Hand Gestures for Intelligent Tutoring Systems: Dataset, Techniques & Evaluation

Suchitra Sathyanarayana, Gwen Littlewort, Marnie Bartlett
University of California, San Diego
Gilman Drive, La Jolla.
ssathyanarayana@ucsd.edu, gwen@mplab.ucsd.edu

Abstract

Analysis of hand gestures in one-to-one tutoring gives a number of characteristics of social interaction and behavior between the tutor and the student. This analysis can not only aid in understanding the effectiveness of the learning methodology and developing new techniques for learning, but also help in developing intelligent and online tutoring systems. Although there exists a comprehensive literature on recognizing hand gestures, there is limited work on recognizing such gestures in the context of one-to-one tutoring systems. In this paper, we first introduce a new dataset that comprises a set of 2166 richly labeled video sequences of multiple subjects, showing 4 different classes of most prominent gestures in one-to-one tutoring. In addition to the dataset, two methods comprising appearance based cues and motion based cues are proposed and evaluated on this dataset. A detection accuracy of over 53% is achieved when the proposed techniques are validated across 6 different subjects, which can be used as a benchmark for future works that can employ the proposed datasets for hand gestures for one-to-one tutoring systems.

1. Introduction

One-to-one tutoring has received recent attention after it was established that this form of tutoring leads to an improvement in the performance of the students by two standard deviations, compared to the conventional group classroom setup [2]. One-to-one tutoring involves extensive social interaction between the tutor and the student, and decoding subtle cues from such social interactions has been of great interest, especially to aid in the development of intelligent and adaptive tutoring systems [6].

Major research efforts in this area have focused on understanding and modeling cognitive processes of one-to-one tutoring [5][6], and the common modalities used are text and speech [22]. 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. It has been established that behaviors that improve social rapport increase information transfer between individuals, and can specifically affect the efficacy of teaching [8]. The role of hand gestures in tutoring has been well established [30] and recent studies have shown that gestures form a major modality in understanding tutor-student interactions [24][31]. This paper focuses on hand gestures commonly used in one-to-one tutoring.

Computer Vision has been widely used for recognizing human hand gestures for applications such as human computer interface (HCI), virtual reality and robotics [7]. [20] is one of the earliest sruveys on visual interpretation of hand gestures for HCI, in which techniques for gesture modeling, analysis and recognition are discussed in detail. A number of visual features in varying combinations have been used to identify the gestures. This includes model-based cues [29][17][21], motion based cues [11][25] and appearance based cues such as skin color [4][18], histogram of oriented gradients (HoG) [9][10] etc. Model-based approaches rely on three dimensional representation of body parts, e.g., [21], whereas appearance-based approaches use two dimensional information such as gray scale images or body silhouettes and edges [26]. IOn the other hand, motion based approaches attempt to recognize the gesture directly from the motion without any structural information about the physical body, e.g., [3]. In all these approaches, the temporal properties of the gesture are typically handled using approaches such as Hidden Markov Models (HMM) or Conditional Random Fields [29]. In [17], a discriminative framework that incorporates hidden state variables is used for continuous gesture sequence segmentation. Skin color detection is one of the most popular methods for hand localization [18]. The skin color cues are combined with the motion cues [4] for improving the efficiency of hand detection. Modified HOG based feature is explored in [18] that introduces a two-dimensional HOG or HOG^2 for hand gesture recognition in vehicle systems. A recent example of a model-based method is described in [19] which employs geometry based normalizations and Krawtchouk moments to locate and identify hand gestures in a rotation invariant manner and in varying backgrounds. Local and global motion based methods using descriptors such as SIFT are discussed in [9]. Both these motion based methods are combined together in [9] to get higher accuracy in gesture identification. An exemplar based gesture recognition method is proposed in [7], which represents each gesture as a sequence of body poses (exemplars) through a probabilistic framework for matching these body poses to the the image data.

Although there are a number of hand gesture recognition techniques, there is limited work done on recognizing hand gestures for tutoring systems. Given that the tutoring process involves a variety of hand gestures that can be directly correlated to interactions and social behavioral patterns between the tutor and the pupil, detecting such hand gestures in one-to-one tutoring environments can aid in developing intelligent systems to evaluate the learning methods, interactions and ultimately improving the education process.

