AI Enabled Tutor for Accessible Training

Ayan Banerjee(✉), Imane Lamrani, Sameena Hossain, Prajwal Paudyal, and Sandeep K. S. Gupta

Arizona State University, Tempe, AZ 85281, USA
{abanerj3,ilamrani,shossai5,ppaudyal,sandeep.gupta}@asu.edu

Abstract. A significant number of jobs require highly skilled labor which necessitate training on pre-requisite knowledge. Examples include jobs in military, technical field such computer science, large scale fulfillment centers such as Amazon. Moreover, making such jobs accessible to the disabled population requires even more pre-requisite training such as knowledge of sign language. An artificial intelligent (AI) agent can potentially act as a tutor for such pre-requisite training. This will not only reduce resource requirements for such training but also decrease the time taken for making personnel job ready. In this paper, we develop an AI tutor that can teach users gestures that are required on the field as a pre-requisite. The AI tutor uses a model learning technique that learns the gestures performed by experts. It then uses a model comparison technique to compare a learner with the expert gesture and provides feedback for the learner to improve.

Keywords: AI enabled tutor · ASL · Explainable AI

1 Introduction

Advances in machine learning, artificial intelligence (AI) and embedded computing is bringing a revolution in human computer communication, where humans and computers will operate in symbiosis for collaborative outcomes and cooperative learning. The applications with collaboration can span over robot assisted military combat [12,22,33], and collaboratory rehabilitation for diseases such as Parkinson's or Alzheimer's [1,30]. Cooperative learning applications include computer aided training of military personnel [23], heavy equipment operators [15], or performance coaching in entertainment applications [29] or tutoring American Sign Language (ASL) for ease of communication between humans with various disability profiles such as deaf or hard-of-hearing [7]. In most of these applications gestures form an important component of communication between the human and the computer or another human. A gesture is composed of multiple components arranged in temporal order with specific transitions from one component to the other. There are typically two components: a) gesture recognition, by a machine or a human, and b) replication by an audience (machine/human). If the audience is a machine, the recognized gesture may not be in any

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12163, pp. 29–42, 2020.
https://doi.org/10.1007/978-3-030-52237-7_3

understandable form since the machine can be programmed to replicate by using the original sensor measurements. But if the audience is a human then gestures need to not only be recognized but understood in more fundamental ways to achieve desired learning outcomes.

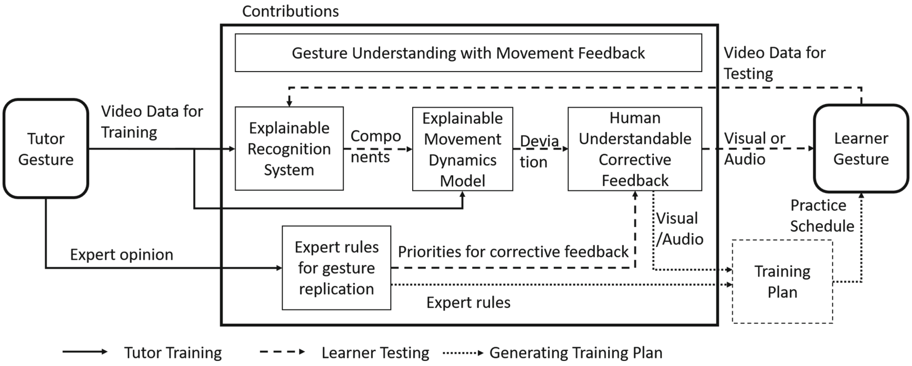


Fig. 1. Co-operative learning application model for understanding and replicating gestures similar to a tutor.

In this work, we consider a co-operative gesture learning application model that not only recognizes errors in a learner but also provides corrective feedback that enables the learner to replicate a gesture with similar qualities of a tutor (Fig. 1). The application model takes multiple iterations of a given gesture from several tutors. It should model not only the individual gesture components potentially using a data-driven machine learning architecture but also the transition from one component to the other. In addition, the tutors will also provide "expert rules" that are essential for expressing the correct or nuanced meaning of a gesture and can be used to guide corrective feedback to the learner. In the testing phase, a learner provides sensor data for a replication of the gesture, which is passed to a recognition system. It results in recognition of the gesture components along with an explanation for correctness. The inter-component movement will be checked against a tutor. The results from the component and movement recognition system will be combined with expert rules to create a prioritized set of corrective feedback for the learner, which will be disseminated through audio-visual means. In an extension of this system, it can also be envisioned that the system generates a personalized training plan for the learner over time. The training plan is considered as an extension for the research.

