CSE 4120 : Technical Writing & Seminar

Title: Online Action Detection

Presented To:

Dr. K. M. Azharul Hasan

Professor
Department of Computer Science and Engineering
Khulna University of Engineering and Technology

Sunanda Das

Assistant Professor
Department of Computer Science and Engineering
Khulna University of Engineering and Technology

Presented By:

Md. Raduan Islam Rian

Roll: 1907117

Department of Computer Science and Engineering

Khulna University of Engineering & Technology

Outline

- ✓ Introduction
- ✓ Publication details
- ✓ Related Work
- ✓ Methodology
- ✓ Experimental Studies
- ✓ Comparison
- ✓ Limitations
- ✓ Conclusion
- ✓ References

Introduction

Problem:
Detecting actions in real-time from untrimmed video streams presents significant challenges:
Complexity of Untrimmed Videos: Untrimmed videos contain multiple action instances mixed with background scenes, making it hard to isolate and identify individual actions accurately.
Limited Information in Online Settings: Online action detection systems rely only on past and current data, lacking access to future observations essential for predicting upcoming actions.
Variability in Human Actions: Human actions exhibit diverse characteristics within the same class, making it difficult to rely on traditional methods based on average action durations.

Introduction (Cont.)

Solution:
Our novel framework tackles these challenges effectively:

Action Subclass Modeling: By representing actions as temporally ordered subclasses, the distinct phases of each action class directly from visual input.

Future Frame Generation: Utilizing a future frame generation network, predicted future frames to compensate for the lack of future information in online settings, enabling accurate anticipation of forthcoming actions.

Data Augmentation: The model's robustness by varying the lengths of training videos, exposing it to a diverse range of temporal characteristics in human actions.

Publication Details

Table 1: Publication Details

Serial No	Title	Author	Source	Published Year
1	E2E-LOAD: End-to-End Long-form Online Action Detection	Shuqiang Cao, Bairui Wang, Wei Zhang, Lin Ma, Weixin Luo	ICCV (The International Conference on Computer Vision)	2023
2	Learning to Discriminate Information for Online Action Detection	Hyunjun Eun , Jinyoung Moon, Jongyoul Park, Chanho Jung, Changick Kim	CVPR (Conference on Computer Vision and Pattern Recognition)	2020
3	A novel online action detection framework from untrimmed video streams	Da-Hye Yoon, Nam-Gyu Cho, and Seong-Whan Lee.	ElSEVIER, Pattern Recognition Journal, VOL. 106	2020

Related Work

Paper 1: E2E-LOAD: End-to-End Long-form Online Action Detection

1.Traditional methods struggle with processing long videos in **real-time** due to computational complexity.

2. **E2E-LOAD** proposes an **end-to-end framework** for online action detection that efficiently processes long-form videos.

3. It utilizes a combination of deep learning architectures, including convolutional and recurrent networks, to capture temporal information and model action sequences.

4. By processing videos in a streaming manner, **E2E-LOAD** enables real-time action detection in long-form videos.

Related Work (Cont.)

Paper 2: Learning to Discriminate Information for Online Action Detection

1. Recent advancements in action detection have seen a shift towards addressing online scenarios, facilitated by datasets like **RGB-based TVSeries** and **skeleton-based OAD**. Baseline models, including **LSTM networks and CNNs**, aim to predict action endings using temporal priors or unsupervised learning approaches.

2. However, these methods often struggle with limited temporal context, impacting their performance in online settings. Newer approaches focus on discriminating relevant information, leveraging discriminative learning techniques to enhance action detection accuracy and mitigate the challenges posed by limited information availability and **high intra-class variation** in action lengths.

Related Work (Cont.)

Paper 3: A Novel Online Action Detection Framework from Untrimmed Video Streams.

1. Deep learning has revolutionized action detection, with recent focus on untrimmed video datasets like **THUMOS'14** and **ActivityNet**. Methods such as improved dense trajectories **(IDT)** encoded by fisher vectors **(FVs) and CNNs** with sliding windows have shown efficacy

2. Advanced techniques like **multi-stage CNNs**, convolution-de-convolutional **(CDC)** networks, and pyramid of score distribution features **(PSDF)** descriptors enhance temporal context capture. Spatial considerations, including **spatio-temporal** action localization methods, and fine-grained action detection approaches have also emerged.

Methodology

Table 2: Model Comparison

Papers	**E2E-LOAD: End-to-End Long-form Online Action Detection**	**Learning to Discriminate Information for Online Action Detection**	**A Novel Online Action Detection Framework from Untrimmed Video Streams**
Dataset Selection	Utilized established untrimmed video datasets like THUMOS'14 and ActivityNet	necessitating tailored preprocessing steps to handle real-time data streams effectively.	suitable for long-form online action detection, involving preprocessing steps optimized for continuous video analysis.
Model Design	Employed multi-stage CNN architectures trained using supervised learning techniques	Developed specialized baseline models such as long short-term memory (LSTM) networks and CNNs.	Designed an E2E framework integrating convolutional and recurrent networks, capture temporal dependencies effectively.

Methodology (Cont.)