In this paper, techniques that are based on appearance and motion history are explored for automatic detection of the most common gestures in the context of one-to-one tutoring. The paper is organised as follows: The types of gestures that are seen in a one-to-one tutoring are first described in Section 2, followed by a detailed description of the dataset that is used in this paper, along with the process of generating it, in Section 3. This is then followed by the techniques that have been employed for gesture recognition in Section 4, after which the evaluation results are presented in Section 5. A discussion on the relevance of this work for social behavior has been included in Section 6, before concluding the paper in Section 7.

2. Types of gestures in one-to-one tutoring

There are many kinds of gestures, including hand gestures, nods and eye gaze, that are commonly used in a one-to-one tutoring system, and these have been studied in detail by William et al. [30]. This paper focuses on automatically recognising hand gestures commonly used in the context of one-to-one tutoring sessions. Hand gestures accompanying speech are termed as speech-gestures [15] and considering that speech is a modality that almost always accompanies gestures in a tutoring set up, speech-gestures are commonly analysed in this context. McNeill [15][16] distinguishes the following four major types of gestures by their relationship to the speech:

- Deictic: refers to the pointing gesture often associated with the index finger, but not limited to it. Deictic gestures are used to direct a listeners attention to a physical reference in course of a conversation. In face-to-face conversation these gestures mostly are limited to the pointing in and often used in a reference to the imaginary placeholder.

- Iconic: Iconic gestures have close relationship to the semantic content of speech. In McNeills definition, the iconicity of gesture is determined by exhibiting the aspects of the same scene described by speech.

- Beats: Beat gestures are possibly the most spontaneous and the smallest hand movement resembling flicks. Unlike other speech gestures, beats are not associated with any particular meaning and they occurs with the rhythm of the speech, mostly placed on stressed syllables.

- Metaphoric: Metaphoric gestures are associated with abstract ideas. Similar to iconic gestures in pictorial manifestation, they represent a metaphor of the speakers idea or feeling about a concrete concept.

Apart from the speech-gestures, other visual gestures such as ‘writing’ or ‘fidgeting’ are expected to commonly occur in tutoring videos. William et al. [30] investigated the various hand gestures involved in a typical one-to-one tutoring system and concluded that *‘deictic’* gestures and *‘writing’* on the work-space are the two main hand gestures, constituting to more than 80% of the hand gestures used, respectively. Max et al. [13] studied the role of deictic gestures in focusing visual attention and conclude that these gestures cannot be ignored in developing intelligent tutoring systems.

Automatic vision-based hand gesture detection in tutoring has a number of challenges. Firstly, the manual labeling of videos for specialized gestures, such as speech gestures, requires the experience of experts in this field. The labeling process involves isolating video segments containing meaningful gestural activity and discarding portions that are not useful. Secondly, the same gesture is represented by a wide range of hand configurations and appearances. For example, although a deictic gesture is typically represented by the pointing of the index finger to the object of interest, other fingers can be used for the same purpose.

3. About the dataset

The hand gesture dataset Tutor-Gesture considered in this paper is part of a bigger multi-modal tutoring dataset that is due for release soon. The full tutoring dataset offers a set of richly labelled data with video and audio modalities, captured using a 4-camera set up, one facing the tutor, another facing the student, a wide angle capturing both, and an aerial camera capturing hand gestures. A sample from this

