Long-term Human Video Activity Quantification of Student Participation

Venkatesh Jatla*, Sravani Teeparthi*, Marios S. Pattichis*, Sylvia Celedón-Pattichis‡, and Carlos LópezLeiva‡
*Department of Electrical and Computer Engineering
The University of New Mexico, Albuquerque, NM, USA. Email:{steeparthi, venkatesh369, pattichi}@unm.edu
‡Department of Language, Literacy, and Sociocultural Studies
The University of New Mexico, Albuquerque, NM, USA. Email: {sceledon, callopez}@unm.edu

***Abstract*—Research on video activity recognition has been primarily focused on differentiating among many diverse activities defined using short video clips. In this paper, we introduce the problem of reliable video activity recognition over long videos to quantify student participation in collaborative learning environments (45 minutes to 2 hours).**

Video activity recognition in collaborative learning environments contains several unique challenges. We introduce participation maps that identify how and when each student performs each activity to quantify student participation. We present a family of low-parameter 3D ConvNet architectures to detect these activities. We then apply spatial clustering to identify each participant and generate student participation maps using the resulting detections.

We demonstrate the effectiveness by training over about 1,000 3-second samples of typing and writing and test our results over ten video sessions of about 10 hours. In terms of activity detection, our methods achieve 80% accuracy for writing and typing that match the recognition performance of TSN, SlowFast, Slowonly, and I3D trained over the same dataset while using 1200x to 1500x fewer parameters. Beyond traditional video activity recognition methods, our video activity participation maps identify how each student participates within each group.

***Index Terms*—Action recognition, temporal modeling, 3D-ConvNets, low-complexity neural networks**

I. Introduction

The paper introduces new methods for long-term human video activity quantification. Ultimately, our goal is to quantify student participation by identifying video segments where a student performs a particular activity (e.g., writing or typing). Thus, our approach extends traditional video analysis methods, which primarily focus on differentiating among short-term activities.

We present examples of classroom video activities in Fig. 1. Here, we are only interested in the group activity that is happening at the table that is closest to the camera. Then, as shown in Fig. 1(a), we define structural clutter as activity occurring in the background. Other challenges include multiple activities (see Fig. 1(b)), concurrent activities (see Fig. 1(c)), and activities associated with different students (see Fig. 1(d)).

Our proposed approach consists of two essential parts. First, we use an object detection system to locate where the video activity may be happening. Second, we apply a video activity recognition system to classify the activity. For general purpose object recognition, we consider the Faster R-CNN [1]. For comparison to our proposed approach to video activity recognition, we consider three alternative methods: TSN [2], Slowfast & Slow-only [3], and I3D [4]. For TSN, the video is divided into multiple segments, and the final score is derived through a combination of three separate classifiers based on RGB, frame differences, and optical flow with 24M trainable parameters [2]. Slowfast & Slow-only [3] is built to recognize slow and fast actions. For comparison purposes, we will consider a version of Slowfast & Slow-only using ResNet [5] with 32M trainable parameters. For I3D [4], we consider the use of an inflated Inception Network [6] with 27M parameters.

We also provide a summary of recent related research on classroom video analysis. In [7], the authors developed fast methods for hand detection. In [8], [9], and [10], the authors developed methods for person detection and talking detection. In [11], the authors used video analysis methods to develop a speech recognition system for English and Spanish.

Compared to previous approaches, the current paper is based on the development of low-complexity systems. For low-complexity, we consider the development of classifiers with substantially smaller numbers of trainable parameters. As we describe here, the approach is successful in the sense that we reduce the number of trainable parameters by more than a thousand times.

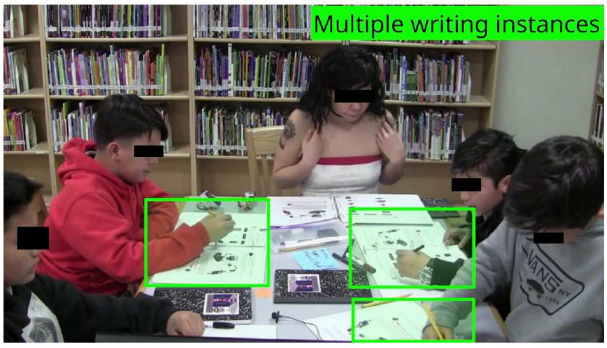
The rest of the paper is organized as follows. We describe the method in section II. We provide results in section III. We provide concluding remarks in section IV.

