An AI-enhanced Pattern Recognition Approach to Analyze Children's Embodied Interactions

Ceren OCAK[a*], Theodore J. KOPCHA[b] & Raunak DEY[c]
[a]*College of Education, Michigan State University, USA*
[b]*College of Education, University of Georgia, USA*
[c]*College of Arts and Sciences, University of Georgia, USA*
*ocakcere@msu.edu

Abstract: This study first presents an approach to the study of computational thinking (CT) as an embodied phenomenon that relies on the creation and analysis of multimodal transcripts. The approach, which incorporates a social semiotic approach to multimodality, is then used to train an artificial intelligence (AI) to recognize patterns in the participant's behaviors that reflect their embodiment of CT during an educational robotics activity. The AI was developed to ease the labor-intensive aspects of creating and analyzing a multimodal transcript. The findings suggested that the AI-enhanced pattern recognition approach identified similar clusters of activity as human analysis, adding a level of confidence to the analysis of children's CT that would be difficult to achieve using human analysis.

Keywords: Computational thinking (CT), social semiotic approach to multimodality, embodied interactions, artificial intelligence (AI)

1. Introduction

Recent perspectives of embodied cognitive science offer new methodological prospects for exploring children's CT, where CT is studied as a process rather than a product of learning. Although some scholars have begun studying CT from an embodied perspective (Black et al., 2012; Chung & Hsiao, 2019; Melcer, 2017), attempts to conceptualize CT from an embodied perspective have not translated into researchers' methodological preferences. The research designs associated with CT have often been reductive, ignoring the chaotic, self-organizing aspects of the process. This study aims to explore how different methodological possibilities can support the analysis of CT from an embodied perspective. To that end, we first present an approach to multimodal transcription using a methodological framework that aligns with the study of CT as an embodied phenomenon. The framework incorporating a social semiotic approach to multimodality will then be applied to the artificial intelligence (AI) pattern recognition approach. This research contributes meaningfully to recent and ongoing questions about how embodied perspectives can be leveraged into a research methodology in CT. To answer this question, we ask:

1. How can we create an AI-enhanced pattern recognition approach to study children's CT through their embodied interactions?

One area in which AI has the potential to improve research and analysis is involving multimodal data (Andrade et al., 2016; Sharma et al., 2019). When combined with machine learning (ML), the analysis of multimodal data can bring additional insight into different aspects of students' learning (Blikstein, 2013). In addition, multimodal learning analytics has been useful to predict learning performance (Giannakos et al., 2019; Junokas et al. 2018), model and assess student behavior (Blikstein, 2011), and enhance the performance of intelligent tutoring systems (Yang et al., 2021).

One reason for AI's success in education design research is when its design strongly embodies the tenets of learning sciences (Järvelä et al., 2020; Luckin & Cukurova, 2019). This highlights the significance of grounding AI-design in a strong methodological framework. In this study, we offer a *supervised deep neural network framework*, which takes as input a set of images and maps them to labels generated via domain experts. The goal is to classify embodied interactions in video data in which multiple modes of communication take place simultaneously (e.g., talking while gesturing;

moving one's body while using objects in the environment). Our paper contributes to the existing literature in the following ways: (1) employs an ML algorithm, which takes on human analysis to understand the participants' meaning-making that is strongly connected to extended cognition (e.g., coupling taking place among the gestures, discourse), (2) details the steps to create a training dataset for a supervised approach, (3) provides a method of studying social aspects of behavior in large-scale data, and (4) helps multimodal learning analytics realize its potential in the field of learning sciences.

2. Methodological Framework

For the current study, we drew on our previous research (Kopcha et al., 2019; Kopcha et al., 2020) to train an AI to identify children's embodied interactions: hand gestures that represented numbers, bodily movements that imitated the robot, use of the computer to program the robot, use of hand-made notes, and use of the robot itself. The data we used to train the algorithm was largely derived from the multimodal transcription in which children's CT was broken down to reveal the modes of interactions (see Kopcha et al., 2020; which also appear in Kopcha et al., 2019).