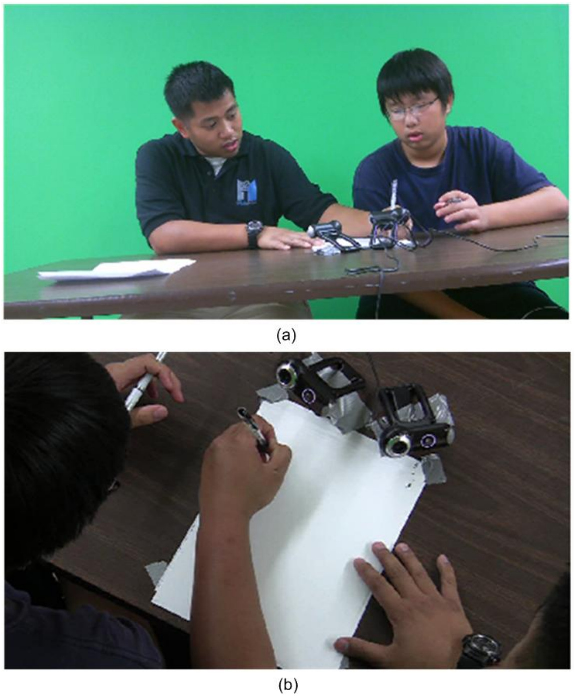


Figure 1. Sample snapshots from the tutoring dataset (a) wide frontal view camera (b) overhead view camera

dataset is shown in Fig. 1, with snapshots from the wide frontal and overhead views.

The full tutoring dataset consists of 20 videos capturing one-to-one mathematics tutoring sessions on the subject of logarithms. The tutors were two accredited middle school math teachers (1M, 1F) and the participants were 20 typically developing 8th graders (10M, 10F). 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 4 camera angles as explained above; a wide frontal view of both the teacher and student, close-up views of the student and teacher faces, and an overhead view which captured the shared workspace. The dataset was extensively hand-labeled in the modalities of speech, gesture, eye gaze and facial expression using ELAN [23]. The speech of the teacher and the student was transcribed and then labeled according to the contextual meaning of each speech unit within the session. Eye gaze direction, Facial Action Coding System (FACS) units, and key gestures were labeled for both teacher and student for the duration of each session. Additional measurements of student and teacher FACS units were automatically extracted using the Computer Expression Recognition Toolbox (CERT).

The Tutor-Gesture dataset used in this paper was derived from the bigger tutoring dataset and currently consists of over 2300 small video clips, of length ranging from few seconds to a few minutes. It was generated using the overhead view camera capturing the workspace and the hand gestures. ELAN label files were used to automatically extract the small video clips corresponding to the length of each gestural activity. The process of generating the dataset included the following steps:

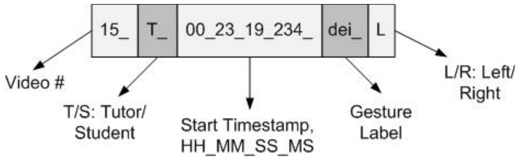


Figure 2. Labeling scheme followed for the Tutor-Gesture dataset, with an example

1. The timestamps of the gestural activity were extracted using the ELAN label files.

2. MATLAB scripts were written to automatically generate commands for commandline video-editing software called FFMPEG, using the timestamps extracted above. These commands were saved as batch files.

3. The batch files were run, resulting in extraction of small clips from the original overhead view camera.

4. This was repeated for 6 videos of the same tutor but with 6 different students - 3 Male and 3 Female.

The labeling scheme that has been employed for the Tutor-Gesture dataset has been illustrated in Fig. 2 with an example. The label starts with the index of the video number. The letter ‘S’ or ‘T’ has been used to indicate whether the gesture belongs to the student or the tutor, respectively. This is followed by the start timestamp of the segment (in HH:MM:SS:MS) format in the original video. The label of the gesture follows next, which include *dei, ico, bea, met, wri* to represent deictic, iconic, beat, metaphoric and writing, respectively. This is finally followed by the letter ‘R’ or ‘L’ to indicate whether it is the right or the left hand that has been used for the gesture.

The distribution of the gesture types in this dataset has been shown in Table 1. In alignment with the observation pointed out in [30], it can be seen that deictic and writing gestures form the majority of the gestures that transpire during the tutoring sessions. It is for this reason that in this paper, the focus has been limited to automatically recognising deictic and writing gestures. The tutor considered in this Gesture Dataset is left-handed, while all the six students are right-handed. The left and right handed deictic and writing gestures, therefore form the four classes considered for classification, namely, dei_L, dei_R, wri_L and wri_R, respectively. These account to a total of 2166 video clips. Sample

