Towards Automated Understanding of Student-Tutor Interactions using Visual Deictic Gestures

Suchitra Sathyanarayana[1], Ravi Kumar Satzoda[1], Amber Carini[2], Monique Lee[2], Linda Salamanca[2], Judy Reilly[2], Deborah Forster[1], Marian Bartlett[1], Gwen Littlewort[1]
[1]University of California San Diego, [2]San Diego State University

Abstract

In this paper, we present techniques for automated understanding of tutor-student behavior through detecting visual deictic gestures, in the context of one-to-one mathematics tutoring. To the best knowledge of the authors, this is the first work in the area of intelligent tutoring systems, which focuses on spatial localization of deictic gestural activity, i.e. where the deictic gesture is pointing on the workspace. A new dataset called SDMATH is first introduced. The motivation for detecting deictic gestures and their spatial properties is established, followed by techniques for automatic localization of deictic gestures in a workspace. The techniques employ computer vision and machine learning steps such as GBVS saliency, binary morphology and HOG-SVM classification. It is shown that the method localizes the deictic tip with an accuracy of over 85 % accuracy for a cut off distance of 12 pixels. Furthermore, a detailed discussion using examples from the proposed dataset is presented on high-level inferences about the student-tutor interactions that can be derived from the integration of spatial and temporal localization of the deictic gestural activity using the proposed techniques.

1. Introduction

In recent years, there has been an increasing interest in one-to-one tutoring systems because it has been shown that one-to-one tutoring leads to an improvement in the performance of students by two standard deviations when compared to the traditional classroom setup [1]. This has motivated the increasing number of studies on developing intelligent tutoring systems [4, 21] with cognitive capabilities that are derived from understanding the social interactions between the student and tutor in one-to-one tutoring. Unlike existing online training modules in which there is little or no feedback between the online tutor and the student, intelligent tutoring systems are envisioned to mimic efficient tutor behavior and adapt to the requirements of the student in real-time. Therefore, studying the behavior of the tutor and the student, and the nature of their interaction in one-to-one tutoring has been of specific interest for developing such cognitive systems [21].

Major research efforts in this area have focused on understanding and modeling cognitive processes of one-to-one tutoring [3][4] using modalities such as text and speech [17]. Another major component of one-to-one tutoring that has been relatively less studied is the nonverbal behavior that accompanies speech, which includes hand gestures, facial expressions, nods, gaze etc. The role of hand gestures in tutoring has been well established [25] and recent studies have shown that gestures form a major modality in understanding tutor-student interactions [19][26]. In particular, it has been recently established that children learn mathematics better if the tutor uses gestures that reinforce and complement speech content [7]. Intricate analyses of the role of verbal and non-verbal elements of communication, in transferring information between participants and in driving observer understanding of an interaction, has revealed that gestures may modulate the meaning of speech, may provide independent information or may be redundant [23]. Gestures that contradict speech content are particularly interesting and have been demonstrated by Goldin-Meadows to be decrease learning in mathematics education settings [23].

The various hand gestures that can occur in the context of one-to-one tutoring are co-speech gestures [12] such as deictic, metaphoric, iconic and beat gestures; additionally writing can be considered [20]. Among the different hand gestures, deictic gestures (i.e. pointing for directing attention to a physical reference in course of a conversation [12]) were found to be of particular interest [25]. This is because deictic gestures and writing on the workspace constitute to more than 80% of the hand gestures [25] in a one-to-one tutoring setup. Max et al. [11] studied the role of deictic gestures in focusing visual attention and concluded that these gestures cannot be ignored in developing intelligent systems.

Computer Vision has been widely used for recogniz-

ing human hand gestures for applications such as human computer interface (HCI), virtual reality and robotics [6]. [15] is one of the earliest surveys on visual interpretation of hand gestures for HCI. A number of visual features in varying combinations have been used to identify the gestures such as model-based cues [24][13][16], motion based cues [10][22] and appearance based cues such as skin color [2][14], histogram of oriented gradients (HoG) [9] etc. Although there are a number of hand gesture recognition techniques, there is limited work done on recognizing hand gestures for tutoring systems. One of the earliest works in this area is [20], which describes appearance and motion based techniques for detecting deictic gestures. However, such methods limit the study to the detection of the presence or absence of deictic gestures only. The significance of spatial localization of the deictic gestural activity has not been explored in the past, especially in terms of extracting the location of the deictic gestures.

