Instructional Activity Recognition Using A Transformer Network with Multi-Semantic Attention

Matthew Korban, Scott T. Acton
C.L. Brown Dept. of Electrical and Computer Engineering
University of Virginia

Peter Youngs
Department of Curriculum, Instruction, and Special Education
University of Virginia

Jonathan Foster
Department of Education Theory & Practice
University at Albany

2024 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI) | 979-8-3503-6011-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/SSIAI59505.2024.10508634

***Abstract*—Instructional activity recognition is an analytical tool for the observation of classroom education. One of the primary challenges in this domain is dealing with the intricate and heterogeneous interactions between teachers, students, and instructional objects. To address these complex dynamics, we present an innovative activity recognition pipeline designed explicitly for instructional videos, leveraging a multi-semantic attention mechanism. Our novel pipeline uses a transformer network that incorporates several types of instructional semantic attention, including teacher-to-students, students-to-students, teacher-to-object, and students-to-object relationships. This comprehensive approach allows us to classify various interactive activity labels effectively. The effectiveness of our proposed algorithm is demonstrated through its evaluation on our annotated instructional activity dataset.**

I. Introduction

Activity recognition is an active area of research in computer vision with numerous practical applications, including surveillance, autonomous vehicles, human-computer interaction, and video annotation [1]. Recently, instructional activity recognition (IAR) has been explored by several researchers [2], [3], [4], [5] to enhance the analysis of classroom education. IAR can provide automated and immediate feedback to instructors in training, reducing the workload of human annotators. IAR, however, is challenging because the interactions between instructional semantics, including the teacher, students, and instructional objects in classrooms, are complex. Moreover, the number and heterogeneity of these interactions can be significant in a classroom setting due to the multitude of instructional semantics. We show that a specialized transformer network can effectively model these intricate classroom interactions using its attention mechanism. Furthermore, the parallelization capability of the transformer network [6] enhances the efficiency of processing such a large number of interactions. To this end, we propose a transformer network encompassing multiple instructional semantic attention types to capture the relations between instructional semantics in classroom videos. We suggest that four types of instructional relationships exist in a classroom setting, including the relations between teacher-to-students, students-to-students, teacher-to-object, and students-to-object modeled using the four proposed semantic attention mechanisms (see Fig. 1 for more details). To our knowledge, we are the first to capture these systematic relations in classroom settings using semantic attention mechanisms.

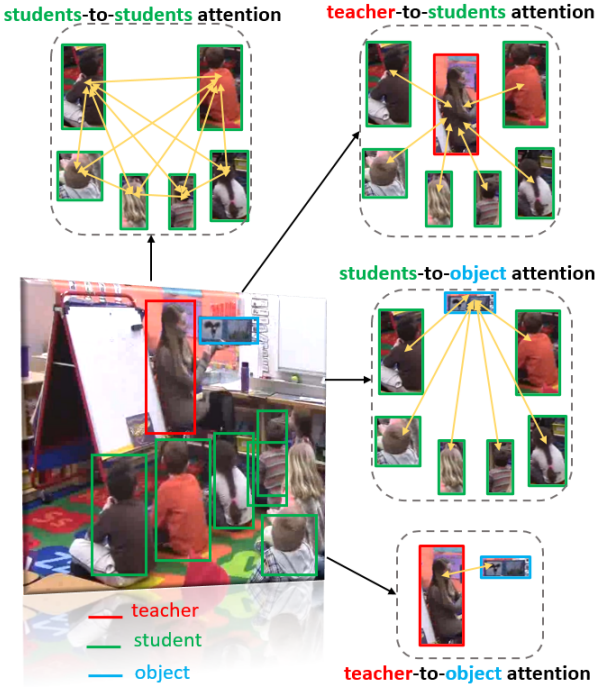

Fig. 1. Four types of multi-semantic attention proposed to model the complex classroom interactions between teacher, students and object in an example where a teacher is discussing a book with students.