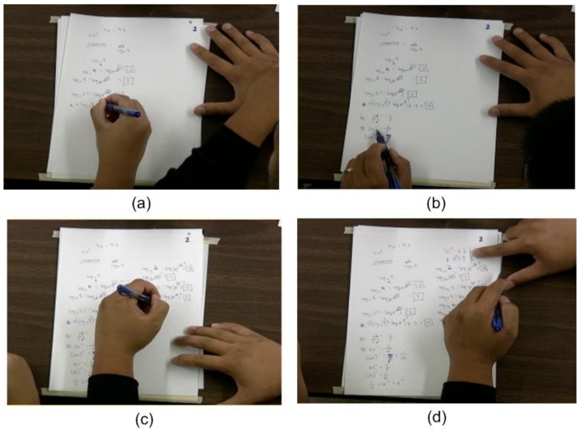


Figure 3. Snapshots of the four types of gestures that will be analyzed in this paper, along with the labels (a): $13S\_00\_04\_28\_548\_wri_R$, (b): $13T\_00\_08\_25\_048\_dei\_L$, (c): $13T\_00\_16\_03\_368\_wri_L$ and (d): $13T\_00\_19\_42\_698\_dei\_R$

	V1	V2	V3	V4	V5	V6
Deictic	167	157	246	231	194	219
Writing	173	140	154	173	152	158
Iconic	15	19	30	8	2	5
Beat	4	9	22	18	17	11
Metaphoric	-	-	3	2	1	2

Table 1. Distribution of hand gestures in Tutor-Gesture dataset

snapshots of the four gesture types, along with their labels are illustrated in Fig. 3.

4. Techniques for Hand Gesture Classification

In this section, techniques are proposed for classifying hand gestures in datasets generated during the one-to-one tutoring sessions. Two different kinds of hand gesture classification techniques are proposed, which will be evaluated using the abovementioned datasets. We will describe an appearance based classification technique, followed by a motion based technique for classifying the four different kinds of hand gestures in one-to-one tutoring sessions that were described in Section 3.

4.1. Bag of Words on SIFT

We use a bag of words (BoW) model on dense SIFT features [27] to generate appearance based descriptors, that are fed into an SVM for recognition. The strength of the proposed approach lies in the combination of dense feature sampling, implicit inclusion of spatial information through a pooling step using spatial grids [12], and state-of-the-art feature encoding using Locality-constrained Linear Coding [28].

In each video segment of the training dataset, we consider the middle frame of the segment for generating appearance based features. First, we sampled dense SIFT features [14][27] using a stride of 4 pixels. The codebook for BoW was generated using approximate K-means clustering, a clustering approach that employed data-to-cluster distances using the Approximate Nearest Neighbor algorithm. Once the codebook of cluster centers was generated, each local SIFT feature was assigned to a codeword using Locality-constrained Linear Coding (LLC) [28]. LLC projected each descriptor to a local linear subspace spanned by a selection of the codewords using an optimization problem.

The traditional Bag of Words model is robust to spatial translation, but sacrifices spatial layout information during the histogramming process. Spatial Pyramid Matching (SPM) implicitly incorporates spatial information into the feature representation through histogramming within different subdivisions of the image [12]. For SPM each image was partitioned into 2 × 2 segments. The BoW representation was then computed within each of these segments, and all of the subsequent BoW histograms were concatenated into a single feature vector. The features are then sent for linear SVM for training. Fig. 4 shows the feature generation and learning steps of the appearance based method employed in this paper.

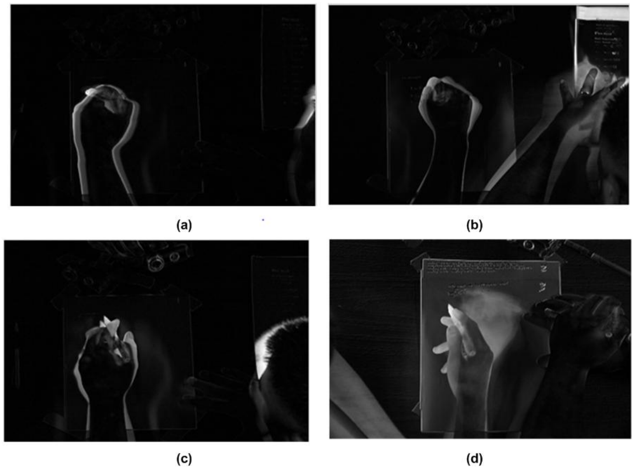


Figure 5. Modified MHIs generated for the four different classes: (a) Writing left, (b) Writing right, (c) Deictic left, and (d) Deictic right.