In this paper, we consider ASL learning example to demonstrate our contributions. ASL signs are poly-componential in nature and is a sophisticated gesture based language [2]. Hence, lessons learned from this example can potentially be applicable to other gesture based communication domains such as ensuring compliance to Center for Disease Control (CDC) guidelines for hand-washing [3].

The ASL tutor is intended to be used in computer science accessible virtual education (CSAVE) architecture for Deaf and Hard of Hearing (DHH) individuals [14]. An IMPACT Lab project, CSAVE architecture facilitates personalized learning environment for deaf and hard of hearing students. It enables DHH students to collaborate with the instructor, interpreter, and their hearing peers seamlessly without them having to reveal their disability. Many of these technical courses require students to work in groups to collaborate on projects. Incorporating ASLTutor within the CSAVE architecture can enable the hearing students with the tool they would need to communicate with their DHH peers.

2 Existing Work and Challenges

To understand the unique challenges associated in answering the abovementioned question, let us contrast two examples: a) an AI tutor for training a person in a foreign spoken language, and b) an AI tutor for training a person in ASL.

Existing Work: For second spoken language learners, many research works point out to the positive relationship between feedback given through interaction and the learning performance [19,20]. The ability to practice and receive feedback is also a positive aspect of immersive environments for second language learning such as study abroad programs and even classroom environment to some extent [21]. Many software applications for spoken languages incorporate some form of feedback to help improve the pronunciation of learners [36]. Applications like DuoLingo also provide interactive chat-bot like environments with feedback to increase immersion [40]. However, such applications are not available for learners of sign languages. This is in part due to the inherent technical difficulties for providing feedback to sign language learners.

2.1 Challenge 1: Explainable Systems

A simple notion of the correctness of a sign execution can be computed using existing sign language recognition systems [5,6,8,9,16,17,32,34,41]. However, for providing more fine-grained feedback, more details are desirable. This is specially so because sign languages, unlike spoken languages, are multi-modal. Thus, if an error is present in execution, feedback should be given that ties back to the erroneous articulator(s). For instance, if a student executes the movement part of a sign correctly, and performs the sign in the right position relative to her body, but she fails to articulate the right shape of the hand, then feedback should be given regarding the incorrect handshape. Thus, blackbox recognition systems are not very useful for feedback and explainable systems that can recognize conceptual elements of the language must be developed.

2.2 Challenge 2: Determination of Appropriate Feedback

Feedback mechanisms for spoken and sign language differ significantly. The differences arise primarily due to the articulators used for speech versus those used for signing. Apart from some research for feedback in rehabilitation for physical therapy, which is conceptually very dissimilar to sign language learning, there are no existing systems in this domain [42]. Thus, the types of feedback to be given to learners must be determined by referring to the linguistics of sign languages, close work with ASL instructors and referring to academic studies. Codifying and automating the suggested feedback into a usable system is a challenging process and a worthy research undertaking.

2.3 Challenge 3: Extension to Unseen Vocabulary

Sign language recognition differs from speech recognition in one crucial aspect: the number of articulatory channels. This is partially an artifact of the medium used for recognition, i.e. audio vs video. Audio is usually represented as two-dimensional signals in amplitude and time, while colored videos are four-dimensional signals: three spatial dimensions, one channel dimension for color and one time dimension. The consequence of this for speech to text systems for spoken language learning such as Rosetta Stone [36] offers some feedback to a learner based on comparisons between their utterances and those of a native speaker. This one-to-one comparison to a gold standard is a desirable way for learning systems where the learner is attempting to get close in performance to a tutor. Such comparison for gesture learning becomes multi-dimensional spatio-temporal problem and hence is more challenging. Moreover, a tutoring system needs to readily extend to new vocabulary as the learner progresses. To extend the capability of a recognition system that is based on a classifier, the entire system will need to be retrained to account for new signs.