In this paper, we introduce novel techniques for spatial localization of the deictic gestures in one-to-one tutoring systems. We first introduce a multimodal dataset called SDMATH that is captured during one-to-one mathematics tutoring sessions. The dataset is richly annotated with speech functions, visual gestures, eye gaze, facial expressions etc. We then present robust techniques that employ visual saliency, image segmentation using binary morphology, and appearance based classification using histograms of oriented gradients (HOG) and support vector machines (SVM) to determine where a deictic gesture points to in the workspace, during the tutoring session. The proposed methods are evaluated for accuracy using the SDMATH dataset, and a detailed discussion is presented on determining higher-order inferences that can be made about the student-tutor interactions using the proposed techniques for spatial localization of the deictic gestural activity.

The paper is organized as follows. Firstly, the dataset that is used in this paper is introduced in Sec. 2. This is followed by the motivation for the localization of deictic gestures in Sec. 3. The proposed method to localize deictic gestures is outlined in Sec. 4. Results and discussion are subsequently presented in Sec. 5, after which the paper is concluded in Sec. 6.

2. Dataset

In a multi-year effort, mathematics tutoring audio-visual data was collected and annotated in great detail on multiple channels, such as speech, function, expression and gesture [5]. This data set (SDMath) will soon be available to the community. This dataset offers a set of richly labeled data with video and audio modalities. Four cameras are used to capture the videos in this dataset, one facing the tutor, another facing the student, a wide angle capturing both, and an aerial camera capturing hand gestures in the workspace.

Figure 1. Snapshot from a video in the SDMATH dataset; in clockwise direction starting from image on left: wideangle camera, overhead camera, student view, tutor view

Samples of this dataset are shown in Fig. 1.

The full SDMATH dataset consists of 20 videos capturing one-to-one mathematics tutoring sessions on the subject of logarithms. Two accredited middle school math teachers (1M, 1F) are the tutors, and 20 typically developing 8th graders (10M, 10F) were the participants or students. Each tutoring session was approximately one hour in duration and consisted of a 10 minute pretest, followed by a 40 minute tutoring session, and concluded with a 10 minute posttest. Video was collected simultaneously from all the four camera angles as explained above.

Different modalities of speech, gesture, eye gaze and facial expression were extensively hand-labeled using ELAN [18]. Transcriptions of the speech of the teacher and the student were labeled according to the contextual meaning of each speech unit within the session. Table 1 lists the different speech functions of the tutor that are used in the dataset [5]. Each tutoring session in the dataset is also divided into different problem segments. Additionally, the direction of eye gaze, Facial Action Coding System (FACS) units for facial expressions, and key gestures (such as hand, head nods etc.) were labeled for both teacher and student. Additional measurements of student and teacher FACS units were automatically extracted using the Computer Expression Recognition Toolbox (CERT).

Hand gestures constitute an important label channel for the video captured by the overhead camera that overlooks the workspace. The video from the overhead camera is manually labeled with different kinds of hand gestures such as deictic, beat, iconic, writing etc. Deictic gestures which form a key component of student-tutor interactions are the primary focus of the study presented in this paper.

3. Deictic Gestures in Mathematics Tutoring

In this section, the significance of detecting and localizing the deictic gestures in one-to-one tutoring is shown using specific examples from the proposed mathematics tutoring dataset.

Table 1. Tutor speech functions description in [5]

Function	Definition
Explanation	Provides information, without prompting the student for any response
Present Problem	Presents the student with a new problem to solve
Solicit Content	Prompts student for information about the problem
Solicit Explanation	Prompts student for information about how they arrived at a step in the problem (during or after student attempt)
Solicit Procedure	Prompts student for information related to how to work through the problem (prior to student attempt)
Request for Participation	Prompts student to accomplish a new step
Provides Hint	Provides incomplete information with an implicit request for student completion
Check for Comprehension	Ask the student about understanding of session-related material
Direct Negation	State that the student is wrong explicitly
Indirect Negation	Suggest that the student is wrong implicitly
Confirmation	Tell the student he/she is correct
Encouragement	Reassure or praise the student
Socializing	Conversation extraneous to session material

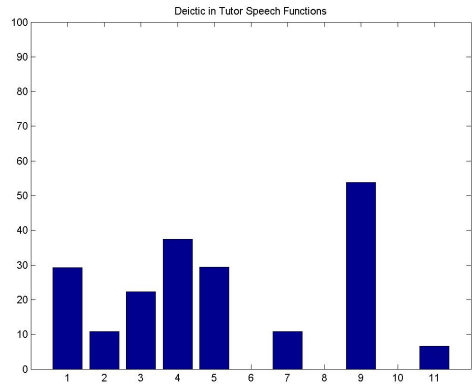

Figure 2. Percentage of cumulative time spent in deictic gesturing within each speech function plotted for an entire 1 hour long tutoring video; Speech functions: (1) prompting (2) confirmation (3) check for comprehension (4) explanation (5) indirect negation (6) encouragement (7) confirmation (8) socializing (9) direct negation (10) check for clarification (11) present problem

The first case is when there are gaps in speech, during which the tutor may communicate non-verbally by pointing at specific regions on the workspace. An example of this can be seen in video 12 of SDMATH dataset in problem segment 8 in the time window of 19 min 29 seconds to 19 min 32 seconds. This time window represents a gap in speech, sandwiched between the two speech functions of present problem ("so if we had x to the b divided by x to the a") and prompting ("guess what do you think") by the tutor. Prior to the prompting speech function, the tutor points at the worksheet where the corresponding logarithm expression was written by him, to support the prompting that follows. In the absence of such speech functions, determining the position of the deictic gesture aids in understanding the interactions better.