4.2. Modified Motion History Image

In this subsection, we will present a motion-based technique for hand gesture recognition in one-to-one tutoring sessions. We employ two modified versions of motion histogram image (MHI) [3][1] to describe the four different hand gestures in one-to-one tutoring. The motion history image (MHI) approach is a view-based temporal template method which is simple but robust in representing move-

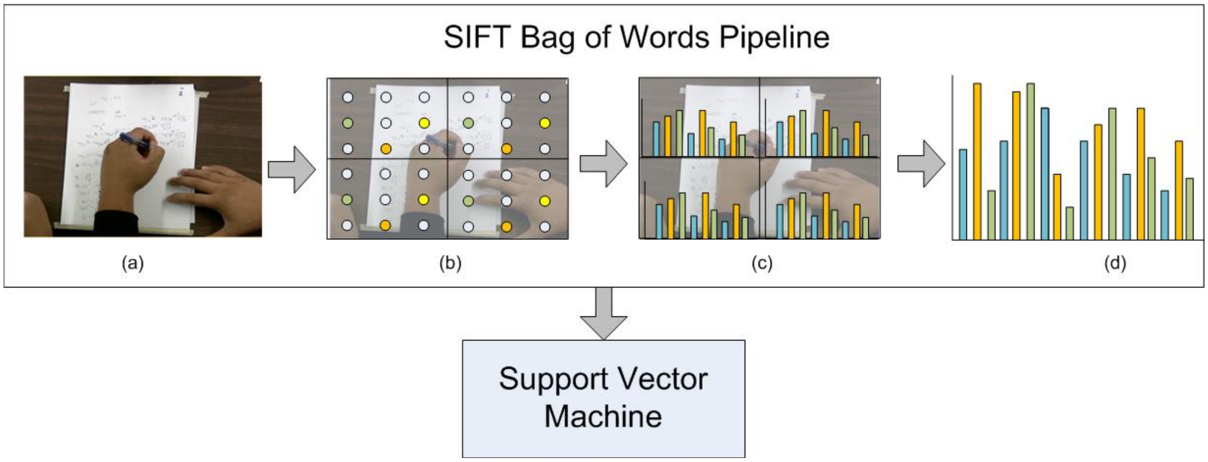


Figure 4. Bag of Words on SIFT with spatial pyramid matching for hand gesture recognition in one-to-one tutoring; (a) Image (b) Local descriptors after encoding (c) Pooling over spatial pyramids and (d) Final feature.

ments and is widely employed for action recognition, motion analysis and other related applications [1]. The MHI $H_\tau(x, y, t)$ is computed from an update function $\psi(x, y, t)$ in the following way:

$$H_\tau(x, y, t) = \begin{cases} \tau \text{ if } \psi(x, y, t) = 1 \\ max(0, H_\tau(x, y, t-1) - \delta) \text{ otherwise} \end{cases} \tag{1}$$

where $\psi(x, y, t)$ is function like background subtraction, frame differencing etc.

In this paper, we use MHI in the following two ways.

Using Background Image: In every video sequence, we consider the first frame $I(x, y, 1)$ of the sequence as the background image, which is then used to compute the binarized frame difference images between the background image and the remaining frames $I(x, y, t)$ in the following way:

$$D(x, y, t) = \begin{cases} 1 & \text{if } (I(x, y, 1) - I(x, y, t)) > \delta \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $2 \leq t \leq N_f$, N_f being the total number of frames in the sequence. The binarized mask $D(x, y, t)$ is used to select the differences in the image pixels between the background image and $I(x, y, t)$, which have changed beyond a threshold δ at time instance t, i.e. the difference image I_D is generated as

$$I_D(x, y, t) = D(x, y, t) \cdot (I(x, y, 1) - (I(x, y, t)) \tag{3}$$

where $\cdot$ denotes pixelwise multiplication. The modified MHI with background image is computed by summing all the difference images and then normalizing it by the maximum value of the summation image, i.e.

$$H^B(x, y) = \frac{1}{S^B_{max}} \sum_{i=2}^{N_f - 1} I_D(x, y, i) \tag{4}$$

where $S^B_{max} = max(\sum_{i=2}^{N_f-1} I_D(x, y, i))$. Fig. 5 shows sample modified MHIs obtained by applying the above equations on video segments containing deictic and non-deictic gestures. After generating H^B for every video sequence in the dataset, H^B is vectorized for training using linear SVM.

