Instructional Activity Detection Using Deep Neural Networks

Matthew Korban
C.L. Brown Dept. of Electrical and Computer Engineering
University of Virginia
Charlottesville, Virginia 22903
Email: acw6ze@virginia.edu

Peter Youngs
Department of Curriculum, Instruction, and Special Education
University of Virginia
Charlottesville, Virginia 22903
Email: pay2n@virginia.edu

Scott T. Acton
C.L. Brown Dept. of Electrical and Computer Engineering
University of Virginia
Charlottesville, Virginia 22903
Email: acton@virginia.edu

***Abstract*—Analyzing instructional videos via computer vision and machine learning holds promise for several tasks, such as assessing teacher performance and classroom climate, evaluating student engagement, and identifying biases in instruction. The traditional way of evaluating instructional videos depends on manual observation with human raters, which is time-consuming and requires a trained labor force. This paper tests several deep network architectures in the automation of instructional video analysis, where the networks are tailored to recognize classroom activity. Our experimental setup includes a set of 250 hours of primary and middle school videos that are annotated by expert human raters. We present several strategies to handle varying lengths of instructional activities, a major challenge in the detection of instructional activity. Based on the proposed strategies, we enhance and compare different deep networks for detecting instructional activity.**

I. Introduction

Evaluation of classroom activities is essential for instructors while improving their teaching skills [1]. As a result, accurate feedback from classroom evaluation drastically affects students' classroom engagement and enhances the quality of education [2]. Traditionally, providing such feedback requires considerable labor and manual rating from trained experts, which is expensive and time-consuming. Using deep learning models is an efficient and effective solution to this issue as they can automate the evaluation of teacher activities, reducing manual labor work and errors caused by humans [3]. Therefore, this paper compares several deep network architectures based on their ability to detect instructional activities. Multiple strategies have also been presented to enhance the effectiveness of deep models in detecting instructional activities. The new strategies include: (1) an adaptive sampling algorithm for selecting critical frames in classroom videos; (2) a new loss function incorporating frame-level and sequence-level prediction; (3) a post-processing algorithm for detecting the start and end frames of long actions, and (4) a motion enhancement algorithm to make the motion features insensitive to camera movements. The first three strategies address the issue of varying lengths of instructional activities, a significant challenge in detecting instructional activities. The fourth strategy makes the pipeline more robust to camera movement.

979-8-3503-3959-8/23/$31.00 ©2023 IEEE

II. Related Work

Earlier methods used hand-crafted features, such as dense trajectory features [4] and bag-of-words (BOW) histograms of motion tubelests [5] combined with traditional classification algorithms including Fisher kernels [4] and support vector machine (SVMs) [5] to detect actions in untrimmed videos. With new advances in deep learning, detecting action has become more effective. [6] suggested an algorithm to localize actions based on the maximum sum of frame-wise classification scores in different temporal segments that are processed through a deep convolutional neural network (CNN). [7] improved this solution by adding a recurrent mechanism that can better model the temporal dependencies in action frames. [8] proposed a more effective approach than [6], [7] using a long-short term transformer that can process longer videos without any bias against older temporal inputs.

There have been several approaches to detect activities in classroom videos. [9] suggested hand-crafted features including elbow angles and movements in the face and hands combined with primitive-based coupled hidden Markov model (PCHMM) to recognize seven teacher activities. [1] presented a more effective deep model with a multimodal attention layer to capture long-term semantic dependencies in instructional videos. [10] suggested that the skeleton pose is a more effective modality than RGB images used by others, as the skeleton pose more compactly represents teacher and students' actions in classrooms.