In the multimodal transcription, we focused on the five major characteristics of children's CT: abstraction, decomposition, algorithmic thinking, pattern recognition, and debugging. These specific CT skills were present during the participant's embodied interactions. For example, Table 1 displays how the children's decomposition was manifested through the participants' dialogue and gestures.

Table 1. *A Moment That Describes the Modes of Interactions Associated With the Children's Decomposition*

Dialogue	Gesture	Transduction
G: Okay! Now we need a turn. B: Yeah! Now a turn to do that. G: Hold on! [*leans forward*] B: So, that means, right forward (#a), left backward [*eyes are fixed upwards*] (#b). G: Right forward... [*both lean towards the computer*]	a. b.	The pair returns from testing to continue **decomposing** the larger task. Both focus on turning left. The boy *moves his hands to show a turn* where the right wheel moves forward, and the left wheel moves backward. (# a-b)

As shown in Table 1, the participants' CT was laid out through their observable behaviors including discourse and gesturing. Looking at the clusters of moments, we planned to identify those sequences of interaction patterns that could be linked to a specific CT characteristic such as debugging. As a result, to expand on our human analysis, we created *an AI-enhanced Pattern Recognition Approach* that can evaluate large and complex data sets to help identify the patterns of embodied interaction that are specific to CT. Table 2, below, details the process we used to support AI analysis from a social semiotic perspective of multimodality.

Table 2. *Steps to Creating an AI-enhanced Pattern Recognition Approach*

Steps	Description with numbered sub-tasks
Preprocessing	Prior to conducting AI-enhanced analysis, the researcher must **(1)** select a theoretical framework that can guide the research process. This theoretical framework provides the foundation for both **(2)** the video segment(s) that are selected for analysis and **(3)** the multimodal transcription process. Each of these components must be aligned such that the theory helps justify the analysis.
Processing	The researcher can now **(4)** extract images. While extracting images is not technically challenging, the researcher must select the number of images per second that makes the most sense for the questions driving the research and the rate of interaction present in the video.

	The researcher then **(5)** categorizes a subset of images based on the theoretical framework. The images selected must show clearly the interaction of interest and the visual indicators of that interaction in order to accurately **(6)** train the AI. After the initial training, the research must determine if and how to improve the training. A common approach to **(7)** improving the training is to reduce ambiguity in the categorized subset of images while also increasing the number of images to include in the AI training. This may entail returning to the extracted images and selecting new images.
Analysis	Once the AI is trained with a high level of accuracy (80% or higher), the researcher can **(8)** analyze the entire video segment and validate output. For short videos, a human can validate the AI results by viewing each image and confirming the AI output. For larger videos, this task may require a larger team to validate the output or validation may take place by confirming a randomly selected subset of images.
Report	Once the AI output has been validated, the researcher can **(9)** identify methods of post-processing and/or visualizing results. A simple way is to use the AI time stamping feature to show sequences of interaction over time. Visualizing the results, in addition to obtaining frequency counts for each interaction, will help the researcher **(10)** draw conclusions and make sense of the results with respect to related prior studies.

1. *Selecting a Theoretical Framework:* We selected a social semiotic approach to multimodality for the analysis of complex, embodied interactions (Bezemer, 2014). Social semiotics is strongly aligned with studying cognition and learning as an embodied meaning-making activity.

2. *Selecting a Video Segment to Begin the Analysis:* The primary data source was a 100-minute video collected from a rural school in the Southeastern US. Two 5th grade participants were recorded as they worked together to program a robot across a 3'x3' grid of obstacles (Choi et al., 2015). We selected a 5-minute segment from a 100-minute video recording because it had high levels of participant interactions and gestures as they used the computer, robot, and written materials to engage in CT.

3. *Design and Create a Multimodal Transcript of the Selected Video Fragment:* The design of the transcript must convey how various interactions and dialogue play out over time *as they are related to the phenomenon of interest.* To create a multimodal transcript, we adopted Bezemer's *five steps to transcribing multimodal interaction* (2014), which also appear in Kopcha et al. (2020). The transcripts contained an image-by-image breakdown of the video segment used in this study. Those transcripts were used to guide our decisions about what images to extract and use to train the AI (see steps 4 and 5, below), as well as to compare and contrast human-based analysis with the AI output (see step 12).