Another case where studying deictic gestures becomes

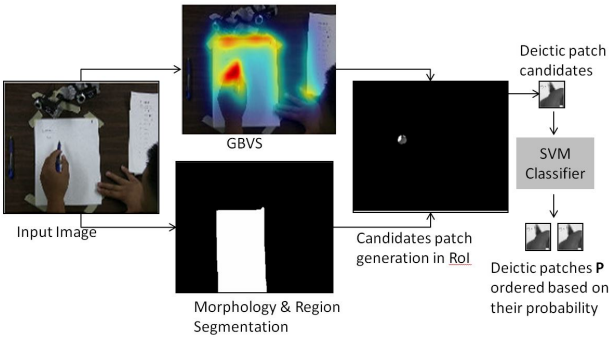

Figure 3. Deictic gesture location hypothesis generation step.

extremely important is when the tutor uses pronouns in his speech, while simultaneously pointing to what is written on the workspace. Without the knowledge of the spatial location that the deictic gesture points to, the information contained in the speech is incomplete. A typical example of this is when the tutor first writes the problem on the worksheet and then follows it up using speech such as "can you tell me what the log of *this* expression is?" accompanied by a deictic gesture that points to the written problem on the workspace.

A study of deictic gestures in relationship with the speech functions can also throw light on the cumulative trends that follow in a typical one-to-one tutoring session. Figure 2 shows a plot of the percentage time spent in deictic gestures by the tutor for each tutor speech function. This is plotted by summing the times for an entire hour-long tutoring session. Apart from explanation, it can be seen that deictic gestures heavily accompany speech functions that involve direct or indirect negation and prompting, all of which indicate instances where the student needs help in understanding the problem. Knowing where on the workspace the tutor is pointing at such time instances may aid in understanding the hot-spots in the learning process.

4. Proposed Method

In this section, we will describe the proposed method in detail. In this paper, it is assumed that the data is labeled as deictic, i.e. given a frame I, it is known that I has a deictic gesture but it is not known where the deictic gesture is. In the absence of such labeled data, the appearance and motion based techniques proposed by Suchitra et al in [20] can be used to determine that I has a deictic gesture.

The proposed algorithm has two steps. The first step is a hypothesis generation (HG) step, wherein possible candidate regions in the image where a deictic gesture occurs are extracted. The second step is the gesture localization step (GL) where spatial location of the deictic tip in the frame is determined (a deictic gesture in a frame is identified by the deictic tip that points to a spatial location in the frame [12]).

4.1. Deictic Gesture Location Hypothesis Generation

In order to hypothesize the location of the deictic gesture, the following property is used. It can be seen from the input image in Fig. 3 that in a deictic gesture, there is a pointing finger or pen that can be perceived as a protrusion from the rest of the hand and extending into the background (worksheet). Therefore, the pointing deictic gesture exhibits salient properties when compared to the background, in terms of the intensity variations and steep gradient changes.

In the first step of the proposed algorithm, we employ graph based visual saliency (GBVS) [8] to detect potential deictic gestural regions in the input image. In GBVS, the dissimilarity between the intensity of a pixel at (i, j) is compared with its surrounding pixels, and a graph is generated with weights of the edges w_1 that are determined using the following equations:

$$\begin{aligned} d((i,j)||(p,q)) &\triangleq \left| log \frac{P(i,j)}{P(p,q)} \right| \\ w_1((i,j),(p,q)) &\triangleq d((i,j)||(p,q)) \cdot F(i-p, j-q) \\ F(a,b) &\triangleq exp\left(-\frac{a^2+b^2}{2\sigma^2}\right) \end{aligned} \tag{1}$$

On applying the above formulations on the entire image followed by a Markovian modeling of each node, a heat map I_S is generated showing the most salient regions as shown in Fig. 3. On applying a threshold T_S on I_S, a set of blobs $B = \{B_i\}$ with each blob centered at C_{B_i} that correspond to the salient regions, are obtained.