2.4 Challenge 4: Ubiquitous Recognition

The growing usage of self-paced learning solutions can be attributed to the effect of the economy of scale as well as to their flexibility in schedule. To achieve these desired advantages, the barrier to access must be reduced as much as possible. This implies that requiring the usage of specialized sensors such as 3-D cameras will hinder the utility. Thus, a proposed solution that can truly scale and have the maximum impact as a learning tool must be accessible without the need to purchase special sensors or to attend in special environments. The sensing device that is most accessible to any user is the smartphone. This is challenging because there is a huge variance in the type, quality, and feed of smartphone-based cameras and webcams. Furthermore, assumptions on adequate lighting conditions, orientations, camera facing directions and other specific configurations cannot be made, and have to either be verified by quality control or accounted for by the recognition and feedback algorithms.

In this paper, we use concept level learning for gesture understanding that can enable: a) extendable recognition, b) corrective explainable feedback to human learners, c) configurable feedback incorporation based on expert rules, and d) ubiquitous operation on any smartphone.

3 AI Tutor Design Goals

In this section, we discuss the design goals and principles for an AI Tutor and show proof-of-concept studies on an ASL tutor.

3.1 Embedding Movement in Models

Hybrid systems encode transient behavior using a set of differential equations that can potentially be used to represent the kinematics of the gesture. For example, the transient behavior of the movement from one hand shape to other is captured from high definition video and then utilizing Posenet to estimate wrist positions [24]. A kinematic model obtained from expert knowledge of human hand movements [35] can express the transient dynamics of movement in between hand shapes. The recognition result of the explainable machine learning system can then be considered as discrete states while the learned kinematic model can be considered as the dynamical component of the hybrid system representation of the gesture. State transitions can be expressed through temporal constraints.

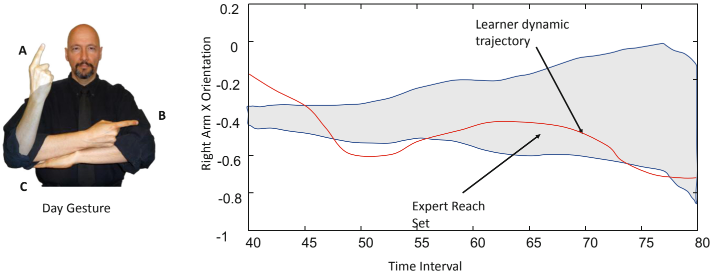


Fig. 2. Day example, the evolution of reach set over time for a tutor, and the execution for a learner.

Proof-of-Concept Example

We consider the gesture for "Day" in ASL. The Day gesture is shown in Fig. 2, it involves two hand shapes: a) the left hand pointing towards the right, and b) the right hand pointing towards the head. Then it has one transient hand movement, where the right arm while pointing pivots on the right elbow and makes a quarter circle and lands on the left elbow.

We generate the hybrid system for Day gesture as shown in Fig. 3. We consider three different components or states of the "Day" gesture: a) Pointing to the head (State A), b) movement from head to the top of the left arm (State B), and c) movement from top of the left arm to the elbow (State C). While transiting from one state to the other, we consider that the center point of the palm of both the left and right arm move following the model described in Eq. 1.

$$\frac{d\overrightarrow{p}}{dt} = \overrightarrow{v}, \frac{d\overrightarrow{v}}{dt} = \overrightarrow{a}, \frac{d\overrightarrow{a}}{dt} = x_1\overrightarrow{a} + x_2\overrightarrow{v} + x_3\overrightarrow{p} + x_4, \tag{1}$$

where $\overrightarrow{p}$ is the position vector for the right arm, $\overrightarrow{v}$ is the velocity vector and $\overrightarrow{a}$ is the acceleration vector and x_is are parameters of the hand motion. This is an overly simplistic model of the palm movement but is used to generate useful feedback relating to arm acceleration.

3.2 Ubiquitous Recognition of Movement

The need for recognition of concepts from data collected using heterogenous sensors prohibits the usage of traditional machine learning systems, which are affected by camera resolution, lighting condition, as well as distance from the lens. Although Convolutional Neural Networks (CNN) or other deep learning systems can perform object recognition under noisy conditions, concepts in a gesture video include much finer details such as handshapes, fine grained location information, and movements, which may not be recognized effectively by a deep learning system [25–27]. Moreover, the amount of available training data for gesture recognition is far less than what needed for reliable performance of deep learning classification systems avoiding the risk of over-fitting [26, 27].