III. Methods

Fig. 1 shows the overall pipeline of our instructional action detection. Given a sequence of RGB frames, $I^R = \{I_t^r \in \mathbb{R}^{H\times W\times 3}, t = 0, 1, ..., T\}$, the goal is to find the action class scores, $\hat{Y}$, and the start of the end of action, Δ and E, respectively. Here, T is the size of the action sequence; and H and W are the height and width of the image, respectively. We first select the keyframes $I^K = \{I_t^k \in \mathbb{R}^{H\times W\times 3}, t = 0, 1, ..., T'\}$, which include important frames in the sequence. Here, T' is the number of keyframes. Next, the optical flow fields, $I^V = \{I_t^v \in \mathbb{R}^{H\times W\times 2}, t = 0, 1, ..., T'\}$, are extracted using a state-of-the-art optical flow estimation algorithm [11]. Using the motion enhancement algorithm, the optical flow fields are enhanced, making them insensitive to camera movements. The enhanced optical flow fields then are converted to optical flow

images $I^{K'} = \{I_t^{k'} \in \mathbb{R}^{H\times W\times 3}, t = 0, 1, ..., T'\}$ using a color-coding technique [12]. I^K and $I^{K'}$ are then converted to RGB features, $I^F = \{I_t^f \in \mathbb{R}^m, t = 0, 1, ..., T'\}$ and motion features $I^{F'} = \{I_t^{f'} \in \mathbb{R}^m, t = 0, 1, ..., T'\}$, using a pre-trained I3D network [13], a widely used model for action recognition. Here, m is the size of features. The enhanced baseline model with a revised loss function process I^F and $I^{F'}$ produces the action class prediction scores. A post-processing algorithm is also utilized to generate the start and end points of action instances.

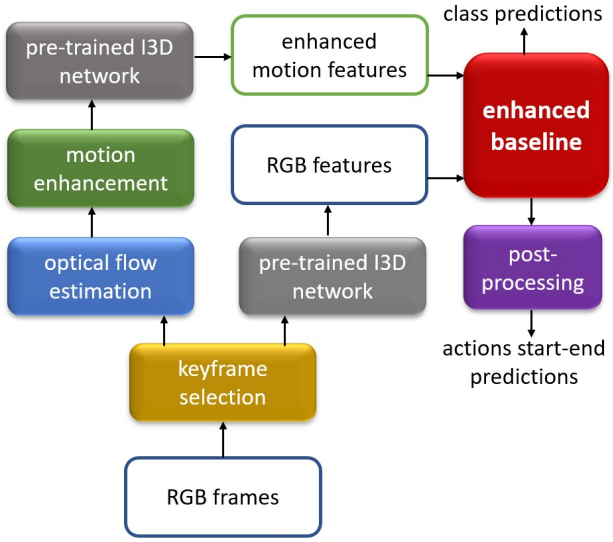

Fig. 1. The pipeline of the presented instructional activity detection algorithm.

A. Keyframe selection

The instructional videos are long and often the important events occur sparsely. So, a keyframe selection algorithm is presented to choose the important frames, making the proposed pipeline more efficient and effective in handling long videos. The keyframes are selected when there is a desirable change in videos that are evaluated by measuring the difference between two consecutive frames as follows:

$$\frac{1}{H \cdot W \cdot 3} \cdot \sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{z=1}^{3}(I_t^r(i,j,z) - I_{t-1}^r(i,j,z)) > T_r, \quad (1)$$

where T_r is a threshold value.

B. Motion enhancement

Camera movement is inevitable in videos captured from classrooms. Such camera movements distort the optical flow fields and reduce the quality of motion features. To solve this we utilize a motion enhancement algorithm. To do such, first Gaussian mixture models (GMMs) are used to model the background motion as $P(\lambda) = \sum_{k=1}^{K} \pi_k N(\lambda|\mu_k, \Sigma_k)$, where $N(\lambda|\mu_k, \Sigma_k)$ is a Gaussian density; μ_k, Σ_m, π_k, and K are mean. covariance, mixing coefficient, and the number of distributions, respectively. The background is modeled by optimizing the GMM parameters using maximum likelihood estimation [14]. With the assumption that the background is only affected by the camera movements, the camera-insensitive foreground then is recovered by subtracting from the corresponding background parts with respect to the Gaussian models.

