Using Vision Based Tracking to Support Real-Time Graphical Instruction for Students Who Have Visual Impairments

Bing Fang Francisco Oliveira Francis Quek
Virginia Polytechnic Institute and State University
2202 Kraft Drive, Blacksburg, VA 24060
{fangb, oliveira, quek}@vt.edu

Abstract

In this paper, we describe a real-time vision-based tracking system to help students who are blind or visually impaired (SBVI) to follow instructional discourse that employs graphical illustrations. The vision system employs a color model based tracking for both the instructor’s pointing behavior and the SBVI’s reading behavior, and maps the pointing positions into the same coordinates. Our Haptic Deictic System - HDS system also employs a haptic glove to provide SBVI access to the situated pointing behavior of instructor that is performed in conjunction with speech, and provides the instructor real-time visual feedback on the SBVI’s reading actions. Thus, our system supports two-way situated multimodal communication. In this paper, we first introduce our system, and then we discuss studies that show the efficacy of our approach.

1. Introduction

To enable lectures using graphical illustration, students who are blind or visually impaired (SBVI) should be able to read the illustration, to listen and to fuse both information streams in real time [7, 2, 5]. SBVI are capable of spatial reasoning and have the capacity for mathematics [6, 8, 4]. A critical problem in instructional discourse is that they cannot resolve the instructors deictic references towards the instructional material as she lectures [3]. This might be one of the key reasons they are typically one to three years behind their seeing counterparts [12].

To enable the SBVI follow the instructor’s pointing, we employ a typical sensor-replacement approach [1, 11]. Our system employs a real-time signal, which informs SBVI about the pointing activity of instructor by applying a pair of vision tracking subsystems. By tracking both the pointing locus of the instructor and the co-temporal reading point of the SBVI’s reading hand, our system is able to convey the direction the student should move his ‘reading finger’ over the raised line version of the figure to access the deictic field (pointing target) of the instructor. The directional signals are conveyed through a haptic glove that was designed and implemented for the purpose.

In this paper, we describe the system design and setup, including the haptic glove and camera settings, the detail of our vision tracking approach, our experimental results and our conclusion.

2. System Overview

Figure 1 (A) (captured from the math instruction experiments in college) demonstrate the classroom setup of our system. Where the instructor is pointing into a graphic in a poster, a pair of seated students (one SBVI and one sighted) are taking in the lecture. The instructors pointing gestures are tracked via the camera in the iMac. Figure 1 (B) shows the two SBVI reading the scaled raised-line version of the graphic on the poster. A down-looking camera tracks the students reading finger (a frame of the down-looking camera video is shown in Figure 1 (E)). Figure 1 (F) shows the internal detail of the haptic glove worn by the SBVI. Figure 1 (D) shows the screen of the iMac in which a video of the instructor is shown with the reading location of the visual impairment highlighted as a green dot.

During the lecture, the built-in iMac camera tracks the instructors pointing, and the down-looking camera tracks the current position of the reading finger. The system determines the disparity between where the instructor is pointing and where the students reading hand is positioned, and computes the current direction the student needs to move her hand to read the target location on the raised-line graphic and the vibrating actuator array in the haptic glove [10] activates in the appropriate pattern to guide the student to where the instructor is pointing, in essence providing the student with awareness of the instructors pointing behavior. Conversely, the iMac screen provides a feedback of SBVI reading behavior to the instructor, in effect, providing a form of ‘gaze awareness’ screen to give the instructor information.
of the state of attention of the student).

3. Vision Tracking Approach and Devices

The overview shown in Figure 1 suggests several design and technology requirements to realize our HDS approach. First, the system must be able to relate and calibrate both spaces so that the system can determine the equivalent disparity of the reading hand over the raised-line graphic vs the pointing on the instructor’s presentation graphic. This calibration must be self-updating because one cannot eliminate the possibility of the SBVI moving the raised-line drawing during reading. Second, the HDS is intended for use by individual SBVI in classrooms. This means that the tracking must be robust and low-cost (both monetarily and in computational load). Third, the tracking systems must function in real-time to provide timely coordination between the pointing, speech, and reading activity.

Figure 2 illustrates the signal provided to the SBVI. We define the Point of Instructional Focus, PIF as the location on the graphic pointed to by the instructor. The location where the SBVI is reading on her corresponding tactile image is the Tactile Point of Access, TPA. The signal that we communicate through the glove is the Disparity Vector, DV.