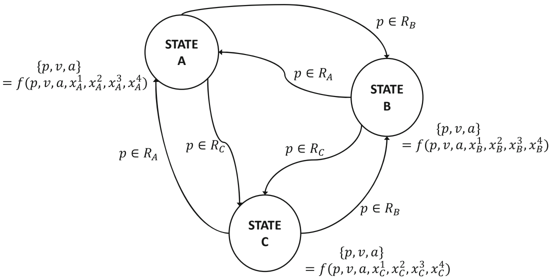

STATE A
STATE B
STATE C
$p \in R_B$
$p \in R_A$
$\{p, v, a\} = f(p, v, a, x_A^1, x_A^2, x_A^3, x_A^4)$
$\{p, v, a\} = f(p, v, a, x_B^1, x_B^2, x_B^3, x_B^4)$
$p \in R_A$
$p \in R_C$
$p \in R_C$
$p \in R_B$
$\{p, v, a\} = f(p, v, a, x_C^1, x_C^2, x_C^3, x_C^4)$

Fig. 3. HA representation of the Day gesture.

We take a different approach through pose estimation [31, 38, 39]. Our approach is to convert the gesture execution into spatio-temporal evolution of continuous variables. The recognition is a composite outcome of simpler similarity-based comparisons. This can potentially contribute to the robustness to changing environmental conditions since the pose estimation step already eliminates background and only focuses on points of interest.

We considered the “X” and “Y” co-ordinate time series of the right and left wrist normalized with respect to torso height and hip width. The location concept was extracted using six location buckets around the face and the chest of a user. This is because as a concept only the proximity to different body parts are important and not the exact pixel level location.

To extract handshape we utilized the wristpoints to crop the palm of the user. We then used the CNN Inception model trained using the ImageNet dataset and retrained using fingerspelling handshapes [37]. The retrained inception model was not used to classify handshapes but instead was used to compare two handshapes: one from the tutorial video and the other from the test user. Only the outputs of the penultimate layer of the Inception model for both the tutorial and the user was compared using the euclidean distance metric. This not only enables concept matching but also provides extensibility, because to compare with a new tutorial sign no training is required.

We explored two different methods of movement concept recognition: a) direct comparison using segmental dynamic time warping strategy [28], and b) comparison with respect to kinematic model parameters [4]. The first strategy is model agnostic and only gives feedback about the correctness of the movement concept. The second approach utilizes a hybrid dynamical system to model gesture concepts. This model driven approach can provide more granular feedback as discussed in our initial work [4].

We evaluated the concept learning methodology on 100 first time learners of ASL users each of them learned 25 ASL gestures and performed three times each gesture. The videos of the ASL gestures were taken using their own smartphones at home. The system has an overall test accuracy of 87.9% on real-world data [28]. We also evaluated our hybrid dynamical model on 60 first time learners of ASL users each of them learned 20 ASL gestures and performed three times each gesture. Results show that kinematic parameters in Eq. 1 can represent each gesture with precision of 83%, and recall of 80%.

3.3 Movement Matching Between Tutor and Learner

In our approach, the hybrid system based representation of a gesture is instantiated for a tutor. The instantiation procedure involved collecting data using wearable and video based sensors from a tutor and running the following hybrid system mining technique.

Hybrid Mining Technique: The input to the model mining methodology are the input output traces, which may contain timed events, and discrete or continuous inputs.

A) First step is I/O segmentation. The discrete mode changes of the hybrid model is triggered by three main causes: a) user generated external events that are accompanied by time stamps and input configurations, b) system generated timed events, and c) events generated due to threshold crossing of observable parameters of the physical system.

B) Second step is to cluster modes in accordance with their triggering mechanism. This clustering step is required to minimize the redundancy in the number of discrete modes of the mined specification.

C) The third step is mining the kinematic equations. Each trace is passed to a Multi-variate Polynomial Regression to obtain the kinematic equations. For the linear case, we utilize Fischer information and Cramer Rao bound to compute the linear coefficients [18]. The output is the flow equation parameters for each trace between modes. A result of the flow equation extraction mechanism is that different traces may have the same flow equation. The corresponding modes are then clustered together using density based approaches on the flow parameters and assigned the same mode labels.

D) The fourth step is guard mining. We derive the guard conditions for each cluster, where each cluster represents a distinct control switch. If the guard condition is not a constant value of actuation and is varying within each data point in the cluster, we employ Fisher information and Cramer Rao bound to derive the linear relation of the input, output, and internal parameters [18]. The Guard conditions are then used to further refine the mode clustering. The output is a Hybrid automata inferred from the input, output, and internal parameters with modes, flow equations, and guards.