II. Methodology

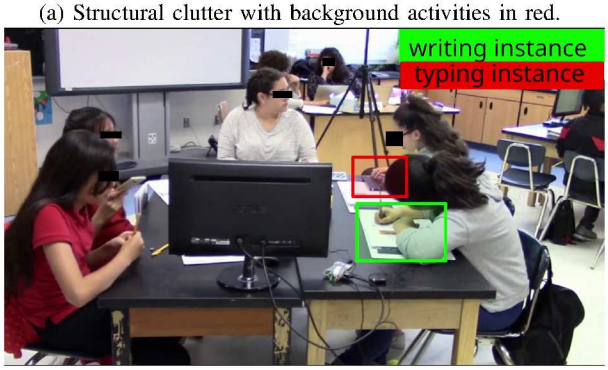
We present the general system diagram in Fig. 2. As described earlier, we process each video in two steps. First, we use object detection to generate activity proposal regions. Second, we use a classifier over each proposal region. The results are used to generate activity maps. For this paper, we only consider writing and typing activities.

For generating activity proposals, we only consider the case for typing detection. To detect typing, we focus on keyboard detection. To extend the results to writing, we will integrate a method for hand detection that we are currently developing [7].

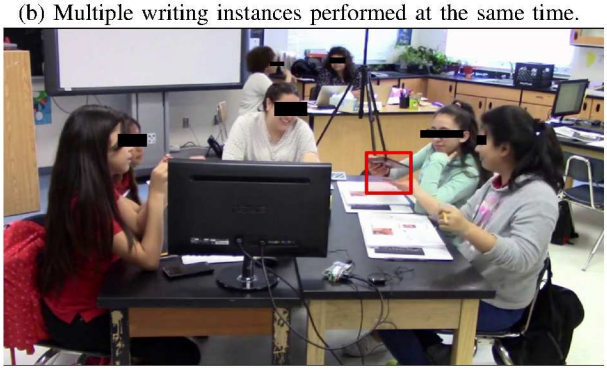
(a) Structural clutter with background activities in red.



(b) Multiple writing instances performed at the same time.



(c) Different activities performed at the same time.



(d) Keyboard movement from student to student.

Fig. 1: Classroom activity challenges.

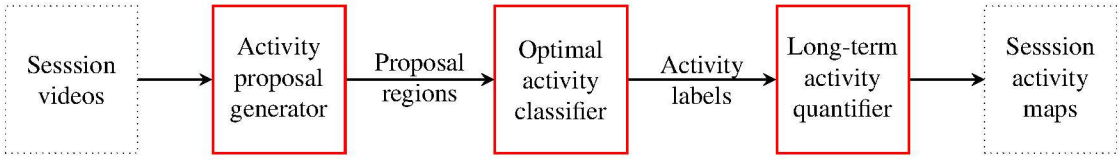


Fig. 2: General system diagram.

For video activity classification, we consider the use of 3D ConvNets that was first introduced by [12]. Here, we build a family of neural network architectures as shown in Fig. 3. Each neural network is made of 3D-ConvNets as shown in Fig. 3(a). The neural network architecture is shown in Fig. 3(b). We want to determine the optimal architecture based on different depth levels.

III. Results

We summarize the results in two sections. First, we describe the dataset and the hardware used to evaluate the activity detection systems in III-A. Second, we compare our low parameter dyadic architecture against the state-of-the-art systems detecting typing and writing. For the comparisons, we measure spatiotemporal accuracy for typing detection. We also provide an example of a quantification map for summarizing typing activity.

A. Dataset

A typical session in AOLME has 1920×1080 resolution and 30 or 60 frames per second (FPS), and lasts around 1 hour to 1.5 hours. Before labeling the activities, we transcode the videos to 848×480 and 30 FPS, standardizing the resolution and frame rate. We then process the transcoded video through MATLAB video labeler, labeling typing and writing instances. We use a spatially fixed bounding box to label an instance of the activity. The videos are broken into 3-second segments for further processing.

We maintain dataset diversity by taking video instances across sessions from different groups over three years. We split the dataset into training, validation, and testing at the session level, ensuring that each set contains samples from a different session. For typing, we used $2,106$ positive and $1,992$ negative samples, 84% for training and 16% for validation. For writing, we used $1,996$ positive and $1,186$ negative samples, 76% for training and 24% for validation. Each video sample was trimmed to 3-second duration as shown in Fig. 4.