With a view to these requirements, we shall discuss our version tracking system in six major parts:

1. Real-time tracking of the instructor’s pointing. The objective of this module is to find the PIF.

2. Real-time tracking of the student’s pointing. This module is responsible to find the TPA.

3. The translation of the PIF to the TPA’s coordinate system. Instructor and student are tracked simultaneously by two different cameras: One observing the instructor and other the student. Each camera has its own coordinate system. We need to have both points on the same coordinate system to calculate the disparity vector (DV).

4. The haptic aided navigation subsystem. This is the part of the system responsible for delivering the DV to the student via the haptic glove.

5. The instructor’s display. The instructor’s display is a window in the system where the image captured by the camera looking at the instructor is displayed along with dots corresponding to both PIF and TPA. This display is updated in real time and can be used by instructor to follow the student’s navigation to the PIF.
3.1. Tracking the instructor’s pointing

For the instructor tracking, we use the background subtraction approach shown as Figure 3. When the system starts, we ask the instructor not to be at the board, to avoid his image being captured by the camera that will track his pointing. So, the initial frames the camera captures do not show the instructor (left figure). The system uses the information from the first several captured frames to calculate what will become the background image. To remove noise, the median value of the pixel from the first several frames are used after they are converted from color to gray.

![Figure 3. Background subtraction for PIF](image)

Assuming the typical pointing morphology of an instructor’s body, one can assume that the extremal point of the body into the instruction graphic is the PIF. Given the background model previously extracted, we can extract the instructor’s body from the graphic by a rapid background subtraction approach. Thus, interest pixel values on the instructor body (right figure) can be calculated. From this, the extremal point can be easily extracted as the PIF.

While this approach proved effective when the only activity performed by the instructor is pointing, it was not able to distinguish pointing from any other behavior when the instructor’s hand was extended from the body (e.g., when she is writing on the board). A simple and robust solution to this problem is to provide the instructor with a ‘pointing wand’. This allows her to signify a ‘point’ by simply moving it over the graphic. The design of this wand can be seen in Figure 1 (D). To simplify processing, the wand is essentially a stick with a round colored disk at the end of it. For robust wand tracking, we employ a Gaussian color model in normalized RG space, and use the Maximum Likelihood Estimation to find the best approximation for normalized RG. The color is comprised of a mean vector $\mu$ and a covariance matrix $\Sigma$. In Eq. 1, $x_k$ is a vector of normalized pixel values from the samples.

$$
\begin{aligned}
\{ & \hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} x_k \\
\Sigma = & \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{\mu})(x_k - \hat{\mu})^T 
\end{aligned}
$$

(1)

$x (x = (R_n, B_n)^T)$ corresponds to a pixel in normalized RG space.

$$
\begin{aligned}
R_n = & \frac{R}{R + G + B} \\
B_n = & \frac{G}{R + G + B} \\
G_n = & 1 - R_n - B_n
\end{aligned}
$$

(2)

Basically, we need to build the color model of the wand disk before using them in the instruction program. To built this model, we need samples of pixels of the wand. When the model is built $(\mu, \Sigma)$, we can classify each pixel, after background subtraction, by the model.

3.2. The tracking of the student’s pointing

Tracking the student’s reading finger can be tricky because we do not know a priori which finger a particular SBVI uses for reading (different individuals choose to use different fingers or even two fingers). Furthermore, there is always a need for anchoring. Anchoring occurs when blind readers leave one finger of the non-reading hand on a fixed position, normally at the beginning of the line or some distance trailing the reading hand to establish a reference. In this scenario, a camera observing the student’s hand and fingers movements might not have the means to identify which finger is actually being used for reading. A simple solution is to ask the user which finger he uses for reading and put a marker on the her fingernail. Hence, we can employ the same color extraction and tracking approach as for the instructor’s wand.

3.3. The translation of the TPA to the PIF’s coordinate system

Once we have found both PIF and PTA, the next step is to translate PTA’s coordinates from the student’s camera coordinate system to the instructor’s camera coordinate system. As we discussed before, the system must constantly update the PIF to PTA transformation because we cannot absolutely avoid the SBVI from bumping and perturbing her raised-line-drawing (to reduce this possibility, we actually provide a wooden frame with rubber cleats to hold the raised-line-drawing, but the SBVI can still inadvertently shift the frame). In Figure 4 the original PTA is shown in green and labeled $TPA$, for orginal. Similarly, Figure 5 shows TPA in red and labeled $TPA$, for transformed. The algorithm presented on table 1 shows how from a set of known values, it is possible to do the transformation.