Tutor and Learner Comparison: The natural variation of a tutor is modeled by computing a reach set of the learned hybrid system. The reach set is the set of all continuous states that is observed from simulating the hybrid system over time for a bounded set of initial conditions, which may represent natural variations in the tutor's execution of the gesture.

Given an execution of the gesture by a learner, the video based hand gesture recognition system provides us with executed hand shapes, the times of transition from one shape to the other, and an identification of wrong executions by the learner. The reach set comparison can provide the deviation from a tutor. For instance if the fingertip data is encompassed by the reach set then, it is tutor level. However, if it is outside the reach set at any point in time, then it the learner has differences with the tutor. The time segments where the learner differed from the tutor can then be passed to a dynamical system mining technique that is programmed with the kinematic model of the human arm. The mining technique will provide a new set of parameters for the learner.

Proof-of-Concept: We collected Kinect data including video and bone movement data from 60 subjects for 20 ASL gestures including "Day". We chose one user, who is a lecturer at ASU on sign language and considered the person as a tutor. We collected data for 20 executions of "Day" and computed the variations in initial positions and angles, speeds, and the parameters of Eq. 1. The sensors used were Kinect video and bone data. In addition the tutors wore an armband that collected accelerometer, orientation, gyroscope and Electromyogram data. The traces were used to derive the parameters of the kinematics described in Eq. 1.

We then derived different initial conditions by performing a statistical analysis of the tutor's speed, initial positions and parameters for Eq. 1. These were used to perform the reachability analysis of the hybrid system using the SpaceEx tool [10]. Figure 2 shows the X orientation reach set evolution of the hybrid system. All the different executions of the tutor are inside the gray area. The reach set is an over approximation because exact computing is intractable.

We then considered another subject's execution of the "Day" gesture, where the learner ended the Day gesture with the right palm near to the left. The X orientation of the right palm of the learner is shown in red in Fig. 2. It clearly shows that the learner violates the reach set and hence is not classified as similar to a tutor, although all the hand signs are correctly executed by the learner. However, the learner executes the same sequence of hand shapes. Hence, a knowledge based feedback system will consider this execution as correct. But the execution has clear differences with the tutor in the transition between gestures in the transition from state B to C.

The dynamics of the learner's execution between state B and C is then used to regenerate the parameters of Eq. 1. The learner is seen to have 40% elevated x_3. This means that as the position of the right arm goes closer to the left arm, the acceleration increases resulting in overshooting of the right arm beyond the left arm position. Hence the feedback that is generated for the learner is to control the learner's right arm so that the velocity is uniform. By practicing one can get the right arm velocity uniform and be on par with a tutor.

3.4 Explainable Feedback

A variety of feedback could be constructed using the information available from the results of the location, movement, and handshape modules. In addition to separate feedback for each of the hands, feedback could also be presented in forms of annotated images or by using animations. For location feedback, the correct and the incorrect locations for each of the hands could be highlighted in different colors. For the handshape feedback, the image of the hand that resulted in the highest difference in similarity could be presented. Each of these types of possible feedback is derived from the information available. However, they should be individually tested for usability and care should be taken not to cognitively overload the learner with too much feedback at once.

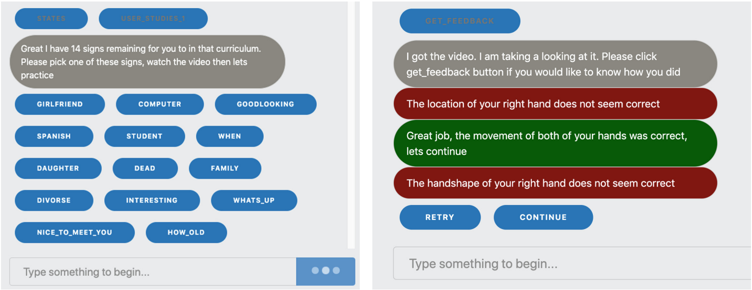
More granular feedback can be provided using kinematic models if each component of the model has direct correlation with a physical manifestation of the human arm. Such correlations and the parameters estimated for the learner can be used to generate understandable feedback that enables the learner to perfect gesture execution. Such feedback will be guided by the expert rules specified by the tutor. Complex models tend to be less amenable towards feedback generation. Hence our goal will be to best exploit the trade-off between model complexity and explainability.

4 Prototype and Evaluation of Learning Outcomes

We first discuss our experimental setup and then evaluation results.

