CAVIAR: A Vibrotactile Device for Accessible Reaching

Sina Bahram, Arpan Chakraborty, Robert St. Amant
Department of Computer Science
North Carolina State University
sbahram | achakra | stamant@ncsu.edu

ABSTRACT
CAVIAR is designed to aid people with vision impairment in locating, identifying, and acquiring objects in their peripersonal space. A mobile phone, worn on the chest, captures video in front of the user; the computer vision component locates the user’s hand and objects in the video stream. The auditory component informs the user about the presence of objects. On user confirmation, the reaching component sends signals to vibrotactile actuators on the user’s wristband, guiding the hand to a specific object. This paper describes an end-to-end prototype of CAVIAR and its formative evaluation.

ACM Classification Keywords
H5.2 [Information interfaces and presentation]: User Interfaces. - Haptic I/O.

Author Keywords
Mobile, haptic, tactile, accessibility.

INTRODUCTION
The Computer-vision Assisted Vibrotactile Interface for Accessible Reaching (CAVIAR) is a wearable tactile system designed to aid people with vision impairment (PWVI) in locating, identifying, and acquiring objects in their peripersonal space (i.e., the space within their reach). The prototype relies on a mobile phone for video input/processing and a wristband with four vibrating actuators. The mobile phone processes video input to recognize objects within its field of view and makes these known to the user via audio. On voice command, directional guidance to a chosen object is given as a sequence of continuous signals to the actuators on the user’s wristband. A brief scenario illustrates the use of CAVIAR.

The user is sitting at a work table, wearing a mobile phone in a harness on his chest (running CAVIAR) and a wristband on his wrist. The phone’s camera scans items on the desk. The phone informs the user of distinct objects it identifies. On user confirmation, CAVIAR sends vibrotactile signals using the wristband to guide the user’s hand to the object.

Work in this area—using modern technology to make the world more accessible to people with vision impairment—is important. Hundreds of thousands of PWVI live in the U.S. alone, and a lightweight, low-cost system like CAVIAR could significantly improve their lives. PWVI have strategies for retrieving nearby objects, relying on common sense knowledge about object locations, grid search techniques, and the sense of touch. But many modern objects are difficult to distinguish based on touch alone; consider a store display of DVDs or packaged food. Computer vision plus guidance techniques can provide useful missing information.

Researchers have developed wearable vibrotactile feedback systems for motor tasks (e.g. [3, 5]); CAVIAR shares some of the goals of such systems. We have implemented a prototype of CAVIAR, running on a mobile phone, with the following:

- Communication: Voice recognition, based on an existing online service, for issuing commands to the device and speech generation for system feedback.
- Computer vision: Object recognition from video on the phone and tracking of objects and the location of the hand.
- Tactile guidance: The wristband, with its vibrotactile actuators, and the embedded software to generate signals to guide the user’s hand. The wristband is directed via Bluetooth from a mobile phone or other similar device.

This paper focuses on the design and formative evaluation of the computer vision and tactile guidance capabilities of the prototype, tested as separate components from the integrated system. To identify relevant design parameters and estimates for their values, we have carried out a non-interactive study with a fixed camera and an interactive, single-participant study using video from an overhead camera. The vision component can locate objects with a Euclidean distance error of 6.1 cm, which is adequate for the constraints of the task. Guided, linear movement is possible, though speed varies across different conditions, ranging from about 7 to 20 cm/s. We find lower deviations from guided paths in rectilinear versus diagonal movements. For rectilinear movement, response times to vibrotactile signals average 700 ms for changes of direction and 300 ms for stopping at a target location. These formative findings were influential on the development of CAVIAR.

SYSTEM
In CAVIAR, a wristband with small coin-type vibrators serves the purpose of generating continuous stimuli for guiding the user. The vibrators are governed by a microcontroller strapped
to the arm, which communicates with the phone over Bluetooth. These physical components are shown in Figure 1.

**Computer vision**

The vision component is a finite state machine for image processing and supporting functions. A master controller receives frames from the phone’s camera, dispatches them to other appropriate components, and orchestrates user interaction. It maintains a common data store. The other parts of the vision system are as follows.