4. *Extract Images:* We used an image extraction program to extract roughly one frame for every seven images, which resulted in 1353 images over 5 minutes. Researchers need to select the recording ratio in terms of the activity rate in the chosen video fragment. We suggest an increase in the extracted frames per second if there is a higher level of activity. In addition, we resized the extracted images to 128 x 128 resolution and normalized the intensity in all the red, blue, and green channels between 0 and 1 to train our network.

5. *Categorize Subset of Images Based on Theoretical Framework to Train the AI:* According to the theory of social semiotics, there were four dominant interaction types: (1) hand gestures when communicating with each other, (2) use of the computer, (3) interaction with the robot, and (4) use of the student guide. For each of the interaction types, we then selected images that were distinct and uniquely showcased the interaction of interest. We ultimately ended up with 20-25 images per interaction type. We used this classification table to train the AI.

6. *Train the AI:* We designed a minimalist *Convolution Neural Network* (Lecun & Bengio, 1995*)*, consisting of two convolution layers with 32 feature maps and a filter size of 3 x 3. Each convolution layer used the *ReLu activation function* and was followed by batch normalization. Next, we flattened the last convolution layer and connected it to the dense block of size five representing five possible classes. The dense layer used the *sigmoid activation function*. Further reading on the layers can be obtained at Yann Lecun's Keras guide. Finally, we used the *Adam Optimizer* (Kingma & Ba, 2014) with a learning rate of 0.00005.

7. *Revise Training as Needed:* The AI ultimately yielded 80% accuracy of a match with our manual classification. The 20% inaccuracy was the result of some images lacking distinct, unique features altogether or a single image resulting in multiple possible categorization outcomes (e.g., she looks at the computer while he also moves like a robot's wheels).

   In the first trial, the AI was unable to label the interactions correctly due to the disproportionate number of categories used for manual coding compared to the number of samples. That is, there were not enough samples of pictures under each category and for AI to make generalizations. As a result, we increased the number of images listed under each category. In addition, certain frames had more than one interaction per frame.

8. *Analyze Entire Video Segment & Validate Output:* We then analyzed the entire 5-minute segment using the AI. The AI was configured so that the output was a movie consisting of each image (~4 images per second) with a display of the interaction type detected in each image. We then independently viewed the movie in a second-by-second fashion to determine if the AI had appropriately identified the interaction type or if it misidentified the interaction type. Figure 1, below, contains two images from the movie. The first displays an appropriate identification in which the girl is looking off-screen at the computer (Com) and the boy is making a gesture (Gest). The second displays a misidentification. The misidentification was largely due to ambiguity in the image. It was difficult to tell whether the boy was looking downward at the workbook (WB) or if it was at the computer (Com) off-screen to the left of the image.

   When compared to human identification of interactions, the AI was able to accurately identify the children's interaction with the computer 84% of the time. Inaccuracies with identifying interaction with the computer were primarily due to ambiguity in the images where it was unclear if the child was looking at the computer or somewhere else off-screen (see Figure 1 below). The AI accurately identified the children's interactions with each other, through gestures, and interaction with the workbook 65% and 61% of the time, respectively. Inaccuracies with gestures were largely due to the computer detecting small hand movements that never became meaningful gestures. With regard to the workbook, the inaccuracies were due to the AI missing instances of the participants looking downward at the workbook or mislabeling those instances as computer use. The most inaccuracies were associated with interaction with the robot. The AI did not detect any instances, whereas the human analysis identified 11 images that contained the robot.
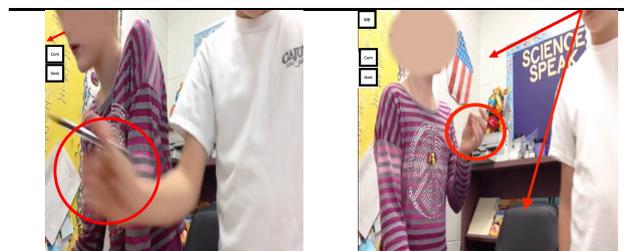