Without Using Background Image: In the second variant of the MHI, differences are computed between every two adjacent frames to generate the binarized difference images and subsequently the difference images. The equations (2) and (3) are now modified as the following:

$$D(x, y, t) = \begin{cases} 1 & \text{if } (I(x, y, t) - I(x, y, t-1)) > \delta \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$I_D(x, y, t) = D(x, y, t) \cdot (I(x, y, t) - (I(x, y, t)) \tag{6}$$

The above I_D is now used to determine H^B, which is vectorized and trained using SVM.

5. Evaluation

In this section, we evaluate the proposed techniques using the one-to-one tutoring hand gesture dataset that we introduced in Section 3. As discussed in Section 3, six videos in the entire dataset are being used in this paper, wherein each video corresponds to a different student subject. The tutor is the same in all the six different videos. The number of test cases corresponding to each subject and the type of gestures are listed in Table 1, wherein it was established that deictic and writing are the two main gestures that are involved in one-to-one tutoring. Therefore, we perform the evaluation of the proposed methods for these two gestures only, namely, deictic and writing. Additionally, we have two different classes for each type of gesture one for right

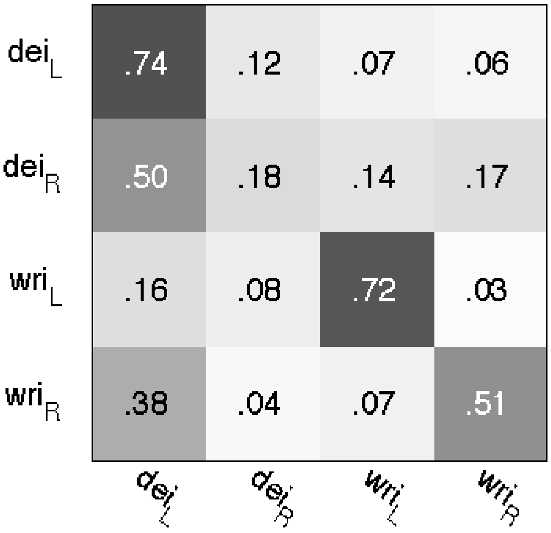


Figure 6. Confusion matrix for hand gesture recognition in one-to-one tutoring using BoW on SIFT features.

and one for left, i.e. there are a total of 4 classes: deictic left, deictic right, writing left and writing right. In these classes, left refers to the gesture being performed using left hand and similarly for the right gestures using right hand. The dietic left and writing left observations are contributed by the tutor and the right hand observations are contributed by the student. The total number of samples used for evaluation is 2166 for the four gesture classes considered.

In order to perform the evaluation, we perform a leave-one-out cross-validation (LOOCV) with respect to the subject. In other words, it is a 6-fold cross validation in which the cross validation is repeated six times such that in each iteration, the observations of one out of the six subjects are used for testing while the observations of the remaining 5 subjects are used for training.

Fig. 6 shows the confusion matrix for the four classes using the appearance based method that employs BoW on SIFT features (Section 4.1). It can be seen that this yields over 70% accuracy for deictic left and writing left classes, that mainly are associated with the tutor. It is noteworthy that that this method gives over 51% accuracy for writing right class, which has been cross-validated using LOOCV across six different student subjects, each with a different writing style. A higher accuracy for the two left classes can be attributed to the fact that the deictic left and writing left classes are associated with the same tutor. Also, an accuracy of only 18% is obtained for deictic right. One of the reasons attributing to this low accuracy is that the total number of samples for deictic right (corresponding to the deictic gestures of the student) is 173 out of 2166 samples (about 8%), and the pointing styles across the 6 different subjects can be quite varied. Fig. 7 and Fig. 8 show the confusion matrices for the motion-based approaches based on the two variants of MHI. It can be seen that the appearance based method using BoW on SIFT features gives a higher accuracy as compared to the motion based approach for all classes. It may be observed that the shape of a hand writing with a pen and a hand pointing with a pen are similar, so higher discrimination may be obtained by considering the temporal dynamics as well. However in the case of deictic left and right gestures, the shape as well as temporal dynamics of left and right hands pointing are both similar to one another. It is therefore understandable that deictic right gestures are often confused as deictic left gestures.