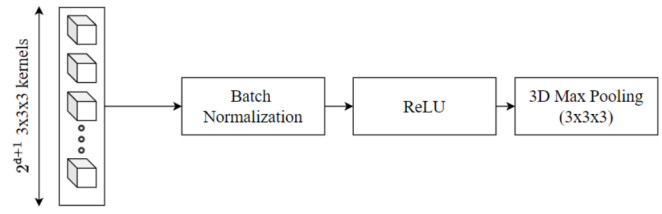
B. Results for Writing and Typing

We summarize our activity recognition results in Table I. Our method performs better in typing recognition with far fewer parameters and is only 3% less accurate in writing

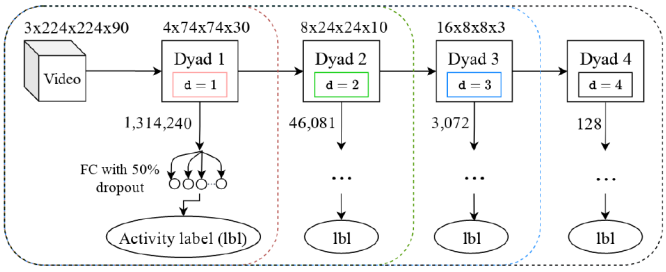
detection with the lowest variation in performance (e.g., smallest standard deviation). Due to the lower-parameter count, our proposed model avoids overfitting and is expected to generalize better. Furthermore, the low number of parameters allows for implementations with low computational resources and limited memory ($\approx$ 350 MB). For training the model, we use a Dell Precision 7920 with Intel Xeon 4208 CPU @ 2.10GHz server and 128 GB DDR4 RAM. The system used an NVIDIA RTX 5000 GPUs each having 16GB DDR6 memory with PyTorch. For the competitive methods, we used the implementations provided in MMACTION2 [13].

We use Faster RCNN [1] to propose regions for typing detection. We then trim the proposed regions and classify them using our activity classifier identifying typing. We process complete sessions and evaluate performance every second.

Table II summarizes the results across 7 sessions. We can see from the table that we are getting good accuracy in identifying regions where there is no typing. On average, we got a sensitivity of 0.58 for typing detection. We found that most of the error is due to pauses during typing. To quantify the detection, we also provide Intersection over Union (IoU) scores. From the results, we can see that typing regions are detected with high accuracy. We present examples in Fig. 5. For the false positive example, we had hand motion near the keyboard that was mistakenly classified as typing. We also provide an example of typing quantification in Fig. 6. From the activity map, we can see that the method is good at detecting long typing instances.



(a) ConvNet.



(b) Neural Network Parameter Optimization.

Fig. 3: 3D-CNN architecture optimization. A family of architectures is generated for depth d. At depth d, we have $2^{d+1}, 3 \times 3 \times 3$ kernels, followed by Batch Normalization, ReLU, and 3D MaxPooling.

TABLE I: Summary of video activity classification results for writing and typing. The table provides classification accuracies. The optimal classifier had $d = 4$.

Activity recognition system	Number of parameters	Saved model size in MB	Typing, No-Typing Mean, SD	Writing, No-Writing Mean, SD
I3D	27M (1421x)	218 (1329x)	89.7, **3.6**	83.5, 7.5
Slowfast (SF)	34M (1789x)	269 (1640x)	89.0, 4.3	**84.2**, 7.5
Slow-only (SO)	32M (1684x)	253 (1542x)	89.3, 4.1	79.7, 8.4
TSN	24M (1264x)	188 (1132x)	85.3, 5.9	76.0, 11.5
3D-CNN (ours)	**0.0190 (1x)**	**0.164 (1x)**	**91.0**, 5.9	81.2, **4.4**

TABLE II: Spatio-temporal performance of the proposed method. The test data included seven complete video sessions with a duration of 1 hour to 1.5 hours, collected over three years. Temporal accuracy is calculated by comparing predicted typing instances against ground truth for every second. Spatial accuracy is estimated using IoU scores of temporal true positive instances.

Session	Temporal accuracy Sens, Spec, Acc	Spatial accuracy # TP	Spatial accuracy median IoU
C1L1P-E, Mar 02	0.45, 0.83, 0.75	516	0.70
C1L1P-C, Apr 13	0.58, 0.82, 0.79	4575	0.71
C1L1P-C, Mar 30	0.37, 0.95, 0.94	39	0.80
C2L1P-B, Feb 23	0.54, 0.88, 0.84	399	0.78
C2L1P-D, Mar 08	0.64, 0.66, 0.66	727	0.77
C3L1P-C, Apr 11	0.63, 0.64, 0.63	251	0.80
C2L1P-D, Mar 08	0.82, 0.52, 0.79	278	0.40

IV. Conclusions

The paper presented a low-parameter method for detecting typing and writing instances in collaborative learning environments. Compared to alternative methods, our low-parameter approach performed better for typing and was competitive in writing detection (at about 3% less) while using 1200x to 1500x fewer parameters. We also provided a complete framework that summarizes typing over an entire session. In future work, we will be working on a complete writing quantification framework and improving our existing typing framework.