Table 2: Model Comparison

Papers	**E2E-LOAD: End-to-End Long-form Online Action Detection**	**Learning to Discriminate Information for Online Action Detection**	**A Novel Online Action Detection Framework from Untrimmed Video Streams**
Feature Extraction & Representation	Utilized improved dense trajectories (IDT) encoded by fisher vectors (FVs) and CNNs with sliding windows, followed by encoding techniques.	Explored various feature extraction methods and prediction strategies to enhance the accuracy and efficiency of real-time action detection.	Emphasizing the importance of temporal context preservation.
Evaluation Metrics and Benchmarking	Evaluated model performance using standard metrics such as mean Average Precision (mAP) on benchmark datasets	Introduced new evaluation metrics to assess the performance in dynamic environments.	Conducted comprehensive evaluations on long-form video datasets for end-to-end online action detection systems.

Experimental Studies

Table 3: Result Analysis

Papers	**E2E-LOAD: End-to-End Long-form Online Action Detection**	**Learning to Discriminate Information for Online Action Detection**	**A Novel Online Action Detection Framework from Untrimmed Video Streams**
Real-world performance	Real-world performance evaluation demonstrated the ability of E2E-LOAD to efficiently process long videos and detect actions in real-time.	Results demonstrated the feasibility of online action detection with baseline models such as LSTM networks and CNNs, albeit with some limitations.	Real-world performance was demonstrated through the ability to process untrimmed video streams and detect actions in real-time.
Accuracy evaluation	Accuracy analysis showcased the effectiveness of the proposed method, in capturing temporal information and accurately modeling action.	Accuracy analysis highlighted the importance of temporal priors and unsupervised learning approaches in improving action detection	Accuracy analysis showed improvements in action detection accuracy compared to traditional methods, handling long video sequences.

Comparison

Table 4: Comparative discussion

Papers	**E2E-LOAD: End-to-End Long-form Online Action Detection**	**Learning to Discriminate Information for Online Action Detection**	**A Novel Online Action Detection Framework from Untrimmed Video Streams**
Focus	Real-time action detection in long-form videos	Adapting offline methods for online action detection	Online action detection from untrimmed video streams
Dataset	Not specified	RGB-based TVSeries, skeleton-based OAD	THUMOS'14, ActivityNet
Methodology	End-to-end framework with convolutional and recurrent networks	LSTM networks, CNNs	Utilizes improved dense trajectories, CNNs
Contribution	Real-time processing.	Feasibility of online action detection with baseline models	Competitive results on benchmark datasets
Performance	Real-time action detection with efficient processing of long videos	Demonstrated feasibility with baseline models	Achieved competitive results on benchmark datasets

Limitations

Table 5: Limitations

E2E-LOAD: End-to-End Long-form Online Action Detection	Learning to Discriminate Information for Online Action Detection	A Novel Online Action Detection Framework from Untrimmed Video Streams
1. Lack of comprehensive evaluation on a wider range of datasets beyond THUMOS'14 and ActivityNet. 2. Limited analysis on the scalability of the proposed framework to handle diverse real-world scenarios.	1. Reliance on existing datasets and lack of exploration of custom datasets for online action detection. 2. Limited discussion on the generalization of the proposed methods across different domains.	1. Relatively limited exploration of alternative architectures beyond CNNs and CDC networks. 2. Lack of in-depth analysis on the impact of hyper-parameters and network configurations. 3. Limited discussion on real-world deployment challenges.

Conclusion

In conclusion, the three papers present innovative approaches to online action detection, showcasing advancements in model architectures, dataset utilization, and real-time processing capabilities. While each framework demonstrates promising results and contributes valuable insights to the field, there exist several avenues for further exploration and improvement. These include the investigation of alternative architectures, extensive evaluation on diverse datasets, analysis of hyperparameters impact, and scalability assessment to real-world deployment scenarios. Overall, these papers collectively push the boundaries of online action detection and pave the way for future research endeavors in this domain.

References

1. Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. "Vivit: A video vision transformer." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6836–6846, 2021.

2. Gedas Bertasius, Heng Wang, and Lorenzo Torresani. "Is space-time attention all you need for video understanding?" In ICML, volume 2, page 4, 2021.

3. Junwen Chen, Gaurav Mittal, Ye Yu, Yu Kong, and Mei Chen. "Gatehub: Gated history unit with background suppression for online action detection." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19925–19934, 2022.

4. Y. Cai, H. Li, J.-F. Hu, and W.-S. Zheng. "Action knowledge transfer for action prediction with partial videos." In Proc. Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence, pages 8118–8125, Jan. 2019. [1]

5. J. Carreira and A. Zisserman. "Quo vadis, action recognition? A new model and the kinetics dataset." In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724–4733, Jul. 2017. [2]

References (Cont.)

3. Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar. "Rethinking the Faster R-CNN architecture for temporal action localization." In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1130–1139, Jun. 2018. [3]

15. H. Maeng, S. Liao, D. Kang, S.-W. Lee, A.K. Jain, "Nighttime face recognition at long distance: cross-distance and cross-spectral matching," in: Asian Conference on Computer Vision, Springer, 2012, pp. 708–721.

16. U. Park, H.-C. Choi, A.K. Jain, S.-W. Lee, "Face tracking and recognition at a distance: a coaxial and concentric PTZ camera system," IEEE Trans. Inf. Forensics Secur. 8 (10) (2013) 1665–1677.

17. R. Poppe, "A survey on vision-based human action recognition," Image Vis. Comput. 28 (6) (2010) 976–990.

Thank You