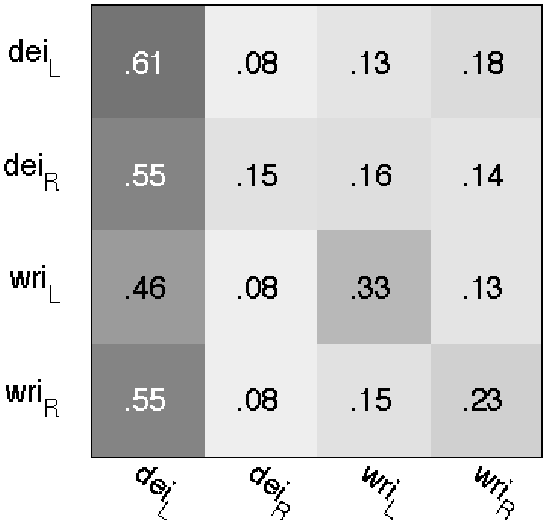
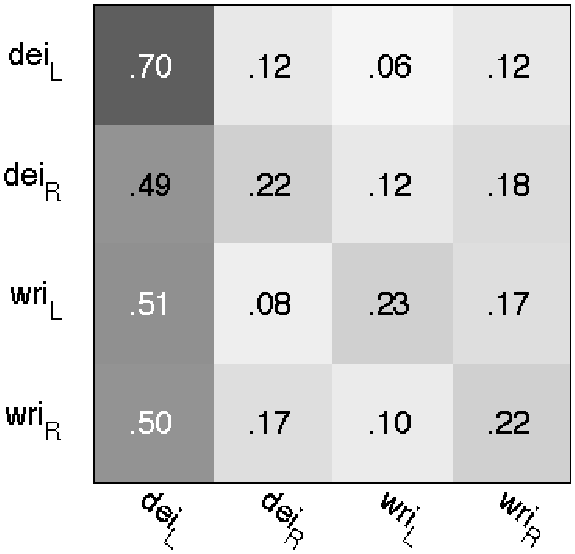


Figure 7. Confusion matrix for hand gesture recognition in one-to-one tutoring using modified motion history image *with* background image.

6. Discussion

In this section, the relevance of automatic gesture classification in the understanding social behavior in a one-to-one tutoring session is briefly discussed. Firstly, gestures constitute a major aspect in the non-verbal communication between the tutor and the student. The presence of, the frequency and distribution of the various gestures transpiring in a tutoring session can directly convey various aspects of the social communication. This coupled with the information from other modalities can be used to effectively decode subtle social cues. For example, if there is a lot of deictic gestural activity by the same tutor with one student compared to another, this could be possibly correlated with the degree of learning or the pace of learning of the students (which can be possibly inferred from the students track record, or from the pre and post test scores that have been recorded as part of the bigger dataset described in this paper).



	dei_L	dei_R	wri_L	wri_R
dei_L	.70	.12	.06	.12
dei_R	.49	.22	.12	.18
wri_L	.51	.08	.23	.17
wri_R	.50	.17	.10	.22

Figure 8. Confusion matrix for hand gesture recognition in one-to-one tutoring using modified motion history image *without* background.