**Calibration.** CAVIAR reasons about spatial quantities, which requires a calibration step to obtain real-world equivalents of information in the system’s image space. Camera calibration techniques exist for robust computation of parameters [2], e.g., to use a planar pattern with distinctive features on it, such as a chessboard, and compute the intrinsic and extrinsic camera parameters from different views of the pattern.

In CAVIAR, two colored squares are placed a known distance apart on the thumb. The centroids of the two detected squares are used in lieu of absolute points. The known and observed distances between these two points (in pixels) can be used to obtain a relative measure of the distance from the hand to the phone’s camera. This information, along with the assumption that a table-top surface is flat and level with the ground, lets us compute the vertical distance of the camera’s line of view from the table’s surface. We assume that the camera’s line of view is parallel to the surface.

Formally, if the projection of a real point \( \hat{\mathbf{p}} \) on a plane surface \( S \) is a point \( \mathbf{p}' \) on the image plane, and \( \hat{\mathbf{O}} \) is the focal point of the camera, then the \( \hat{\mathbf{p}} \) is the point of intersection of the line \( L \) (passing through \( \hat{\mathbf{O}} \) and \( \mathbf{p}' \)) and plane \( S \). Any two-dimensional plane \( S \) can be characterized by a point \( \mathbf{p}_0 \) lying in it and a unit normal vector \( \hat{\mathbf{n}}_s \), such that all points \( \mathbf{p} \) lying in the plane satisfy the equation:

\[
(\mathbf{p} - \mathbf{p}_0) \cdot \hat{\mathbf{n}}_s = 0
\]

Similarly, the equation for a line \( L \) passing through two points, say \( \hat{\mathbf{O}} \) and another point \( \mathbf{p} \) is

\[
\mathbf{p} = d \cdot \hat{\mathbf{n}}_l + \hat{\mathbf{O}}.
\]

Here \( d \) is the Euclidean distance between \( \hat{\mathbf{O}} \) and \( \mathbf{p} \), and \( \hat{\mathbf{n}}_l \) is the unit vector representing the direction of \( L \). Given any point \( \mathbf{p}' \) on the line, \( \hat{\mathbf{n}}_l \) can be computed as

\[
\hat{\mathbf{n}}_l = \frac{\mathbf{p}' - \hat{\mathbf{O}}}{|\mathbf{p}' - \hat{\mathbf{O}}|}.
\]

Given the observed separation between the two calibration points on the image plane \( \mathbf{r}_{obs} \) and the known real separation \( \mathbf{r}_{real} \), \( d \) can be estimated as \( k \times \mathbf{r}_{real}/\mathbf{r}_{obs} \), where \( k \) is a constant of proportionality. Following that, the intersection point \( \mathbf{p}' \) is simply \( \mathbf{p}' = d \cdot \hat{\mathbf{n}}_l + \hat{\mathbf{O}} \). If \( \hat{\mathbf{O}} \) is taken to be the origin, then the vertical coordinate of \( \mathbf{p}' \) is the distance between the line of view of the camera and the surface, or in other words, how far the table-top is below the phone’s camera.

**Target Acquisition.** This component integrates speech recognition, robust object detection, and affine-invariant recognition. Target acquisition is simplified by a smoothed background model and detection of foreground regions associated with changes over time. These regions are analyzed to extract colored connected components to detect object boundaries.

**Hand Location and Tracking.** Locating a familiar object is much more straightforward than detecting unknowns. There are several techniques for detecting known objects that usually require some form of limited training. In our current prototype, the user’s hand is located simply by performing a template match to look for a known marker. We re-use one of the squares from the calibration stage for this purpose. The user gets appropriate messages via the speech output system to know what is to be done at different points of time. In this case, the user needs to keep his hand in view. Once the user’s hand is located, we extract the color histogram of the region where the marker is present, and use that as our primary observation for tracking. At each frame processed from then on, we essentially look for image regions with the same histogram composition and find the most likely current position. This process is optimized by the position filtering component.