4.1 Prototype

A chat bot enabled web based gesture learning interface is developed (Fig. 4). In this chatbot, the learner chooses a sign and learns the gesture. Then the learner chooses to practice when the video of the learner executing is recorded and compared with the expert. Textual feedback is then provided to the learner to improve gesture execution capability.



Great I have 14 signs remaining for you to in that curriculum. Please pick one of these signs, watch the video then lets practice

I got the video. I am taking a looking at it. Please click get_feedback button if you would like to know how you did

The location of your right hand does not seem correct

Great job, the movement of both of your hands was correct, lets continue

The handshape of your right hand does not seem correct

Fig. 4. Interactive chat-bot interface. Right: The movement of both the hands were correct (green), but the location and right hand handshape were not correct. (Color figure online)

4.2 Learning Outcomes

The purpose of assessment tests are to evaluate learning outcomes. Two types of tests are considered: a) retention tests and b) execution tests. We recruited 15 subjects who were tested on 50 ASL signs of their choice from a pool of ASL signs for the states of USA.

Each learner is given randomly selected signs for retention tests. The learner either chose to practice a given sign multiple times or move on. For each test, the learner is shown a video and is asked to choose among 4 options for the correct one. Thus, the baseline performance for random guessing would be 25%. The performance of the learner with and without feedback is used as a metric for feedback effectiveness.

For execution tests each of the learners is given randomly selected signs to execute. During the test the learner is given a sign and asked to begin recording its execution. The execution tests is manually scored offline by the research team. If the learner had any two of location, movement or handshape correct on both hands, then she receives a score of 0.5 for that sign. If all three were correct, she receives 1. Otherwise, she receives 0. The performance on execution tests with and without feedback is considered to evaluate effectiveness of feedback.

Our results show that retention of the signs did not improve with feedback. In fact retention was already upwards of 80% with or without feedback. However, there was significant improvement in execution accuracy. It improved from 63% without feedback to 85% with feedback. This indicates that overall feedback has a significant effect on learning outcome. The effectiveness of different types of feedback however could not be evaluated given the less number of participants. However, an end of study survey showed that majority of the participants preferred fine grained feedback.

5 Conclusions and Discussions

Feedback in gesture based learning is of utmost importance as evidences in our evaluation results. An AI tutor hence not only has to disseminate knowledge and evaluate students, but also provide feedback to ensure learning. In this paper, we have demonstrated through a proof-of-concept study of an AI tutor of ASL, that AI tutor has to be explainable, ubiquitous and extensible. The concepts learned in this project can be employed in other gesture based training applications such as military, physiotherapy, medical surgery training. Through proof-of-concept implementations we have shown the importance of feedback in AI tutor, however, there are significant hurdles before it can be realized in practice.

Usability: The system requires no extra sensor, just a mobile phone camera is enough. The system could achieve this operation, because of the modular representation and identification of gestures in terms of their components.

Extensibility: The system only compares a test gesture to one expert video and does not need training for new gesture classes. Hence it is extensible with only the inclusion of an expert gesture video.

Difference in Feedback Generation Methods: Generation of explanation heavily depends on the model. Complex models may be more accurate but not be explainable. Dynamical models of the human arm of different complexity can be broadly classified into the following categories:

a) Differential equation models derived from kinematics of human fingers and arms: These models are typically derived from Magnetic Resonance Imaging (MRI) [35] or CT [13] scans of the human hand and can go to the level of minute finger movements. In these methods a kinematic model is developed from a general understanding of human hand and the parameters are estimated from the imaging data. Authors in [35] use a parameterized models such that each parameter has a direct visual manifestation. A deviation in a parameter hence can be easily converted into explanations considering the visual signatures. A big problem is that the model is the dimensionality, and learning the appropriate parameters from MRI images is computationally expensive.
b) Data-driven models derived using data glove or frictional sensors: Such models typically utilize predictors such as Kalman filters [11]. The model parameters have no direct relation to any understandable component of the human hand. But the overall model can be used to predict hand motion given a configuration of the parameters. Results from these models are difficult to explain.

Constraints on Generation of Feedback: Another significant hurdle is the feasibility of using the feedback for a person given their unique constraints. A difference in model parameters between the learner and the tutor is intended to be used to generate correctional feedback. However, the low dimensional dynamical model is not accurate for larger time horizons. This means that there can be cases where the model may generate inviable feedback. Such as requesting extremely large acceleration or bending the arm at infeasible angles. Hence, every feedback has to be validated against a set of constraints that express viable feedback. Moreover in case feedback is invalid, we have to modify the model such that it can generate a feasible feedback.