II. Related Work

Earlier methods focused on designing a manual feature representation encompassing both the spatial and temporal aspects of human actions. These features include spatiotemporal changes in videos, such as keypoint trajectories [7], human pose deformations [8], and differences in depth images [9]. Based on data-driven mechanisms, deep network solutions offer more robust feature representations than those with handcrafted features [10]. Several deep architectures have been applied including the convolutional neural network (CNN) [11], the recurrent neural network (RNN) [12], the long-short term memory (LSTM) network [13], and the graph convolutional network (GCN) [14]. A state-of-the-art deep network is the transformer, that has been widely used in the last few years for activity recognition [15], [16].

Numerous methods have been developed for identifying activities in educational video recordings. [2] employed sil-

houettes of the educator to derive motion characteristics. [3] introduced a more enhanced feature set that includes elbow angles and facial and hand movements, utilizing a primitive-based coupled hidden Markov model (PCHMM) to identify seven specific teaching activities. [5] advanced this field with a sophisticated deep learning architecture, integrating a multi-modal attention layer to understand long-term semantic relationships in instructional videos. Lastly, [4] proposed focusing on significant video segments, such as the skeletal posture, to model activities in classroom settings effectively.

However, an effective method for systematically calculating the interactions among instructional semantics, including teachers, students, and objects, has not yet been established.

III. Methods

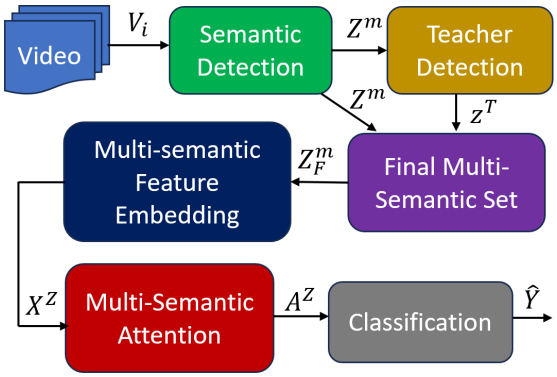

Fig. 2. The overview of the proposed pipeline of instructional activity recognition.

Fig. 2 shows the overview of the suggested pipeline for instructional activity recognition. The inputs of the proposed pipeline are instructional videos, where each video $V_i = \{f_m, m = 0, 1, ..., M\}$ consists of M action frames $f_m \in \mathbb{R}^{H \times W \times 3}$. Here H and W are the height and width of the frames. The outputs of the pipeline are the predicted action class labels for each frame as $\hat{Y} = \{\hat{y_m}, m = 0, 1, 2, ..., M\}$. The proposed pipeline includes several steps that are explained as follows:

A. Semantic Detection

We first use a state-of-the-art network [17] to detect the semantics, specifically the persons and objects. So that for each frame f_m, we will have N' detected semantics as $Z^m = \{z_i, i = 0, 1, 2, ..., N'\}$, where $z_i \in \mathbb{R}^{h \times w \times 3}$, that h and w are the height and width of the each semantic. Each s_i is then resized to the fixed dimension of $h' \times w'$.

B. Final Multi-semantic Set

Next, an age estimation algorithm [18] is used to detect the teacher by way of selecting the person with the highest estimated age. The aforementioned age estimation algorithm can estimate age in cluttered environments and in the event of partial face occlusion, a common occurrence in classroom videos. In this stage the instructional semantics set is $Z^m = \{z^T, z_i^S, z_j^O, i = 0, 1, 2, ..., N^S, j = 0, 1, 2, ..., N^O\}$, that includes one teacher (z^T), N^S number of students (z_i^S), and N^O number of objects (z_j^O). In this study the closest object to the teacher, z^O, is considered as the primary instructional tool for further analysis as shown as follows:

$$z^O = \underset{j}{\text{argmin}}(\left\| Center(\mathbf{z_j^O}) - Center(\mathbf{z^T}) \right\|_2) \quad (1)$$

where the operator *Center* defines the center of each semantic image. So, the final set of instructional semantics provided to the network becomes $Z_F^m = \{z^T, z_i^S, z^O, i = 0, 1, 2, ..., N^S\}$.

C. Multi-semantic Feature Embedding

The feature embedding step is formulated as follows:

$$X^Z = Conv(Z_F^m, W^S) : \mathbb{R}^{N \times h' \times w' \times 3} \rightarrow \mathbb{R}^{N \times d_f}. \quad (2)$$

where, *Conv* is a convolutional operator, $N = N^S + 2$ is the total number of instructional semantics, and W^S is the convolutional kernel weights. So, the subsequent embedded semantic feature set becomes $X^Z = \{X^T, X_i^S, X^O, i = 0, 1, 2, ..., N^S\}$, where X^T, X_i^S, and $X^O \in \mathbb{R}^{d_f}$ are embedded semantic features for the teacher, students, and objects respectively.