**Position Filtering.** Tracking can be enhanced if we have a model for the expected movement of objects. We use a particle filter to smooth out the information generated by the tracking step, which can be quite noisy. We initialize several particles in and around the initial location of the marker. The positions of these particles are predicted using the limited motion model and updated using observations made from the frames being processed. Each of these particles is then weighed by the similarity of a histogram extracted around it to the initial object histogram. The weighted sum of the particles’ position is used as the current position of the object. Resampling is done as necessary when particle locations degenerate. The detected hand location is then transformed into world coordinates using the transformation parameters from calibration.

**The Wristband**

The CAVIAR wristband has four 10mm shaftless vibrating motors. Each is connected via an NPN 2N3904 transistor and a 1K Ohm resistor to its own pin on an Arduino board. The wristband puts the vibrating motors in contact with the center of the dorsal (top), ventral (bottom/palm), inner and outer
lateral sides of the wrist. Assuming that the reaching is carried out with the palm of the hand facing downward, the vibrations felt by the user map to the direction indicated by the system: the wristband can signal movement to the right or left (which we'll refer to as the $x$ axis) and forward or backward (the $z$ axis); activation of two adjacent motors signals a diagonal movement. These conventions help reduce cognitive load in translating vibrotactile feedback to hand movement. PWVI need their hands, especially their fingers, for exploration and interaction with their environment. A wristband does not restrict or prevent movement of the fingers. Justification for wristband guidance can also be found in models of movement in HCI: the hand defines a frame of reference and dominates the effectiveness of movement. Unfortunately, it is known [1] that the sensitivity that the user may receive from the four quadrant regions of the wrist, as discussed above, is much less than in areas such as the knuckles or other parts of the hand closer to the fingers and fingertips; nevertheless wristbands are still commonly used in vibrotactile systems (e.g. [4]). This issue is explored below.

EVALUATION

Our formative evaluation covered two areas: computer vision for object localization and interaction with the wristband.

Computer vision

Figure 2, left, shows the camera, mounted on a tripod, and a sampling of objects within its field of view. This experimental set-up more closely approximates the intended use of the system. For a set of objects, a marker standing in for the hand was used for calibration, in particular to establish the distance of the surface. Other objects for testing included an iPhone, a coffee mug, a water bottle, and a pair of scissors.

In the coordinate system of the camera, the line of sight is parallel to the surface, with the $x^+$ axis to the right, $y^+$ down, and $z^+$ directly away from the camera. In this work, we are interested in estimates of $z$ and $x$ values for objects; however, CAVIAR is not limited to this plane for operation. We tested the computer vision component by comparing estimated locations against the actual measured locations of 18 objects, treated singly. Three cases resulted in failure due to image segmentation. Figure 3 diagrams the estimated locations of the remaining objects (shown as filled circles) with arrows pointing to actual locations.

The mean error in Euclidean distance for successful cases is 6.1 cm (in Manhattan distance, 8.1 cm), which can clearly be improved. Manhattan distance is provided because of its importance when discussing rectilinear cardinal directions without diagonal supercardinal directions. Estimation errors are due to two main sources. In some cases only the upper part of an object was captured; in other cases, shadows and other lighting effects influenced estimation. These problems are well-understood, if not completely resolved, in the computer vision literature, and we expect to improve these results.

Interaction

A set of interaction studies was carried out with a PWVI, one of the authors of this paper, acting as the single participant. An overhead camera tracked a visual marker on the wristband; Figure 2, right, shows the camera view. All signals during the experiment were based on processing of video data from the camera, as was the computation of user performance metrics after the experiment. The target of the movement was a virtual object, a square $24 \times 24$ pixels (6.4 cm) in size. A trial was completed when overlap is detected between the wristband marker and the target.

Our initial experimentation was with a beacon approach. The participant was instructed to treat the wristband as a beacon sending signals to move in any of the cardinal directions or diagonals, with the signalled direction changing based on the location of the hand relative to the target. Figure 4 shows the movement of the hand in one representative trial. The green dot indicates the starting point, the red square in the dotted circle the target. The pattern in the movement indicates that a beacon approach is impractical. A stair-stepping pattern and many movement irregularities can be seen; one 18-second trial involved over a dozen changes in direction. The problems are error from so many direction changes, and higher frustration and cognitive load.