One of the important future work is to apply this AI tutor for training DHH students, gestures related to technical concepts of computer science so that they can then take CS courses in the future and have a career in the technical field. CS courses have several technical terms which do not have gestures for them. Utilizing AI tutor to not only teach but organically generate signs for these technical gestures is one of our future goals.

References

1. Alwardat, M., et al.: Effectiveness of robot-assisted gait training on motor impairments in people with Parkinson's disease: a systematic review and meta-analysis. Int. J. Rehabil. Res. **41**(4), 287–296 (2018)
2. Anthimopoulos, M., Dehais, J., Diem, P., Mougiakakou, S.: Segmentation and recognition of multi-food meal images for carbohydrate counting. In: 13th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 1–4. IEEE (2013)
3. Banerjee, A., Amperyani, V.S.A., Gupta, S.K.: Hand hygiene compliance checking system with explainable feedback. In: 18th ACM International Conference on Mobile Systems Applications and Services, WearSys Workshop (2020)
4. Banerjee, A., Lamrani, I., Paudyal, P., Gupta, S.K.S.: Generation of movement explanations for testing gesture based co-operative learning applications. In: IEEE International Conference on Artificial Intelligence Testing, AITest 2019, Newark, CA, USA, 4–9 April 2019, pp. 9–16 (2019). https://doi.org/10.1109/AITest.2019.00-15
5. Camgöz, N.C., Kındıroğlu, A.A., Karabüklü, S., Kelepir, M., Özsoy, A.S., Akarun, L.: BosphorusSign: a Turkish sign language recognition corpus in health and finance domains. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016, pp. 1383–1388 (2016)
6. Chai, X., et al.: Sign language recognition and translation with Kinect. In: IEEE Conference on AFGR, vol. 655, p. 4 (2013)
7. Chen, T.L., et al.: Older adults' acceptance of a robot for partner dance-based exercise. PloS One **12**(10), e0182736 (2017)
8. Cooper, H., Bowden, R.: Learning signs from subtitles: a weakly supervised approach to sign language recognition. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2568–2574. IEEE (2009)
9. Forster, J., Oberdörfer, C., Koller, O., Ney, H.: Modality combination techniques for continuous sign language recognition. In: Sanches, J.M., Micó, L., Cardoso, J.S. (eds.) IbPRIA 2013. LNCS, vol. 7887, pp. 89–99. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38628-2_10
10. Frehse, G., Kateja, R., Le Guernic, C.: Flowpipe approximation and clustering in space-time. In: Proceedings of the Hybrid Systems: Computation and Control, HSCC 2013, pp. 203–212. ACM (2013)
11. Fu, Q., Santello, M.: Tracking whole hand kinematics using extended Kalman filter. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4606–4609. IEEE (2010)
12. Galliott, J.: Military Robots: Mapping the Moral Landscape. Routledge, Abingdon (2016)
13. Harih, G., Tada, M.: Development of a finite element digital human hand model. In: 7th International Conference on 3D Body Scanning Technologies (2016)