Considering that the deictic gestures are pointed on the workspace, the region of interest (RoI) is selected as the worksheet on which the tutor is explaining the student. This eliminates false positives that may be shown as salient regions by GBVS. For example, the heat map shown in Fig. 3 shows the edges of the sheet also as salient regions but they do not correspond to the deictic gesture occurring within the worksheet. Therefore, such false salient regions need to be eliminated for the next steps. In order to do this, image morphology is used to select the sheet as the region of interest. Fig. 4 shows the steps graphically. The following steps are undertaken to get the region of interest (RoI):

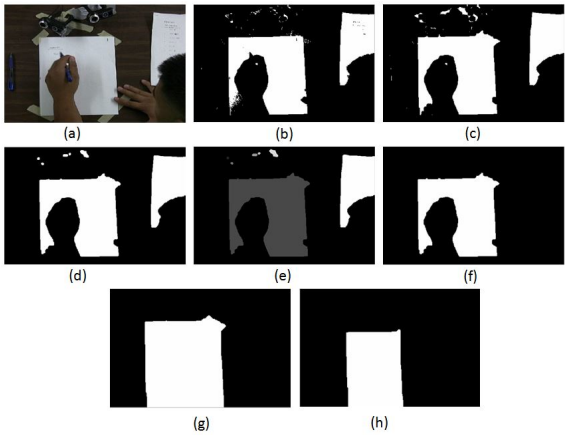

Figure 4. Steps in segmentation of workspace; (a)Input image (b) Binarization (c) Closing to fill in gaps (d) Opening to remove noisy blobs (e) Connected components labeling (f) Identify biggest component (g) Closing with a big structural element to remove hand (h) Erosion to remove borders

$$\begin{aligned} \text{1-Binarization: } I_b &= I > T_i \\ \text{2-Closing: } I_b &= I_b \bullet S_{e1} \\ \text{3-Opening: } I_b &= I_b \circ S_{e2} \\ \text{4-CCA: } S_B &= set\ \ of\ \ components \\ \text{5-Closing: } I'_b &= I'_b \bullet S_{e3} \\ \text{6-Erosion: } I'_b &= I'_b \bullet S_{e4} \end{aligned} \tag{2}$$

where: S_{e1} and S_{e2} are structuring elements such that size of S_{e1} is greater than S_{e2}; CCA denotes connected component analysis on I_B resulting in a set of connected components S_B; I'_b is the binarized with the largest connected component in S_B; S_{e3} is a large structuring element such that it fills the large hole in the workspace I'_b due to the hand; and S_{e4} is a small structuring element to remove edges of the worksheet. The above steps result in the workspace or the region of interest (RoI) as shown in Fig. 4(h).

Combining the above RoI with blobs B obtained from saliency, the blobs in B with centroids inside the RoI only are selected for further processing. We consider a bounding box of size $h_B \times w_B$ (rows-by-columns) around each centroid of the blob and a set of image patches $\mathbf{P} = \{\mathbf{I_{B_i}}\}$ is extracted, where each $\mathbf{I_{B_i}}$ corresponds to an image patch around C_{B_i} in I.

Considering that the GBVS could generate salient regions that may not contain the deictic gesture in I, an elimination step is introduced to select the most probable candidate patch in $\mathbf{P}$ that has the deictic gesture. In order to do this, HOG feature is computed for each $\mathbf{I_{B_i}}$. HOG feature was selected because a study of the tutoring videos shows that deictic gestures show a strong appearance correlation in terms of the gradients of the edges of the fingers and objects such as pens that form the deictic gesture. An SVM (support vector machine) classifier is then used to determine the score of the each $\mathbf{B_i}$ being one of the two classes - deictic or non-deictic patch in the following way:

$$p_{B_i} = \mathbf{w}^T \mathbf{x}_{B_i} + b \tag{3}$$

where $\mathbf{w}^T$ is the SVM model vector obtained from training, $\mathbf{x}_{B_i}$ is the HOG feature vector and p_{B_i} is the score of

B_i containing the deictic gesture. Therefore, a set of scores $P = \{p_j\}$ is generated comprising of scores determined using (3) for each B_i in I such that $p_j > p_{j+1}$. In other words, we arrange the image patches in $\mathbf{P}$ in decreasing order of the score of the image patch $\mathbf{I_{B_i}}$ containing a deictic gesture. Fig. 3 shows the different steps involved in the first stage of the proposed method.

It is to be noted that the scores p_{B_i} generated in (3) gives the score whether the patch contains a deictic gesture tip. However, it does not localize the deictic tip itself. In the next step, we localize the deictic tip in the patches that are order according to p_{B_i} in an orderly manner.