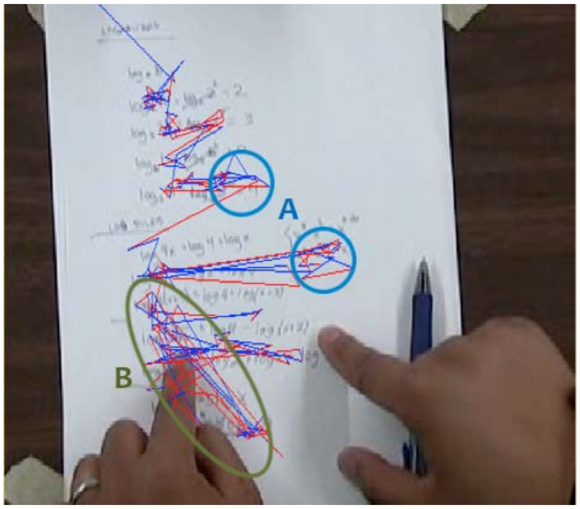


Figure 9. Deictic trace generated by capturing coordinates of the pointing finger of tutor during deictic gesture occurances; clusters of deictic activity are marked as A and B.

It would also be interesting to correlate the gestural activity spatially on the workspace to study emerging properties. One such example is illustrated in Fig. 9, that shows a trace of the deictic activity by the tutor for a stretch of approximately 25 minutes that the tutor spent on that sheet of paper. The trace in this figure was created by manually recording the coordinates of deictic gestures of the tutor, once for each occurance of the gesture, throughout the time when the workspace displayed that particicular page. This process can also be automated by using fingertip detection techniques once the deictic gestures have been identified temporally, although this is out of scope of this paper. It can be seen that there are clusters of high deictic activity that can be observed, which might potentially indicate complex problem segments or eleborate explanations that might be necessary to teach those problem segments.

When the hand gesture classification outputs are studied in combination with other modalities such as speech, other body gestures such as gaze, nods etc., this opens up many interesting work packages for future study.

7. Conclusions

In this paper, a new dataset called Tutor-Gesture is firstly introduced that contains a set of over 2300 video clips taken from six one-to-one tutoring sessions, involving a single tutor and six different students. The commonly used hand gestures, namely, deictic and writing, have been considered for automatic classification. Techniques based on shape properties using SIFT bag of words features, and based on motion, using a few variants of the Motion History Image (MHI) have been considered. It has been shown that SIFT bag of words yields a detection rate of over 70% for the left handed deictic and writing gestures corresponding to the tutor, and over 50% on right handed writing gestures, corresponding to six different writing styles of the students. Future work includes combining hand gestures from video data with other modalities such as speech, gaze, nods etc.

8. Acknowledgements

Support for this work was provided by NSF grant IIS 0905622 SOCS. This data was collected in San Diego Unified School District, at Reilly Lab, SDSU and the Machine Perception Lab at UCSD, and annotated by both groups. Thanks to Judy Reilly, Linda Salamanca, Jake Whitehill, Michael Salamanca, Karmen Dykstra and Amber Carini.

References

[1] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa. Motion history image: its variants and applications. *Mach. Vision Appl.*, 23(2):255–281, Mar. 2012.

[2] 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.

[3] A. F. Bobick, J. W. Davis, I. C. Society, and I. C. Society. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267, 2001.

[4] 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.

[5] 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.

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

[7] 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.

[8] B. F. J. Coordinated movement and rapport in teacher-student interactions. *Journal of Nonverbal behavior*, 12(2):120–138, 1988.

[9] M. Kaaniche and F. Bremond. Gesture recognition by learning local motion signatures. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2745–2752, 2010.

[10] 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.

[11] 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.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

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

[14] D. G. Lowe. Object recognition from local scale-invariant features. In *In Computer vision, 1999. The proceedings of the seventh IEEE international conference on, Volume 2*, page 11501157, 1999.

[15] D. McNeill. Gesture: A psycholinguistic approach.

[16] D. McNeill. Hand and mind. 1992.

[17] 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.

[18] 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.

[19] S. Padam Priyal and P. K. Bora. A robust static hand gesture recognition system using geometry based normalizations and krawtchouk moments. *Pattern Recogn.*, 46(8):2202–2219, Aug. 2013.

[20] 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.

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

[22] 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.

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

[24] 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.

[25] 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.

[26] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[27] A. Vedaldi and B. Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*, MM '10, pages 1469–1472, New York, NY, USA, 2010. ACM.

[28] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *IN: IEEE Conference on Computer Vision and Pattern Classification*, 2010.

[29] 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.

[30] 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.

[31] 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.