D. Multi-Semantic Attention

We hypothesize that instructional activity labels can be defined in light of four types of relationships among instructional semantics. These relationships are teacher-to-students, students-to-students, teacher-to-object, and students-to-object. Hence, we define four types of attention to capture such relations among instructional semantics.

The attention aims to compute the correlations between the *query* (Q) and *keys* (K) and then map the correlation results to *values* (V). The four suggested attention types to represent the correlations among instructional semantic features are formulated as follows:

$$\begin{aligned} A^{T-S} &= Softmax(\frac{Q^T(K^S)^T}{\sqrt{d_h}})V^S, \\ A^{S-S} &= Softmax(\frac{Q^S(K^S)^T}{\sqrt{d_h}})V^S, \\ A^{T-O} &= Softmax(\frac{Q^T(K^O)^T}{\sqrt{d_h}})V^O, \\ A^{S-O} &= Softmax(\frac{Q^S(K^O)^T}{\sqrt{d_h}})V^O, \end{aligned} \quad (3)$$

where A^{T-S}, A^{S-S}, A^{T-O}, and A^{S-O} are teacher-to-students, students-to-students, teacher-to-object, and students-to-object attention types, respectively. d_h is the size of the attention head and T is a transpose operation. The multi-semantic queries (Q^Z), keys (K^Z), and values (V^Z) are calculated as follows:

$$Q^Z = X^Z W_Q^Z,\ K^Z = X^Z W_K^Z,\ V^Z = X^Z W_V^Z \quad (4)$$

where W_Q^Z, W_K^Z, and W_V^Z are projecting weights.

The set of multi-semantic attention is used as the final representation of multi-semantic features:

$$A^Z = \{\alpha A^{T-S}, \beta A^{S-S}, \gamma A^{T-O}, \zeta A^{S-O}\} \quad (5)$$

where α, β, γ, and ζ are the adjusting parameters that control the importance of multi-semantic attention types. Some examples of multi-semantic relations in classroom videos and corresponding attention types are shown in Fig. 1. Later Table I illustrates the relevance between the multi-semantic attention types and the corresponding activity class labels.

E. Multi-layer Transformer

The suggested transformer network includes multiple layers where the relation between layers $l-1$ and l is illustrated as follows:

$$\begin{aligned} \hat{U}^l &= MSA(U^{l-1}) + U^{l-1}, \quad l \in \{1, ..., L\}, \\ U^l &= MLP(Norm(\hat{U}^l)) + \hat{U}^l, \quad l \in \{1, ..., L\}, \end{aligned} \quad (6)$$

where MSA is the multi-semantic attention, U is the output of the layer, $\hat{U}$ is the output of the intermediate layer output, L is the number of layers, $Norm$ is a normalization layer, and *MLP* is a multilayer perception layer.

F. Classification

Finally, the classification phase can be formulated as follows:

$$\hat{Y} = Softmax(Conv(A^Z, W^F)), \quad (7)$$

where $\hat{Y}$ is the frame prediction scores set, $Conv : \mathbb{R}^{N \times df} \rightarrow \mathbb{R}^C$, W^F is the kernel weights, and C is the number of activity classes. The loss function of the transformer network is shown as follows:

$$Loss = -\sum_{c=1}^{C} y^{(c)} log\hat{y}^{(c)} \quad (8)$$

IV. Experimental Results

A. Dataset

We utilized 50 hours of annotated instructional videos acquired by the University of Virginia School of Education of Human Development. Twelve instructional activity labels are used in these experiments. These activity labels with their relevant multi-semantic attention are shown in Table I.