V. Acknowledgments

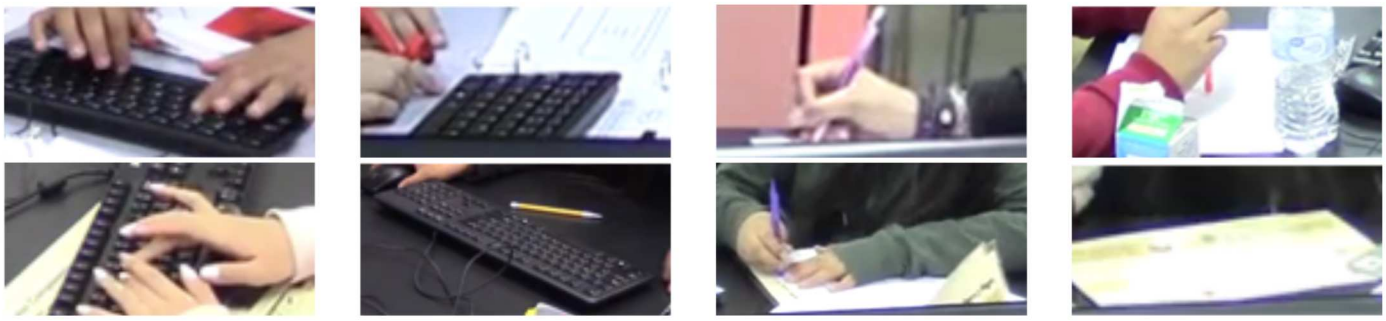
This material is based upon work supported by the National Science Foundation under the AOLME project (Grant No. 1613637), the AOLME Video Analysis project (Grant No. 1842220), and the ESTRELLA project (Grant No. 1949230). Any opinions or findings of this paper reflect the views of the author. They do not necessarily reflect the views of NSF.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[2] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.

[3] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6202–6211.



(a) Typing instances. (b) No-Typing instances. (c) Writing instances. (d) No-Writing instances.

Fig. 4: Examples from 3-second video samples used for training, validation, and testing.

(a) TP with IoU of 0.87. (b) FP caused due to a book having similar pattern as keyboard.

Fig. 5: Typing detection examples. True positives (TP) are denoted by green bounding boxes. False positives (FP) are denoted by yellow bounding boxes.

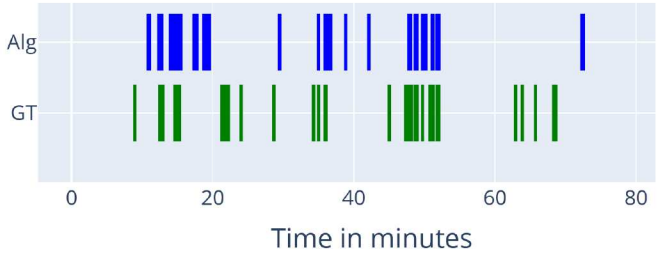


Fig. 6: Typing activity map for a session (C1L1P-E, Mar 02). In this plot, we compare our framework predictions (blue) against ground truth (green).

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[5] R. Christoph and F. A. Pinz, "Spatiotemporal residual networks for video action recognition," *Advances in Neural Information Processing Systems*, pp. 3468–3476, 2016.

[6] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*. PMLR, 2015, pp. 448–456.

[7] S. Teeparthi, V. Jatla, M. S. Pattichis, S. Celedón-Pattichis, and C. L. Leiva, "Fast hand detection in collaborative learning environments," in *19th International Conference CAIP*. Springer, 2021.

[8] W. Shi, M. S. Pattichis, S. Celedón-Pattichis, and C. L. Leiva, "Talking detection in collaborative learning environments," in *19th International Conference CAIP*. Springer, 2021.

[9] ——, "Person detection in collaborative group learning environments using multiple representations," in *2021 55th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2021.

[10] P. Tran, M. S. Pattichis, S. Celedón-Pattichis, and C. L. Leiva, "Facial recognition in collaborative learning videos," in *19th International Conference CAIP*. Springer, 2021.

[11] L. S. Tapia, M. S. Pattichis, S. Celedón-Pattichis, and C. L. Leiva, "Bilingual speech recognition by estimating speaker geometry from video data," in *19th International Conference CAIP*. Springer, 2021.

[12] A. E. U. Cerna, L. Jing, C. W. Good, S. Raghunath, J. D. Suever, C. D. Nevius, G. J. Wehner, D. N. Hartzel, J. B. Leader, A. Alsaid *et al.*, "Deep-learning-assisted analysis of echocardiographic videos improves predictions of all-cause mortality," *Nature Biomedical Engineering*, pp. 1–9, 2021.

[13] M. Contributors, "Openmmlab's next generation video understanding toolbox and benchmark," https://github.com/open-mmlab/mmaction2, 2020.