C. Baselines

Three baselines are selected in this paper, including the background suppression network [15], multi-label action dependencies [16], and long short-term transformer [8]. All of these networks performed exceptionally well on the THUMOUS [17] and ActivityNet [18] datasets. Our instructional activity datasets share several characteristics with these two datasets, including (1) videos that are long in length; (2) videos are continuous streams, which means they are not segmented; (3) several instances of class labeling can occur simultaneously with co-occurring labels in the annotation data. The background suppression network produces weighted scores for background and foreground frames. Such a weighting strategy handles the crowded scenes in classroom videos that may include a significant amount of irrelevant information. The multi-label action dependencies model captures the multi-class dependencies between different action classes. This is useful in our experimental setup since multiple instructional activities may co-occur. The long short-term transformer can capture short and long-term dependencies in action videos. This will help to include the critical temporal dependencies regardless of their temporal distances.

D. Enhanced loss function

The instructional videos include both long and short action instances, such as "teacher sitting" and "student raising hand," respectively. To accommodate the videos with varying lengths, we suggest adding a new loss function to enhance the baseline models:

$$L = -\alpha \sum_{t=1}^{T}\sum_{c=1}^{C} y_t^{(c)} log\hat{y}_t^{(c)} - \beta \sum_{c=1}^{C} Y^{(c)} log\hat{Y}^{(c)}, \quad (2)$$

where y, $\hat{y}$, Y, and $\hat{Y}$ are the ground truth per frame, predicted values per action frame, ground truth per action sequence, and predicted values per action sequence, respectively. Moreover, c, C, α, and β are predicted class, the number of classes, and the loss adjustment parameters for frame-level, and loss adjustment parameters for sequence-level, respectively.

E. Post-processing

Many instructional activities such as "teaching sitting" are significantly longer than standard actions. So, they cannot be entirely processed within the standard deep learning models to regress the start and end frames. It is because these deep models have limited temporal receive fields. Therefore, we propose a post-processing algorithm to find the start and end frames of action instances after the frame-level prediction stage.

Given the action detection scores $\hat{Y} \in \mathbb{R}^{T\times C}$, the goal is to find the start and end frames of action instances, $\{\Delta, E\} = \{\delta_{n,c}, \epsilon_{n,c}, n = 0, 1..., N'; c = 0, 1, ...C\}$, where N' is the number of segmented class instances. Moreover, $\delta_{n,c}$ and $\epsilon_{n,c}$ are the start and end frame for class c of instance n.

Our post-processing algorithm consists of two phases of action scores thresholding and the start and end frames detection as shown in Algorithm 1. We used our post-processing algorithm to visualize the start and end of actions for teachers based on a developed teacher dashboard.

IV. Experimental Results

A. Implementation details

The size of RGB and motion features in our experiments for all baselines is 1024. For the background suppression model, six convolutional layers are used. The learning rate is $1e^{-5}$ which is decayed by 0.1, for every 1500 iterations. The multi-label dependencies network includes five layers. The temporal length is 132, and the initial learning rate is $1e^{-4}$. The long-short term transformer consists of four layers and 16 heads. Moreover, the learning rate is increased from zero to $5e^{-5}$ for half of the training iterations that took 50 epochs. 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.

B. Dataset.

An elementary school dataset was collected to analyze instructional activities. 250 hours of instructional activity videos were annotated by a team of nine professional annotators. Fig. 2 shows the labels for our 24 instructional activity classes. In our experiments, 80% of the data is used for training and 20% for testing.

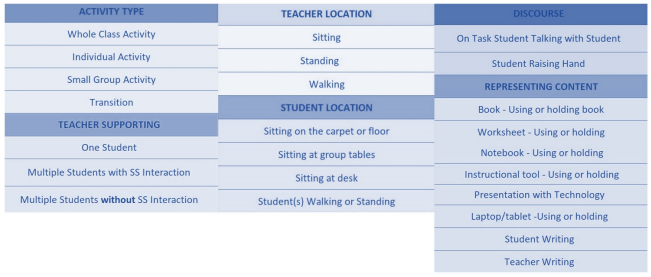

Fig. 2. The 24 instructional activity class labels in our annotated dataset.