B. Results

In our experiments, we used the F1 score metric based on frame-level prediction as

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (9)$$

TABLE I. Twelve instructional activity labels and their relevant multi-semantic attention (MSA) types

Activity Class Label	Relevant MSA
Whole Class Activity	A^{T-S}, A^{S-S}
Individual Activity	A^{T-S}, A^{S-S}
Small Group Activity	A^{T-S}, A^{S-S}
Transition	A^{T-S}, A^{S-S}
Teacher Supporting with Students	A^{T-S}, A^{S-S}
Teacher Supporting without Students	A^{T-S}
Teacher Supporting one Student	A^{T-S}
Using or Holding Book	A^{T-O}, A^{S-O}
Using or Holding Worksheet	A^{T-O}, A^{S-O}
Using or Holding Instructional Tool	A^{T-O}, A^{S-O}
Presenting with Technology	A^{T-O}, A^{S-O}

where TP, FP, and FN are true positive, false positive, and false negative predicted frames, respectively.

Fig. 3 shows the performance of our proposed pipeline for twelve instructional activity labels based on the F1 score. The average F1 score for all instructional class labels was *0.56*.

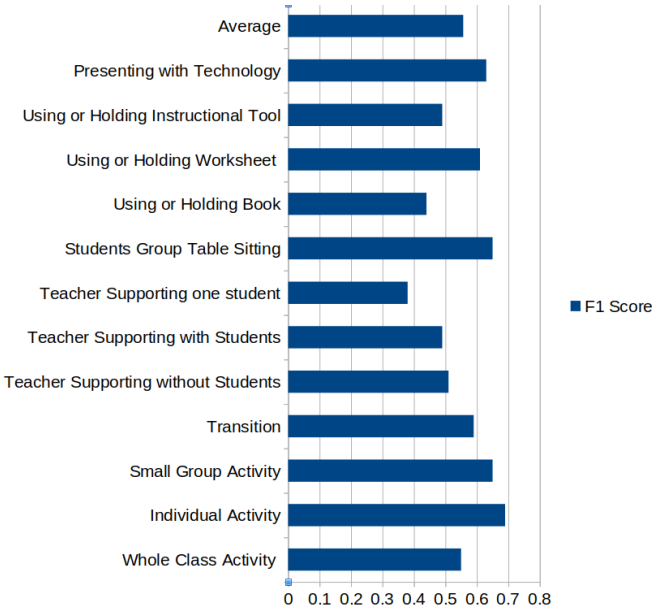

Fig. 3. The performance of the proposed pipeline for twelve activity labels.

Fig. 5 illustrates an ablation study of different types of multi-semantic attention. As can be seen, using the full model with all the attention types led to maximum performance.

Fig. 5 shows the impact of different types of multi-semantic attention on the activity prediction scores for two examples of our classroom videos. As can be seen, for the activity label, "teacher using instructional tool", the highest prediction score is achieved by using the most relevant attention type, teacher-to-object ($T-O$). On the other hand, the most relevant attention type for the activity label, "small group activity" was teacher-to-students ($T-S$) based on the highest prediction score.

Six transformer layers were used in our implementation. The learning rate was $1e^{-4}$. All the weights are initialized

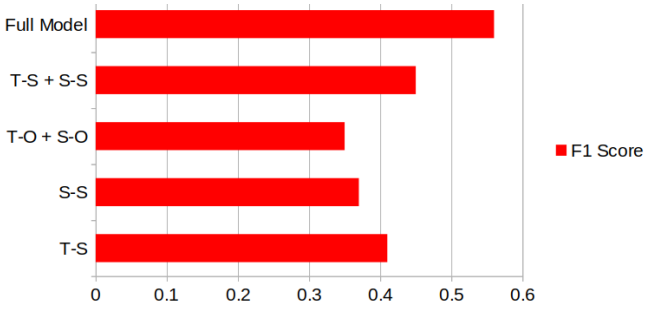

Fig. 4. The ablation study on the impact of various combinations of multi-semantic attention types of teacher-to-students ($T-S$), students-to-students ($S-S$), teacher-to-object ($T-O$), and students-to-object ($S-O$) on the overall performance of instructional activity recognition.