3.4. The calculation of the disparity Vector (DV)

The final step is to calculate the Disparity Vector – DV – that goes from $PTA_i$ to $PIF$.

$$DV = \begin{pmatrix} \text{Direction} \\ \text{Distance} \end{pmatrix}$$

Where

$$direction = \frac{TPA_i(x) - PIF(x)}{\pi} \times 180^{\circ} - 90^{\circ};$$

(3)
3.5. The haptic aided navigation subsystem

During the very first stages of this research, we studied how to use the different vibration intensities to convey useful information. Originally, we had 7 different levels of vibration intensity. Our perception studies showed, however, that it was more important to provide the SBVI with a easily perceived signal rapidly, and to handle overshoots by providing a fresh DV for fast correction. Hence, we just set the motors to vibrate at the highest intensity to make the direction deciphering easier and therefore faster. It was a conservative move since the most important information to deliver is direction and not distance. However, the new glove models yielded better perception salience and encouraged us to review our original decision [10] for future designs.

3.6. The instructor’s display

The instructor’s display, Figure 1 (D) helps the instructor in collecting evidence that the student is following his explanation. The image that appears on that window is constructed from the frames captured from the cameras tracking both instructor and student. The green (light colored) circle at the tip of the instructor’s wand corresponds to the PIF, whereas the blue (dark colored) one is the TPAI. The blue (dark) rectangle is the instructor’s gesturing space (TGS). All pointing done inside this area will be delivered to the student and conversely, any pointing done outside this area will not be delivered. The instructor is aware of the behavior of visual impairment through this feedback.

4. Experiment of Phrase Charade

While the specifics of the haptic design is beyond the scope of the current paper, we provide a brief discussion for context. Our design of the haptic system took several iterations where we performed as set of ‘perception-response’ experiments to determine how the SBVI would perceive the haptic DV signal, whether the vibration of the glove interfered with finger-tip reading, and whether a SBVI can simultaneously listen to spoken information, read a raised-line graphic, and attend to the glove (for deictic resolution) [10]. Upon finalizing our design that satisfied our three basic conditions, we conducted a second set of experiments to evaluate the interaction configuration of a “instructor” (or guide) pointing at a visual on the board with the student using the glove on a parallel graphic, we devised a game we call the phrase charade that probes the capacity for such ac-

\[
\text{distance} = \sqrt{(TPA_i(y) - PIF(y))^2 + (TPA_i(x) - PIF(x))^2}
\]

(4)

DV is updated in real time as the cameras feed the system at a rate of 30 frames per second. The DV is then sent to the student through the haptic glove. Note that there is a very tightly closed feedback loop on the system. As new DV is computed, it is compared with the DV computed just before it. The new DV is delivered immediately replacing the directional signal being sent by the previous DV.

Table 1. Transforming PTA coordinates from the student’s camera coordinates system to the instructor’s camera coordinate system

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>Let TPAi in Figure 4 be the point originally tracked by the system.</td>
</tr>
<tr>
<td>2)</td>
<td>Let TPAi in Figure 5 be TPAt, translated to the instructor’s camera coordinate system.</td>
</tr>
<tr>
<td>3)</td>
<td>S0 through S3 in Figure 4, T0 through T3 in Figure 5 are known.</td>
</tr>
<tr>
<td>4)</td>
<td>Calculate ds1, ds2, dt1, dt2 in Figure 4.</td>
</tr>
<tr>
<td>5)</td>
<td>[w_1 = \frac{\text{ds}_1}{\text{dt}_1}, w_2 = \frac{\text{ds}_2}{\text{dt}_2}]</td>
</tr>
<tr>
<td>6)</td>
<td>[\text{dt}_1 = \text{tc}_1[x] - \text{tc}_0[x]; \text{dt}_2 = \text{tc}_2[y] - \text{tc}_0[y]; \text{dt}_3 = \text{tc}_3[x] - \text{tc}_1[x]; \text{dt}_4 = \text{tc}_2[y] - \text{tc}_1[y]]</td>
</tr>
<tr>
<td>7)</td>
<td>[\text{dt}_5 = \text{tc}_2[x] - \text{tc}_3[x]; \text{dt}_6 = \text{tc}_2[y] - \text{tc}_3[y]; \text{dt}_7 = \text{tc}_3[x] - \text{tc}_0[x]; \text{dt}_8 = \text{tc}_3[y] - \text{tc}_0[y]]</td>
</tr>
</tbody>
</table>
| 8)   | \[p_i[x] = \text{tc}_0[x] + \text{integer}(w_1 \times \text{dt}_5)\]
| 9)   | \[p_i[y] = \text{tc}_0[y] + \text{integer}(w_2 \times \text{dt}_6)\] |
| 10)  | \[p_i[x] = \text{tc}_3[x] + \text{integer}(w_1 \times \text{dt}_7)\]
| 11)  | \[p_i[y] = \text{tc}_0[y] + \text{integer}(w_2 \times \text{dt}_8)\] |
| 12)  | TPAt is the point where the line segments formed by 0i, p2i and 0i, p4i meet.