Figure 1. Example Images from the AI Output Displaying Appropriate Identification (a) And Misidentification (b) of Interaction Types

9. *Identify Method of Post-processing & Visualizing Results:* The frequencies can be displayed visually so that it is clear how the interactions played out over time. Our AI provided us with two outputs that supported the creation of a visual display. The first involved a frequency display with a time-stamp (see Figure 2).

```
(array([0, 0, 0, 0, 1]), array([0.01217304, 0.00140641, 0.03665041, 0.      , 1.      ],
      dtype=float32))
(array([0, 0, 0, 0, 1]), array([2.9266805e-05, 1.9357228e-06, 1.4561087e-02, 0.0000000e+00,
      1.0000000e+00], dtype=float32))
(array([0, 0, 0, 1, 0]), array([1.0383282e-04, 0.0000000e+00, 3.3084176e-05, 1.0000000e+00,
      1.0227008e-04], dtype=float32))
(array([0, 0, 0, 1, 0]), array([0.0000000e+00, 6.8170101e-07, 8.1757090e-09, 1.0000000e+00,
      1.1531353e-06], dtype=float32))
(array([0, 0, 1, 0, 0]), array([3.5270514e-07, 1.9708584e-01, 1.0000000e+00, 0.0000000e+00,
      1.2114414e-03], dtype=float32))
(array([0, 0, 1, 0, 0]), array([1.2092181e-07, 7.0519286e-06, 1.0000000e+00, 0.0000000e+00,
      2.7588658e-02], dtype=float32))
(array([0, 1, 0, 0, 0]), array([2.0210858e-04, 1.7966072e-03, 1.0000000e+00, 0.0000000e+00,
      2.4145457e-01], dtype=float32))
(array([0, 1, 0, 0, 0]), array([2.6700931e-05, 1.0000000e+00, 2.7112640e-06, 0.0000000e+00,
      2.8806555e-04], dtype=float32))
(array([1, 0, 0, 0, 0]), array([1.0000000e+00, 1.8151298e-04, 9.7207826e-01, 6.7893684e-02,
      0.0000000e+00], dtype=float32))
(array([1, 0, 0, 0, 0]), array([3.7065128e-04, 7.1923489e-05, 1.0000000e+00, 0.0000000e+00,
      1.2906876e-05], dtype=float32))
(8, 10, 0.8)
```

Figure 2. Visual Display of The Computer Output

The computer output displayed what mode of interaction emerged at what time: 0.0 represented the absence of interaction, whereas a result of 1.0 pointed to the existence of interaction. By looking at this output, we could tell the frequency of a specific type of social interaction (e.g., most frequent; least frequent) as part of a progressive timeline. For example, it was clear that the participants made extensive use of their workbooks as they worked on programming the computer in the 5-minute video segment. This output was particularly useful when determining what mode of interaction became more prevalent in a specific data fragment.

In the context of our AI analysis, our theoretical framework suggested that the use of gestures by our participants was an important interaction to focus on. One way we could visualize the results would be to isolate key moments of gesturing and display them on a timeline. Figure 3, below, illustrates visually one moment of gesturing from our AI results.

Timeline	1:08:13	**1:09:06 -1:09:08**	01:12:03
Interaction Type	#1561	#1591	#1621
Com	✓ (1.0)	✓ (1.0)	✓ (1.0)
Gest	✓ (1.0)	✓ (1.0)	✓ (1.0)
WB	— (0.0)	— (0.0)	✓ (1.0)

Figure 3. Visually One Moment of Gesturing from Our AI Results

10. *Draw Conclusions:* We compared the visual display (see step 9) with our multimodal transcripts (see step 3). Our purpose was to see if the AI was detecting the moments that we had already identified in our multimodal transcript. The frequency input from the AI indicated that the students spent roughly 3.75 minutes in the 5-minute segment working with the computer and workbook as a mode. This was the most prevalent in a specific fragment of the data. The next most prominent interaction was the participants working at the computer with no other modes present (0.98 minutes). The next most frequent result was of the participants holding the robot and referencing their workbook; this occurred 18 times (0.15 minutes).