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

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (3)$$

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

C. Comparative Results

Fig. 3 shows the comparative results of our pipeline based on three baseline models. The average per class performance for the background impression, multi-label dependencies, and long-short term transformer models are 0.47, 0.49, and 0.4, respectively. The average per sample performance for the background impression, multi-label dependencies, and long-short term transformer models are 0.52, 0.57, and 0.47, respectively

Algorithm 1 Post-processing

Require: Action detection scores, $\hat{Y} \in \mathbb{R}^{T \times C}$
Ensure: start and end frames of action instances, Δ, E

Phase 1 – action scores thresholding

$t = 0$
while $t \leq T$ **do** ▷ T is # of frames
  $c = 0$
  while $c \leq C$ **do** ▷ C is # of classes
    if $\hat{Y}_{t,c} \geq \theta$ **then** ▷ θ is the detection threshold
      $\hat{Y}_{t,c} = 1$
    else if $\hat{Y}_{t,c} < \theta$ **then**
      $\hat{Y}_{t,c} = 0$
    end if
    $c \leftarrow c + 1$
  end while
  $t \leftarrow t + 1$
end while

Phase 2 – Start and end frames detection

$Y^S = SPLIT(\hat{Y})$ ▷ splitting $\hat{Y}$ to N segments
$Y^S = \{y_n, n = 1, 2, ..., N\}$
$L \in \mathbb{R}^{N \times C} \leftarrow 0$ ▷ initializing labels for each segment
$n = 0$
while $n \leq N$ **do**
  $c = 0$
  while $c \leq C$ **do**
    if $\sum_{i=1}^{i=Q} y_{n,i,c} \geq Q/2$ **then** ▷ $Q = T/N$
      $L_{n,c} = 1$
    else if $\sum_{i=1}^{i=Q} y_{n,i,c} < Q/2$ **then**
      $L_{n,c} = 0$
    end if
    $n \leftarrow n + 1$
  end while
  $c \leftarrow c + 1$
end while
$L^M = MERGE(L)$ ▷ merging if $L_{n,c} = L_{n+1,c}$
$Y^M = RESPLIT(\hat{Y})$ ▷ re-splitting $\hat{Y}$ based on L^M
$n = 0$
while $n \leq N'$ **do** ▷ N' is # of segments for L^M
  $c = 0$
  while $c \leq C$ **do**
    $\delta_{n,i,c} = START(Y^M_{n,i,c})$ ▷ i: 1st time $Y^M_{n,i,c} = 1$
    $\epsilon_{n,j,c} = END(Y^M_{n,j,c})$ ▷ j: last time $Y^M_{n,j,c} = 1$
    Add $\delta_{n,i,c}$ to Δ ▷ adding the start frame
    Add $\epsilon_{n,j,c}$ to E ▷ adding the end frame
    $n \leftarrow n + 1$
  end while
  $c \leftarrow c + 1$
end while

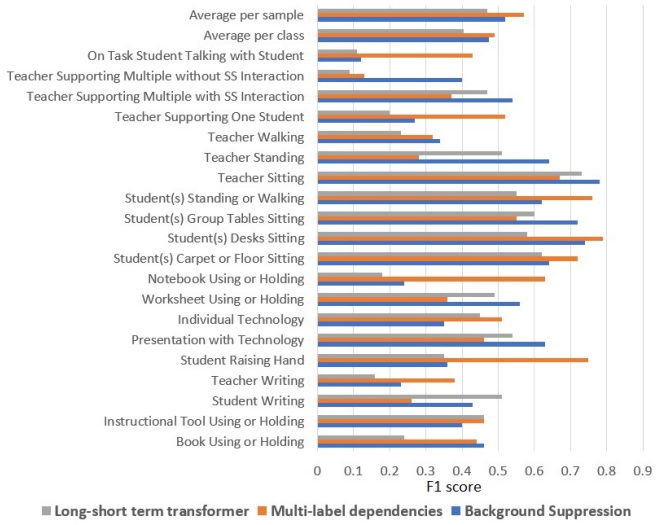

Fig. 3. Comparative results for three baselines and 21 instructional activity class labels.

