*. AFB Press, 2003.

[2] C. M. Baker, L. R. Milne, J. Scofield, C. L. Bennett, and R. E. Ladner. 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, pages 75–82, New York, NY, USA, 2014. ACM.

[3] A. Brock, S. Lebaz, B. Oriola, D. Picard, C. Jouffrais, and P. Truillet. Kin'touch: understanding how visually impaired people explore tactile maps. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 2471–2476. ACM, 2012.

[4] F. Buonamici, R. Furferi, L. Governi, and Y. Volpe. *Universal Access in Human-Computer Interaction. Access to Interaction: 9th International Conference, UAHCI 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part II*, chapter Making Blind People Autonomous in the Exploration of Tactile Models: A Feasibility Study, pages 82–93. Springer International Publishing, Cham, 2015.

[5] F. D'Agnano, C. Balletti, F. Guerra, and P. Vernier. Tooteko: a case study of augmented reality for an accessible cultural heritage. Digitization, 3D printing and sensors for an audio-tactile experience. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, volume XL-5/W4, pages 207–213, 2015.

[6] S. Dixon-Warren. Inside the Intel RealSense Gesture Camera. http://www.chipworks.com/about-chipworks/overview/blog/inside-the-intel-realsense-gesture-camera, accessed May 2016.

[7] B. Drost and S. Ilic. 3d object detection and localization using multimodal point pair features. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 9–16. IEEE, 2012.

[8] P. K. Edman. *Tactile Graphics*. American Foundation for the Blind, New York, 1992.

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

[10] R. Furferi, L. Governi, Y. Volpe, et al. Tactile 3D Bas-relief from Single-point Perspective Paintings: A Computer Based Method. *Journal of Information & Computational Science*, 11(16):5667–5680.

[11] R. Furferi, L. Governi, Y. Volpe, L. Puggelli, N. Vanni, and M. Carfagni. From 2D to 2.5D i.e. from painting to tactile model. *Graph. Models*, 76(6):706–723, 2014.

[12] G. Fusco and V. S. Morash. 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*, pages 97–106. ACM, 2015.

[13] J. A. Gardner and V. Bulatov. Scientific diagrams made easy with IVEO^{TM}. In K. Miesenberger, J. Klaus, W. L. Zagler, and A. I. Karshmer, editors, *ICCHP 2006*, volume 4061 of *LNCS*, pages 1243–1250, Heidelberg, 2006. Springer.

[14] R. J. K. Jacob. The Use of Eye Movements in Human-computer Interaction Techniques: What You Look at is What You Get. *ACM Trans. Inf. Syst.*, 9(2):152–169, Apr. 1991.

[15] S. K. Kane, B. Frey, and J. O. Wobbrock. Access lens: a gesture-based screen reader for real-world documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 347–350. ACM, 2013.

[16] S. K. Kane, J. O. Wobbrock, and R. E. Ladner. Usable Gestures for Blind People: Understanding Preference and Performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 413–422, New York, NY, USA, 2011. ACM.

[17] R. L. Klatzky, N. A. Giudice, C. R. Bennett, and J. M. Loomis. Touch-screen technology for the dynamic display of 2D spatial information without vision: Promise and progress. *Multisensory Research*, 27(5–6):359–378, 2014.

[18] N. Kyriazis and A. A. Argyros. Scalable 3D Tracking of Multiple Interacting Objects. In *IEEE Computer Vision and Pattern Recognition (CVPR 2014)*, pages 3430–3437, Columbus, Ohio, USA, June 2014. IEEE.

[19] S. Landau and K. Gourgey. Development of a Talking Tactile Tablet. *Inf. Technol. Disabil.* , 7(2), 2001.

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

[21] J. S. Olson and A. R. Quattrociocchi. Method and apparatus for three-dimensional digital printing, June 23 2015. US Patent 9,061,521.

[22] S. OÂŠModhrain, N. A. Giudice, J. A. Gardner, and G. E. Legge. Designing Media for Visually-Impaired Users of Refreshable Touch Displays: Possibilities and Pitfalls. *IEEE Transactions on Haptics*, 8(3):248–257, July 2015.

[23] S. Oouchi, K. Yamazawa, and L. Secchi. 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 K. Miesenberger, J. Klaus, W. Zagler, and A. Karshmer, editors, *ICCHP 2010, Part II*, volume 6180 of *LNCS*, pages 527–533, Heidelberg, 2010. Springer.

[24] A. Reichinger, A. Fuhrmann, S. Maierhofer, and W. Purgathofer. A Concept for Re-Usable Interactive Tactile Reliefs. In K. Miesenberger, C. BÃijhler, and P. Penaz, editors, *ICCHP 2016, Part II*, volume 9759 of *LNCS*, pages 108–115, Heidelberg, 2016. Springer.

[25] A. Reichinger, S. Maierhofer, and W. Purgathofer. High-Quality Tactile Paintings. *J. Comput. Cult. Herit.*, 4(2):5:1–5:13, 2011.

[26] A. Reichinger, M. Neumüller, F. Rist, S. Maierhofer, and W. Purgathofer. Computer-Aided Design of Tactile Models. In K. Miesenberger, A. Karshmer, P. Penaz, and W. Zagler, editors, *ICCHP 2012, Part II*, volume 7383 of *LNCS*, pages 497–504, Heidelberg, 2012. Springer.

[27] L. Secchi. Seeing with the Hands - Touching with the Eyes, Work of Art Reading as a Hermeneutical Act.

[28] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3633–3642. ACM, 2015.

[29] H. Shen, O. Edwards, J. Miele, and J. M. Coughlan. 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 and Accessibility*, page 41. ACM, 2013.

[30] Y. Teshima, A. Matsuoka, M. Fujiyoshi, Y. Ikegami, T. Kaneko, S. Oouchi, Y. Watanabe, and K. Yamazawa. Enlarged Skeleton Models of Plankton for Tactile Teaching. In K. Miesenberger, J. Klaus, W. Zagler, and A. Karshmer, editors, *ICCHP 2010, Part II*, volume 6180 of *LNCS*, pages 523–526, Heidelberg, 2010. Springer.

[31] A. D. Wilson. Using a Depth Camera As a Touch Sensor. In *ACM International Conference on Interactive Tabletops and Surfaces*, ITS '10, pages 69–72, New York, NY, USA, 2010. ACM.

[32] J. Wing. Ancient hieroglyphics meet cutting-edge technology at Loughborough University. http://www.lboro.ac.uk/service/publicity/news-releases/2012/197_Manchester-Museum.html, accessed March 2015, November 2012.

[33] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1083–1092, New York, NY, USA, 2009. ACM.

[34] K. Yamazawa, Y. Teshima, Y. Watanabe, Y. Ikegami, M. Fujiyoshi, S. Oouchi, and T. Kaneko. Three-Dimensional Model Fabricated by Layered Manufacturing for Visually Handicapped Persons to Trace Heart Shape. In K. Miesenberger, A. Karshmer, P. Penaz, and W. Zagler, editors, *ICCHP 2012, Part II*, volume 7383 of *LNCS*, pages 505–508, Heidelberg, 2012. Springer.

[35] H.-S. Yeo, B.-G. Lee, and H. Lim. Hand Tracking and Gesture Recognition System for Human-computer Interaction Using Low-cost Hardware. *Multimedia Tools Appl.*, 74(8):2687–2715, Apr. 2015.