Through further experimentation we developed a refined approach for guidance. The separation between the current location of the hand and the target is decomposed into its rectilinear components, and the components are reduced in alternation: first the $x$ direction, then $z$, then $x$, and so forth. Movement in a given direction is continuously signaled until the system detects that a threshold distance from the target has been passed, and then the alternate direction is signaled. Alternation is also triggered when excessive movement away from the target is detected in the non-signed direction. This approach effectively eliminated stair-stepping.
We ran a study in which the participant moved his hand to a virtual target, in two blocks. In the Rectilinear block, only rectilinear movements were signaled; in the Diagonal block, diagonal signals were also included. The participant carried out 100 trials in the Rectilinear block and 50 trials in the Diagonal block. Distances ranged from 27 to 103 cm, with a mean of 52 cm. Trial duration ranged from 516 ms to 41 s, with a mean of 12 s. Camera tracking failed in four of the Rectilinear trials and two of the Diagonal trials; the remaining trials contained only one diagonal failure to reach the target. The resulting interaction was much improved. Figure 5 shows a representative path in a Rectilinear trial. The blue dot shows the start of a movement, the pink dot the point at which a stopping signal was issued, and the red dot overshoot. Statistics are shown in Table 1, the top row for the Rectilinear block and the bottom row for the Diagonal trials. $T_{\text{change}}$ shows the response time to a signal for a new direction. $T_{\text{end}}$ is the interval between a signal to stop and the actual stopping of the hand; Over is the distance moved (overshoot) during that time. $\text{Speed}_i$ is the instantaneous speed along the path followed by the hand. $\text{Dev}_a$ is the instantaneous angular deviation from the signalled direction, $\#\text{Ch}$ the mean number of changes in direction signaled, and Miss the percentage of those signals for which a response was not detected (via Deviation$_a$). The new guidance strategy produced fewer directional changes and was much easier to follow.

Our formative evaluation suggests guidelines for future development: rectilinear signals show shorter response times, smaller overshoot, and lower deviation. We attribute the differences to lower cognitive effort for interpreting signals in the Rectilinear block and a subtle form of vibrotactile masking: Diagonal and Rectilinear signals can overlap in the actuators they use, which delays the recognition of a signal change. Diagonal speed is 20% higher than straight speed, surprisingly; in principle, diagonal movement also results in shorter paths. The higher speed does not entirely account for the increase in overshoot, however.

Our work suggests that movement with vibrotactile guidance is feasible. Speed of movement is variable, even for a single user, with single sessions under different conditions varying between 6 and 18 cm/s. Response times are also variable, but higher for changes of direction than for stopping. These numbers change with practice, and individual differences will doubtless appear in more extensive experiments, but the results give us preliminary guidelines for the use of CAVIAR. Narrow corridors of movement are impractical and movement along rectilinear paths appears to be more reliable than movement along diagonals. If CAVIAR is to support movement to targets while avoiding obstacles, its signals must be predictive, given the speed of movement and delays in response to signal changes. Prediction of movement at least 300 ms into the future appears to be a reasonable requirement for CAVIAR. Qualitative feedback indicates that CAVIAR is workable, though not quite ready for practical application.

ACKNOWLEDGMENTS

This work was supported by a gift from Google.

REFERENCES


<table>
<thead>
<tr>
<th>$T_{\text{change}}$</th>
<th>$T_{\text{end}}$</th>
<th>Over</th>
<th>$\text{Speed}_i$</th>
<th>$\text{Dev}_a$</th>
<th>$#\text{Ch}$</th>
<th>Miss</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.691s</td>
<td>0.303s</td>
<td>1.6cm</td>
<td>6.7 cm/s</td>
<td>26.0°</td>
<td>1.84</td>
<td>13%</td>
</tr>
<tr>
<td>0.969s</td>
<td>0.375s</td>
<td>3.2cm</td>
<td>8.2 cm/s</td>
<td>45.7°</td>
<td>1.38</td>
<td>11%</td>
</tr>
</tbody>
</table>

Table 1. Study movement characteristics