Figure 4. Frame captured from the student’s camera – its corners (sc0 sc3), are known.

Figure 5. Frame captured from the instructor’s camera – its corners (tc0 tc3), are known.
tive engagement. The critical aspect of these experiments is to ascertain the SBVI’s ability to handle the significant cognitive load of handling the mechanics of the system (listening, reading, and perceiving the ‘DV’) while engaging in fluid problem-solving/teaching-learning discourse. For reasons that will become evident, the study was performed twice [9].

In the game, the role of the instructor is replaced by a seeing guide who points into a poster that shows a matrix of letters. The SBVI has a braille version of the same diagram. A ‘clue phrase’ is hidden in the letter matrix (horizontally, vertically, or diagonally in either the forward or reverse direction). An example of a clue phrase is: “Blink Blink small sun”. The goal of the game is for the SBVI to follow the guide’s directions to find the braille version of the phrase, and to realize that it represents the common ‘catch phrase’: “Twinkle, twinkle, little star”. The guide and SBVI engage in speech during the game process. The game is designed to simulate the condition where an instructor provides directing discourse to the student, and where the student, after deciphering the instructional discourse has to perform the cognitive task of understanding what it means.

In Spring 2007, three SBVI were engaged to solve three phrase-charades each, giving us 9 discourse datasets supported by the HDS. These studies showed that while the SBVI were able to solve the charade, the interaction was very mechanical, and not at all fluid. Much of the discourse was directed at the use of the technology, and not at the charade problem. While one might argue that the ability to solve the charade at all is typically beyond the capacity of a SBVI, the interaction did not satisfy our goal of fluid discourse support to the degree at which it may be deployed in ‘inclusive classrooms’ where SBVI learn side-by-side with sighted students.

We hypothesized that while the SBVI were able to perform our three basic conditions for using the HDS, they required far more fluency with the interaction to enable fluid instruction-laden discourse. The HDS use had to become far more automatic to enable the student to devote more cognitive resources to learning. We developed an arcade-style game using the HDS that was made available to the SBVI at their leisure to play and gain the necessary embodied skill [9]. The development of the game and its deployment took a little more than a year.

In the Spring of 2009, we repeated the phrase-charade experiment. This time, we had 5 SBVI (two new participants were recruited in the interim). Again, each participant performed three charades. Owing to a technical malfunction, we had to exclude one dataset was excluded, yielding 8 datasets from the repeat participants and 6 from the new participants.

The performance times were markedly increased in the second charade. For our repeat subjects, the new average completion time was 238.50 seconds as opposed to 443.88 seconds from the 2007 study. This is a 1.86-fold time difference. A paired Student’s T test at a 95% confidence interval show the obvious significance of this result (paired Student’s t, p-value=0.0020). When we include all 5 subjects in the 2009 study, we have an average completion time of 251.92 seconds.