4.2. Gesture Localization Step

Let $\mathbf{P'} = \{\mathbf{I_{B_j}}\}$ be the set of order patches based on their scores computed in (3). We call such patches deictic patches. Starting from the patch with highest score of being a deictic patch, each patch in $\mathbf{P'}$ is examined to extract the location of the deictic gesture tip. If a patch $\mathbf{I_{B_j}}$ does not give a deictic gesture with acceptable score, the next patch in $\mathbf{P'}$ is examined for the deictic gesture tip.

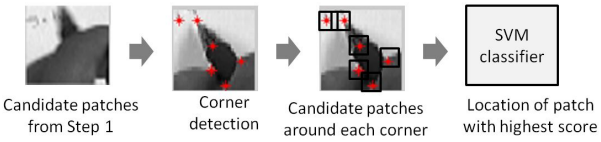

Figure 5. Gesture localization step.

Given a deictic patch $\mathbf{I_{B_j}}$, a set of corners $S_C = \{S_k(x_k, y_k)\}$ are detected using the following formulation:

$$C = \frac{I_x I_y}{I_{xy}} \tag{4}$$

where I_x, I_y and I_{xy} are the gradients along x, y and both $x-y$ directions respectively. Corners are selected based on the cornerness measure T_c, i.e. if $C > T_c$ at (x, y), then (x, y) is considered a corner. Each corner selected in S_C is then subjected to deictic tip classification. We consider a $w_c \times w_c$ window I_C around each $S_k(x_k, y_k) \in S_C$ and a HOG feature is computed I_C, which is classified using SVM. Using a similar formulation as (3), a set of scores $P_C = \{p_k\}$ is generated with p_k giving the score of the corner S_k being a deictic tip.

If $p_k > T_t$ then the corner is considered to be a deictic tip. If multiple corners are found in the image patch $\mathbf{I_{B_j}}$ which are classified as deictic tips, then the corner with the higher score p_k is considered as the correct deictic gesture tip in $\mathbf{I_{B_j}}$. If an image patch $\mathbf{I_{B_j}} \in \mathbf{P'}$ does not give any corners that satisfy scores greater than T_t, then the next deictic image patch in $\mathbf{P'}$ is considered for processing.

5. Performance Evaluation & Discussion

In this section, the performance of the proposed method is evaluated. As described in the previous section, one of the objectives of the work presented in this paper is to determine the visual localization of the deictic gestural activity in input video frames, given that the data is labeled as deictic. In other words, each frame of the input frame is supposed to have a deictic gesture and the proposed algorithm described in Section 4 should determine the location of the tip of the deictic gesture.

Fig. 6 shows some sample frames with results from the proposed method marked in them. The top row shows frames with image patches marked in blue and red. A blue box indicates a patch that is classified as a deictic patch. A red box indicates a patch is shown as a salient region by GBVS but classified as a non-deictic patch by the SVM classifier in the first stage of the proposed algorithm.

The second row shows sample results of the deictic gesture localization step, i.e. detecting the tip of the deictic gesture in the second stage of the proposed method. The blue filled box indicates the estimated position from the second stage classifier that determines if a corner point detected in the patch is a deictic tip.

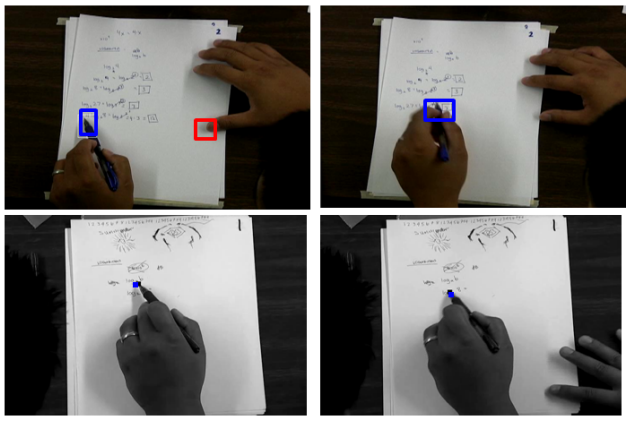

Figure 6. Sample results of the proposed technique. Top row: boxes showing patches that are selected as deictic patches by the first step of proposed algorithm. Blue boxes indicate true positive deictic patches, whereas red boxes indicate salient patches from GBVS that are eliminated as true negatives after the first stage of the proposed algorithm. Bottom row: Deictic tip detection results by the proposed algorithm, which are classified in the second stage.