14. Hossain, S., Banerjee, A., Gupta, S.K.S.: Personalized technical learning assistance for deaf and hard of hearing students. In: Thirty Fourth AAAI Conference, AI4EDU Workshop (2020)
15. Jiang, Q., Liu, M., Wang, X., Ge, M., Lin, L.: Human motion segmentation and recognition using machine vision for mechanical assembly operation. SpringerPlus **5**(1), 1–18 (2016). https://doi.org/10.1186/s40064-016-3279-x
16. Koller, O., Zargaran, S., Ney, H., Bowden, R.: Deep sign: enabling robust statistical continuous sign language recognition via hybrid CNN-HMMS. Int. J. Comput. Vis. **126**(12), 1311–1325 (2018). https://doi.org/10.1007/s11263-018-1121-3
17. Kumar, S.S., Wangyal, T., Saboo, V., Srinath, R.: Time series neural networks for real time sign language translation. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 243–248. IEEE (2018)
18. Lamrani, I., Banerjee, A., Gupta, S.K.: HyMn: mining linear hybrid automata from input output traces of cyber-physical systems. In: IEEE Industrial Cyber-Physical Systems (ICPS), pp. 264–269. IEEE (2018)
19. Lightbown, P.M., Spada, N.: Focus-on-form and corrective feedback in communicative language teaching: effects on second language learning. Stud. Second Lang. Acquisit. **12**(4), 429–448 (1990)
20. Mackey, A.: Feedback, noticing and instructed second language learning. Appl. Linguist. **27**(3), 405–430 (2006)
21. Magnan, S.S., Back, M.: Social interaction and linguistic gain during study abroad. Foreign Lang. Ann. **40**(1), 43–61 (2007)
22. Min, H., Morales, D.R., Orgill, D., Smink, D.S., Yule, S.: Systematic review of coaching to enhance surgeons' operative performance. Surgery **158**(5), 1168–1191 (2015)
23. Noble, D.D.: The Classroom Arsenal: Military Research, Information Technology and Public Education. Routledge, Abingdon (2017)
24. Papandreou, G., et al.: Towards accurate multi-person pose estimation in the wild. In: CVPR, vol. 3, p. 6 (2017)
25. Paudyal, P., Banerjee, A., Gupta, S.K.: SCEPTRE: a pervasive, non-invasive, and programmable gesture recognition technology. In: Proceedings of the 21st International Conference on Intelligent User Interfaces, pp. 282–293. ACM (2016)
26. Paudyal, P., Lee, J., Banerjee, A., Gupta, S.K.: DyFAV: dynamic feature selection and voting for real-time recognition of fingerspelled alphabet using wearables. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces, pp. 457–467. ACM (2017)
27. Paudyal, P., Lee, J., Banerjee, A., Gupta, S.K.: A comparison of techniques for sign language alphabet recognition using arm-band wearables. ACM Trans. Interact. Intell. Syst. (TiiS) (2018, accepted)
28. Paudyal, P., Lee, J., Kamzin, A., Soudki, M., Banerjee, A., Gupta, S.K.: Learn2Sign: explainable AI for sign language learning. In: Proceedings of the 24nd International Conference on Intelligent User Interfaces, pp. 457–467. ACM (2019)
29. Riley, M., Ude, A., Atkeson, C., Cheng, G.: Coaching: an approach to efficiently and intuitively create humanoid robot behaviors. In: 2006 6th IEEE-RAS International Conference on Humanoid Robots, pp. 567–574. IEEE (2006)
30. Salichs, M.A., Encinar, I.P., Salichs, E., Castro-González, Á., Malfaz, M.: Study of scenarios and technical requirements of a social assistive robot for Alzheimer's disease patients and their caregivers. Int. J. Soc. Robot. **8**(1), 85–102 (2016). https://doi.org/10.1007/s12369-015-0319-6

31. Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A.: 3D human pose estimation: a review of the literature and analysis of covariates. Comput. Vis. Image Underst. **152**, 1–20 (2016)
32. Schmidt, C., Koller, O., Ney, H., Hoyoux, T., Piater, J.: Using viseme recognition to improve a sign language translation system. In: International Workshop on Spoken Language Translation, pp. 197–203 (2013)
33. Sharkey, N.E.: The evitability of autonomous robot warfare. Int. Rev. Red Cross **94**(886), 787–799 (2012)
34. Starner, T., Pentland, A.: Real-time American sign language visual recognition from video using hidden Markov models. Master's Thesis, MIT Program in Media Arts (1995)
35. Stillfried, G., Hillenbrand, U., Settles, M., van der Smagt, P.: MRI-based skeletal hand movement model. In: Balasubramanian, R., Santos, V.J. (eds.) The Human Hand as an Inspiration for Robot Hand Development. STAR, vol. 95, pp. 49–75. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-03017-3_3
36. Stone, R.: Talking back required (2016). https://www.rosettastone.com/speech-recognition. Accessed 28 Sept 2018
37. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
38. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: convolutional 3D pose estimation from a single image. In: CVPR 2017 Proceedings, pp. 2500–2509 (2017)
39. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems, pp. 1799–1807 (2014)
40. Vesselinov, R., Grego, J.: Duolingo effectiveness study, vol. 28. City University of New York, USA (2012)
41. Zhang, Q., Wang, D., Zhao, R., Yu, Y.: MyoSign: enabling end-to-end sign language recognition with wearables. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 650–660. ACM (2019)
42. Zhao, W.: On automatic assessment of rehabilitation exercises with realtime feedback. In: 2016 IEEE International Conference on Electro Information Technology (EIT), pp. 0376–0381. IEEE (2016)