Table 2. Blind followers’ experience - A comparison between 1st and 2nd charade studies

<table>
<thead>
<tr>
<th>Grp</th>
<th>Question</th>
<th>1st</th>
<th>2nd</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>I felt comfortable using the glove</td>
<td>4.25</td>
<td>4.40</td>
</tr>
<tr>
<td>C</td>
<td>I’d rather use this glove than have someone physically move my hand to the letter</td>
<td>3.75</td>
<td>4.60</td>
</tr>
<tr>
<td>Cd</td>
<td>If I had someone physically holding my hand and putting it over the document, I would have performed better</td>
<td>3.75</td>
<td>2.80</td>
</tr>
<tr>
<td>Cd</td>
<td>I would perform better with practice</td>
<td>5.00</td>
<td>4.60</td>
</tr>
<tr>
<td>Cd</td>
<td>I would like to participate in future experiments because I believe this technology will help students who are blind</td>
<td>5.00</td>
<td>4.80</td>
</tr>
<tr>
<td>MT</td>
<td>I could perfectly listen to the guide while using the glove</td>
<td>4.50</td>
<td>5.00</td>
</tr>
<tr>
<td>I</td>
<td>Using the system did not interfere on my thinking of the solution</td>
<td>3.20</td>
<td>3.80</td>
</tr>
<tr>
<td>I</td>
<td>The conversation between the guide and myself flowed naturally</td>
<td>4.25</td>
<td>4.80</td>
</tr>
<tr>
<td>I</td>
<td>I was able to point at my chart and ask questions</td>
<td>4.25</td>
<td>4.60</td>
</tr>
<tr>
<td>I</td>
<td>I used pointing to reduce misunderstanding in what I said</td>
<td>3.25</td>
<td>4.00</td>
</tr>
<tr>
<td>I</td>
<td>Because of the system, I perceived that my communication was better understood</td>
<td>4.00</td>
<td>4.40</td>
</tr>
</tbody>
</table>

Our observations of the discourse content showed virtually no references to the technology in the 2009 study. Almost all the speech was dedicated to solving the charade. In other words, the average of conversational turns devoted to problem was 96.68% among the repeat subjects. In the first charade trial, the average was 66.70%, and significantly different (Paired Student’s t test, p-value<0.0001). When we include all participants, we have an average of 97.14% of the turns dedicated to problem solving.

Post-questionnaires were conducted verbally after both charade studies. These took the form of a set of statements to which the participants responded on an agreement Lickert scale (1=strongly disagree to 5=strongly agree, 3 being no opinion). Statements were kept the same to facilitate the comparison between both studies. Table 2 shows a selected
set of these questions and the averages of the answers for both first and second charade studies. The questions were grouped into five different categories: Comfort (C), Confidence (Cd), Multi-task (MT) and Interaction (I). We shall discuss the data on this table by groups.

By comfort, we mean one being comfortable with both wearing the haptic glove and interacting with the guide through the system. One can see that our participants are much more inclined to wear the glove than to have someone physically holding their hand, as normally happens in traditional instruction. As for confidence, we mean the confidence participants have that the system would bring gains in interactions similar to the charade. In this group, one can observe that after playing the game, participants are more inclined to believe that they would perform better using the system than with human help. We have argued earlier on the multi-modal, multi-task demands of a instructor/student-like interaction. In this group, numbers have also improved. When it comes to interaction, one can observe that our blind participants perceived that: 1) the interaction with the guide flowed more naturally, 2) they could also benefit from pointing, 3) the system helps the conversants.

5. Conclusion

We have presented our vision based system together with haptic gloves to support instructional discourse where the verbal component is situated within a graphical presentation with deixis. We demonstrate our system design, vision based approach and the results of our perception-action tests that indicated that the approach is able to provide direction in conjunction with speech and fingertip reading. When we applied the system to a real discourse situation with heavier cognitive requirements represented by our phrase charade game with our participants who are visual impaired, we found that the discourse was labored, and that there was inordinate attention paid to the technology, and insufficient resources were dedicated to the substance of the discourse. However, after a proper training with the HDS, the new results of our second charade study show an almost two-fold increase in speed over our first charade study. Furthermore, commitment of discourse to discussing the technology practically disappeared. The assistive technology was no longer the focus, and the problem at hand (the charade) and the dynamics of regular discourse maintenance dominated.

The speed of interaction and the capacity of our technology to assist communication between a seeing instructor and SBVI are critical in the eventual use of such technology in the classroom. This is made even more critical by the prevailing model of inclusive classrooms, where people with disabilities attend the same class as those who do not. The vision approach described in this paper were designed to satisfy the real-time, robustness, and cost constraints posed by the real-world requirements of such deployment. Our results show that the HDS that incorporates our vision systems have a positive impact on such deployment.

References