In Fig. 7, true positive rate (TPR) is plotted against the distance between the positions estimated from the proposed algorithm and the annotations in ground truth. As described earlier in this section, all the frames are supposed to have a deictic gesture tip. Therefore, true positive rate is determined by finding number of frames, in which the deictic

tip is found to be within a distance $d_e = ||S^{GT} - S^{EX}||$, where S^{GT} is the the coordinate of the decitic tip from the ground truth annotation, and S^{EX} is the estimated coordinate of the tip from the proposed method. All the points that are outside d_e are considered as false positives. Fig. 7 plots TPR against varying d_e in pixels for an input image frame resolution of 720×480. The testing was done on 1600 frames containing deictic gestures, and none of the 1600 frames were used for training either of the two classifiers in the proposed method. Additionally, the 1600 frames for testing are taken from a different tutoring session, which was not used for extracting training frames. All ground truth annotations were done manually. It can be seen that a TPR of over 85% is achieved for $d_e \leq 12$, i.e. in over 85% of the frames, the proposed algorithm estimates the deictic tip within 12 pixels (euclidean distance) from the ground truth annotation. Considering that this analysis is the first of its kind, the proposed annotated dataset will be made available in public domain for future evaluations and comparisons.

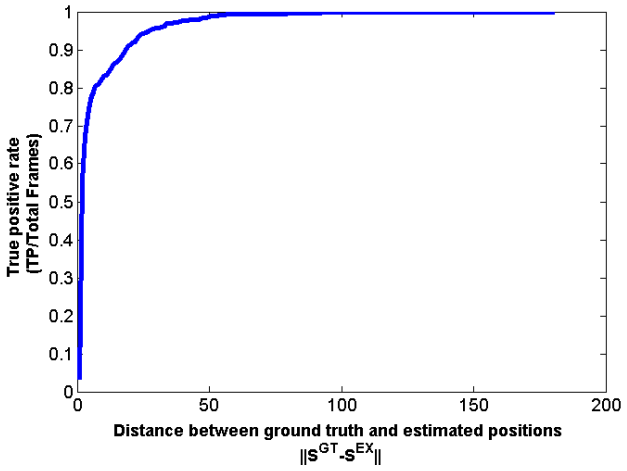

Figure 7. Accuracy curve: true positive rate versus distance from ground truth.

5.1. Discussion

In this sub-section, we present a discussion on higher-order inferences that can be made about the interactions between the student and tutor using the spatial information of deictic gestural activity. We present different future directions in which the proposed study in this paper can be directed towards developing intelligent tutoring systems.

Firstly, the spatial locations of the deictic gestures can point towards hotspots in the learning and teaching process during the tutoring session. For example, consider Fig. 8 which shows the last frame of problem segment that is overlaid with markers at the coordinates where deictic tips are found in a sequence of 300 frames that are sampled from the problem segment. While Fig. 8(a) shows the markers at hand-annotated ground truth positions, Fig. 8(b) shows the markers that are marked at the coordinates obtained from the proposed method for the same set of frames. The first observation that can be made from Fig. 8 is that the proposed automated vision-based method shows similar clusters of deictic activity in spatial domain as compared to manually labeled ground truth annotations. Therefore, the proposed automated method can be used reliably to understand the student-tutor interactions that involve deictic gestural activity. Next, the clusters in Fig. 8 also point towards some insightful conclusions about the teaching and learning process during the one-to-one tutoring session. The spatial locations of the cluster centers can be used to identify the specific parts of the problem segment that involved a higher degree of interaction between the student and the tutor. Given the definition of the deictic activity, a cluster of deictic tips centered around a particular spatial position indicates that the emphasis that the tutor is giving to specific parts of the problem segment. The density of the clusters also indicates the amount of time spent on a specific part of the problem, which can be used to extract specific concepts or problems where the student and tutor spent more time in learning/teaching. The spatial locations thus obtained could also be related with other modalities such as speech functions, and other visual gestures such as nods etc. to further understand the interactions between the student and the tutor.

Secondly, the spatial information of the deictic gesture can be combined with the temporal information to understand and even evaluate the teaching process of the tutor (and the learning process of the student). This is demonstrated using the example segment shown in Fig. 9. The blue circles show the position of the deictic tip that is detected by the proposed method and the red line is the trace of the deictic tip obtained by joining the blue circles. It can be seen from Fig. 9 that the tutor is referring back to what he has written before (P_2 on the workspace in Fig. 9) on the workspace to explain a concept (at P_1 on the workspace in Fig. 9) to the student. Therefore, the integration of spatial and temporal information of the deictic gesture can be used to develop techniques that can extract effective teaching/teacher behaviors, and the interactions involved therein between the student and tutor.