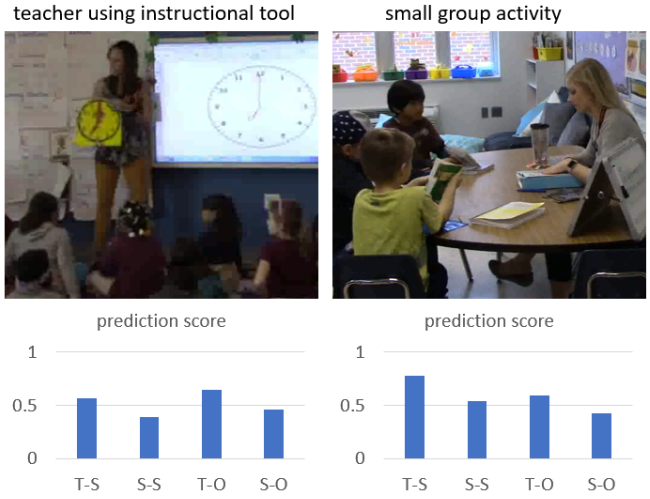

Fig. 5. The ablation study on the impact of various attention types of teacher-to-students ($T-S$), students-to-students ($S-S$), teacher-to-object ($T-O$), and students-to-object ($S-O$) on the activity prediction scores for two examples of instructional video.

with random values. All the experiments are conducted using PyTorch 1.7 on a server PC with dual Nvidia RTX 3090 GPUs (24GB VRAM), AMD Ryzen Threadripper 3990X 64-Core Processor, and 256GB of RAM.

V. Conclusion

This paper introduces various multi-semantic types of attention including teacher-to-students, students-to-students, teacher-to-object, and students-to-object. Such a framework can model complex relationships within instructional videos. Our proposed pipeline encompasses several stages: detecting semantics, identifying teachers, recognizing instructional objects, generating multi-semantic sets, applying multi-semantic attention, and finally, classifying data. We evaluated this pipeline using our unique annotated instructional activity dataset, and the results demonstrate the effectiveness of our approach.

References

[1] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017.

[2] N. Nida, M. H. Yousaf, A. Irtaza, S. A. Velastin *et al.*, "Instructor activity recognition through deep spatiotemporal features and feedforward extreme learning machines," *Mathematical Problems in Engineering*, vol. 2019, 2019.

[3] H. Ren and G. Xu, "Human action recognition in smart classroom," in *Proceedings of fifth IEEE international conference on automatic face gesture recognition*. IEEE, 2002, pp. 417–422.

[4] F.-C. Lin, H.-H. Ngo, C.-R. Dow, K.-H. Lam, and H. L. Le, "Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection," *Sensors*, vol. 21, no. 16, p. 5314, 2021.

[5] H. Li, Y. Kang, W. Ding, S. Yang, S. Yang, G. Y. Huang, and Z. Liu, "Multimodal learning for classroom activity detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 9234–9238.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] G. Burghouts, K. Schutte, R. J.-M. ten Hove, S. van den Broek, J. Baan, O. Rajadell, J. van Huis, J. van Rest, P. Hanckmann, H. Bouma *et al.*, "Instantaneous threat detection based on a semantic representation of activities, zones and trajectories," *Signal, Image and Video Processing*, vol. 8, pp. 191–200, 2014.

[8] X. Yang and Y. Tian, "Effective 3d action recognition using eigenjoints," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 2–11, 2014.

[9] ——, "Super normal vector for activity recognition using depth sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 804–811.

[10] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: A survey," *Image and Vision Computing*, vol. 55, pp. 42–52, 2016.

[11] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "Multi-stream cnn: Learning representations based on human-related regions for action recognition," *Pattern Recognition*, vol. 79, pp. 32–43, 2018.

[12] W. Li, L. Wen, M.-C. Chang, S. Nam Lim, and S. Lyu, "Adaptive rnn tree for large-scale human action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1444–1452.

[13] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[14] M. Korban and X. Li, "Ddgcn: A dynamic directed graph convolutional network for action recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 761–776.

[15] M. Korban, P. Youngs, and S. T. Acton, "A semantic and motion-aware spatiotemporal transformer network for action detection," *in submission, IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2023.

[16] ——, "A multi-modal transformer network for action detection," *Pattern Recognition*, vol. 142, p. 109713, 2023.

[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[18] M. Korban, P. Youngs, and S. T. Acton, "Taa-gcn: A temporally aware adaptive graph convolutional network for age estimation," *Pattern Recognition*, vol. 134, p. 109066, 2023.