The only other arrangement identified by the AI was of the participants working at the computer while making a gesture while working with the workbook; this occurred 3 times (0.03 minutes) (e.g., a numerical representation of the numbers with fingers) while working with the workbook. The AI identified no other arrangements. A visual inspection of the 5-minute video clip confirmed that these frequency counts largely reflected the modes of interaction present; the entire clip entailed the participants working to program the computer, drawing on different tools throughout the process (e.g., the robot, their workbook).

3. Conclusion

To conclude, the technology-driven, AI-enhanced approach to analyzing CT went beyond the current, often reductive methods. The results of the AI analysis are promising and have significant implications for further research. Through the display of relative frequencies, the researchers could easily interpret shifts across the modes in the designated categories. Likewise, spoken dialogue could be added to that timeline to enhance further a transduction approach. In this way, AI would help establish the behavioral indicators of CT that may be missed or overlooked through traditional pre-post test methods, offering new possibilities for integrating, assessing, and studying computer science in K-12 settings.

References

Andrade, A., Delandshere, G., & Danish, J. A. (2016). Using Multimodal Learning Analytics to Model Student Behaviour: A Systematic Analysis of Behavioural Framing. Journal of Learning Analytics, 3(2), 282-306

Bezemer, J. (2014). Multimodal transcription: A case study. In S. Norris & C. D. Maier (Eds.), Interactions, images and texts (pp. 155-169). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9781614511175.155

Black, J. B., Segal, A., Vitale, J., & Fadjo, C. L. (2012). Embodied cognition and learning environment design. *Theoretical foundations of learning environments*, *2*, 198-223.

Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. In Proceedings of the 1st international conference on learning analytics and knowledge.

Blikstein, P. (2013). Multimodal learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 102-106).

Choi I, Hill R, Kopcha T, Mativo J, Bae Y, Hodge E, Way W, Mcgregor J, Shin S, Kim S, Choi J, Um K. (2015). Danger zone: A STEM-integrated robotics unit – My design journal (student guide). Seoul, Korea: RoboRobo Co., Ltd.

Chung, C. Y., & Hsiao, I. H. (2019). An exploratory study of augmented embodiment for computational thinking. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion* (pp. 37-38). https://doi.org/10.1145/3308557.3308676

Järvelä, S., Gašević, D., Seppänen, T., Pechenizkiy, M., & Kirschner, P. A. (2020). Bridging learning sciences, machine learning and affective computing for understanding cognition and affect in collaborative learning. *British Journal of Educational Technology*, *51*(6), 2391-2406.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kopcha, T. & Ocak, C. (2019). Embodiment of Computational Thinking During Collaborative Robotics Activity. In Lund, K., Niccolai, G. P., Lavoué, E., Hmelo-Silver, C., Gweon, G., & Baker, M. (Eds.), *A Wide Lens: Combining Embodied, Enactive, Extended, and Embedded Learning in Collaborative Settings*, 13th International Conference on Computer Supported Collaborative Learning (CSCL) 2019, Volume 1 (pp. 464-471). Lyon, France: International Society of the Learning Sciences. https://doi.dx.org/10.22318/cscl2019.464

Kopcha, T. J., Ocak, C., & Qian, Y. (2020). Analyzing children's computational thinking through embodied interaction with technology: A multimodal perspective. *Educational Technology Research and Development*, 1-26. https://doi.org/10.1007/s11423-020-09832-y

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

Luckin, R., & Cukurova, M. (2019). Designing educational technologies in the age of AI: A learning sciences-driven approach. *British Journal of Educational Technology*, *50*(6), 2824-2838.

Melcer, E. (2017). Moving to learn: Exploring the impact of physical embodiment in educational programming games. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems* (pp. 301-306). https://doi.org/10.1145/3027063.3027129

Sharma, K., Papamitsiou, Z., & Giannakos, M. (2019). Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach. *British Journal of Educational Technology*, *50*(6), 3004-3031.

Yang, C., Chiang, F. K., Cheng, Q., & Ji, J. (2021). Machine Learning-Based Student Modeling Methodology for Intelligent Tutoring Systems. *Journal of Educational Computing Research*, 0735633120986256.