Thirdly, another important observation that can be made is about the salient movements that the tutor does while showing the deictic gesture. We explain this with Fig. 10, which shows a similar deitic trace as explained above. The important observation in this trace is that the tutor is moving from point P_1 to P_2, where the deictic trace can be seen like a circle. A visual inspection of the video shows that the tutor is pointing to the problem at P_2 and circling around P_2 while maintaining the deictic gesture. The deictic trace generated by the proposed method is able to capture that move-

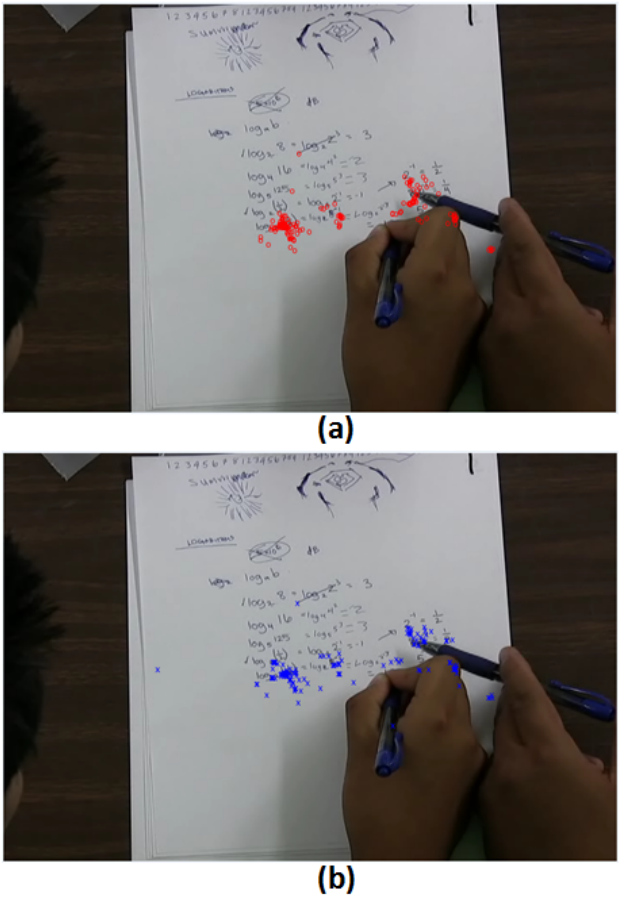

Figure 8. Cluster of deictic tips for a problem segment involving 300 frames with deictic gesture (a) from ground truth annotation, (b) estimated by proposed method.

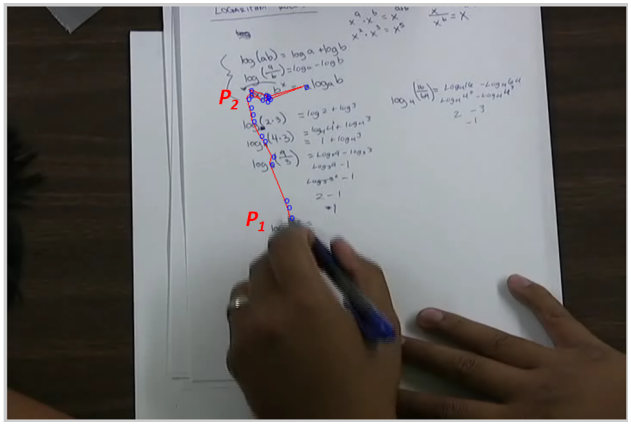

Figure 9. Deictic trace showing movement across the problem explanation.

ment of the tutor Fig. 10. Further studies can be done to relate this visual inferences to higher order semantics about the interactions and behaviors of the tutor and student during the one-to-one tutoring session. For example, circling could mean over emphasis on a particular problem, etc.

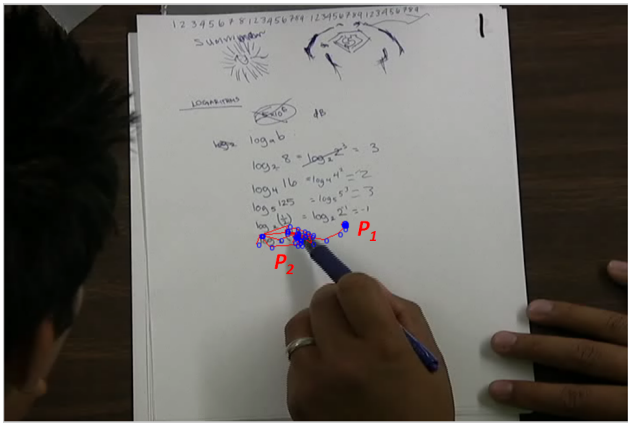

Figure 10. Deictic trace showing salient movements. Notice the deitic trace is like a circle at P_2.