D. Ablation study

Fig. 4 illustrates the impact of our proposed strategies, keyframe selection (KS), motion enhancement (MH), and enhanced loss (EL) function on the overall performances of two baseline models. Using our proposed strategies, the average F1 performance for background suppression model has a significant improvement of 0.2, from 0.37 to 0.57. Similarly, for the long-short term transformer, the F1 performance has been increased from 0.52 to 0.65.

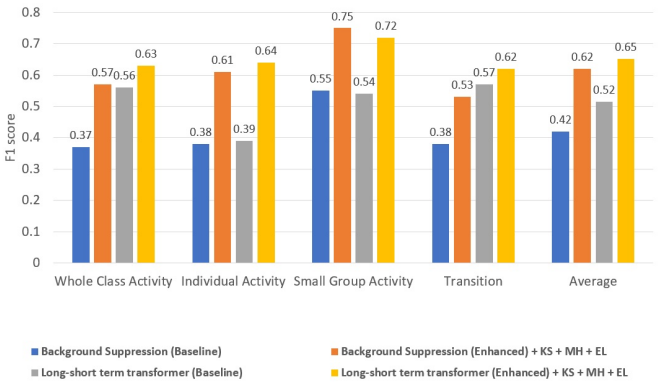

Fig. 4. The impact of the keyframe selection (KS), motion enhancement (MH), and enhanced loss (EL) function on the overall performances of baseline models on three activity types + transition class labels.

V. Conclusion

This paper proposes several strategies to improve the performance of multiple state-of-the-art action detection networks on instructional activity videos. The presented strategies mainly focus on improving the network in dealing with varying-size activity sequences and camera movements. Such a enhanced deep learning framework will facilitate teachers to receive feedback more effectively and efficiently than using manual observation. The experimental results have been promising when the enhanced deep models are evaluated on 250 hours of our annotated instructional videos.

References

[1] 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.

[2] H. Li, Z. Wang, J. Tang, W. Ding, and Z. Liu, "Siamese neural networks for class activity detection," in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*. Springer, 2020, pp. 162–167.

[3] P. J. Donnelly, N. Blanchard, B. Samei, A. M. Olney, X. Sun, B. Ward, S. Kelly, M. Nystrand, and S. K. D'Mello, "Multi-sensor modeling of teacher instructional segments in live classrooms," in *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016, pp. 177–184.

[4] D. Oneata, J. Verbeek, and C. Schmid, "Efficient action localization with approximately normalized fisher vectors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2545–2552.

[5] M. Jain, J. Van Gemert, H. Jégou, P. Bouthemy, and C. G. Snoek, "Action localization with tubelets from motion," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 740–747.

[6] Z. Yuan, J. C. Stroud, T. Lu, and J. Deng, "Temporal action localization by structured maximal sums," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3684–3692.

[7] M. Xu, M. Gao, Y.-T. Chen, L. S. Davis, and D. J. Crandall, "Temporal recurrent networks for online action detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5532–5541.

[8] M. Xu, Y. Xiong, H. Chen, X. Li, W. Xia, Z. Tu, and S. Soatto, "Long short-term transformer for online action detection," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[9] 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.

[10] 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.

[11] L. Kong, C. Shen, and J. Yang, "Fastflownet: A lightweight network for fast optical flow estimation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 10 310–10 316.

[12] K. Jia, X. Wang, and X. Tang, "Optical flow estimation using learned sparse model," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2391–2398.

[13] 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.

[14] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[15] P. Lee, Y. Uh, and H. Byun, "Background suppression network for weakly-supervised temporal action localization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 320–11 327.

[16] P. Tirupattur, K. Duarte, Y. S. Rawat, and M. Shah, "Modeling multi-label action dependencies for temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1460–1470.

[17] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The thumos challenge on action recognition for videos "in the wild"," *Computer Vision and Image Understanding*, vol. 155, pp. 1–23, 2017.

[18] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *2015 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2015, pp. 961–970.