6. Conclusions

This paper introduces the significance of understanding the *where* of deictic gesture in one-to-one tutoring systems. We have proposed novel techniques that robustly localize the deictic gestures spatially on the workspace. The proposed techniques were also evaluated on the SDMATH dataset and show over 85% accuracy. We demonstrated with examples and detailed discussions how the spatial position of deictic gestural activity combined with its temporal information can be used to determine the different characteristics of the teaching process of the tutor, learning behavior of the student, learning/teaching hot spots in a tutoring session etc. It has been shown that the visual localization of the deictic gestural activity can also aid in inferring non-verbal interactions between the student and tutor, which is especially significant in the absence of speech or in the presence of ambiguous speech. Also, the proposed techniques can be further combined with other modalities and gestures to understand the interactions and learning/teaching methodologies better. The proposed work is the first step towards many interesting future directions, some of which have been discussed in this paper.

References

[1] B. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. In *Educational Researcher*, volume 13, page 416, 1984. 1

[2] L. Bretzner, I. Laptev, and T. Lindeberg. Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 423–428, 2002. 2

[3] V. K. Conati. C, Gertner. A. Using bayesian networks to manage uncertainty in student modeling. In *User modeling and user-adapted interaction*, volume 12, page 371417, 2002. 1

[4] K. K. Corbett. A and A. J. R. Intelligent tutoring systems. In *Handbook of humancomputer interaction*, page 849874, 1997. 1

[5] K. Dykstra, L. Salamanca, M. Salamanca, A. Carini, M. Lee, D. Forster, G. Littlewort, J. Reilly, M. Bartlett, and J. Whitehill. Sdmath: A dataset to inform affect-aware intelligent tutoring systems. *Submitted for Journal publication*. 2, 3

[6] A. Elgammal, V. Shet, Y. Yacoob, and L. S. Davis. Learning dynamics for exemplar-based gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 571–578, 2003. 2

[7] L. Goldin-Meadow, S. and S. Jacobs. Gestures role in learning arithmetic. *Emerging perspectives on gesture and embodiment in mathematics, (in press)*. 1

[8] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 19*, pages 545–552. MIT Press, 2007. 4

[9] M. Kaaniche and F. Bremond. Recognizing gestures by learning local motion signatures of hog descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2247–2258, 2012. 2

[10] D. Kelly, J. McDonald, and C. Markham. Continuous recognition of motion based gestures in sign language. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1073–1080, 2009. 2

[11] M. M. Louwerse and A. Bangerter. Focusing attention with deictic gestures and linguistic expressions. 2005. 1

[12] D. McNeill. Gesture: A psycholinguistic approach. 1, 3

[13] L. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007. 2

[14] E. Ohn-Bar and M. M. Trivedi. The power is in your hands: 3d analysis of hand gestures in naturalistic video. In *Computer Vision and Pattern Recognition Workshops*, 2013. 2

[15] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:677–695, 1997. 2

[16] J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *In ICCV*, pages 612–617, 1995. 2

[17] C. P. Rosé, D. Litman, D. Bhembe, K. Forbes, S. Silliman, R. Srivastava, and K. VanLehn. A comparison of tutor and student behavior in speech versus text based tutoring. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing - Volume 2*, HLT-NAACL-EDUC '03, pages 30–37, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. 1

[18] I. Rosenfelder. A short introduction to transcribing with elan. Technical report, University of Pennsylvania, 2011. 2

[19] A. Sarrafzadeh, S. Alexander, F. Dadgostar, C. Fan, and A. Bigdeli. See Me, Teach Me: Facial Expression and Gesture Recognition for Intelligent Tutoring Systems. *2006 Innovations in Information Technology*, pages 1–5, Nov. 2006. 1

[20] S. Sathyanarayana, G. Littlewort, and M. Bartlett. Hand gestures for intelligent tutoring systems: Dataset, techniques amp;amp; evaluation. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 769–776, Dec 2013. 1, 2, 3

[21] G. P. Schiaffino, S. and A. Amandi. eTeacher: Providing personalized assistance to e-learning students. *Computers and Education*, 51:1744–1754, 2008. 1

[22] X. Shen, G. Hua, L. Williams, and Y. Wu. Dynamic hand gesture recognition: An exemplar-based approach from motion divergence fields. *Image and Vision Computing*, 30(3):227–235, Mar. 2012. 2

[23] M. A. Singer and S. Goldin-Meadow. Children learn when their teacher s gestures and speech differ. *Psychological Science*, 16(2):85–89, 2005. 1

[24] S. B. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1521–1527, 2006. 2

[25] B. Williams, C. Williams, N. Volgas, B. Yuan, and N. Person. Examining the role of gestures in expert tutoring. In *Proceedings of the 10th international conference on Intelligent Tutoring Systems - Volume Part I*, ITS'10, pages 235–244, Berlin, Heidelberg, 2010. Springer-Verlag. 1

[26] J. R. Zhang, K. Guo, C. Herwana, and J. R. Kender. Annotation and taxonomy of gestures in lecture videos. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 1–8, June 2010. 1