IRSTI 28.23.37 https://doi.org/10.26577/jpcsit2024-v2-i4-a3

N. Azatbekuly* , **B. Matkerim** , **A. Mukhanbet**

Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: nurtugang17@gmail.com

ANALYSIS OF SPATIO-TEMPORAL CONVOLUTIONAL NEURAL NETWORKS FOR THE ACTION DETECTION TASKS

Abstract. This study investigates the effectiveness of Spatio-Temporal Convolutional Neural Networks (ST-CNNs) for action detection tasks, with a comprehensive comparison of state-of-the-art models including You Only Watch Once (YOWO), YOWOv2, YOWO-Frame, and YOWO-Plus. Through extensive experiments conducted on benchmark datasets such as UCF-101, HMDB-51, and AVA, we evaluate these architectures using metrics like frame-based Mean Average Precision (frame-mAP), video-mAP, computational efficiency (FPS), and scalability. The experiments also include real-time testing of the YOWO family using an IP camera and RTSP protocol to assess their practical applicability. Results highlight the superior accuracy of YOWO-Plus in capturing complex spatio-temporal dynamics, albeit at the cost of processing speed, and the efficiency of YOWO-Frame for live applications. This analysis underscores the trade-offs between speed and accuracy inherent in single-stage ST-CNN architectures. Our findings from the comparative analysis provide a robust foundation for the development of real-time systems capable of efficient and reliable operation in action detection tasks.

Key words: action detection, convolutional neural networks, spatio-temporal convolutional neural networks, YOWO.

1. Introduction

Recent success in solving problems of classification or localization of objects in an image is due to the long-term development of two-dimensional spatial convolutional neural networks, over the years two-dimensional CNNs have achieved excellent results. The application of such a deep neural network to recognize complex images, such as actions, has been done in the following works [1], [2]. But, as the name implies, such a deep neural network architecture is mainly used to find spatial features in data, which means that in most cases they can identify some patterns only judging by a single image. But to identify complex features in the data, such an architecture is often not enough to solve the problem of detecting the effect. To do this, spatiotemporal convolutional neural networks are used to process a more complex data structure, which is four-dimensional data, where in addition to the channel, height and width of the image, there is also a fourth dimension – the frame. Three-dimensional convolutional neural networks are used for convolution for such a complex data structure. They make it possible to identify spatio-temporal characteristics in the data for further detection of actions [3-6].

Existing architectures using 3D convolution for the action detection tasks, should be classified according to the expected environment used:

1. Causality-driven architecture, which relies only on the current or previous frames and does not depend on future frames for its operation (allowing for online processing) [7], [8].

2. Predictive(anticipatory) architecture. This type of architecture incorporates future frames alongside current and past ones to make decisions or predictions. So, that's why we can't use them for online processing [9], [10].

Based on their approach to data processing and analysis, architectures for detecting actions can be divided into two varieties:

1. Single-stage Architectures. Single-stage architectures are aimed at detecting actions directly from video frames without the preliminary stage of selecting candidate areas. The only single-stage architectures for detecting actions are the You Only Watch Once (YOWO) family of architectures. They work by directly predicting classes of actions and their localizations in one pass, which ensures high processing speed.

2. Two-Stage Architectures. Unlike single-stage architectures, they work in two stages: first they

generate region proposals that may contain actions, and then classify these proposed areas and clarify their localization. Examples of such system is [1], a two-dimensional action recognition model for surveillance systems. Another example is [11], which is a tracking-based two-stage framework for spatio-temporal action detection. Although these models can achieve high accuracy through a more thorough analysis of each proposed area, they face several challenges. The problem with such an architecture is low speed, since several box proposals are made for each frame, considering the time coupling in 3D CNN-s, as well as the presence of a local optimum in the presence of several CNN blocks (many modern architectures take advantage of both 2D and 3D CNN), which leads to difficulties in optimizing the architecture.

While the field of Spatio-Temporal Convolutional Neural Networks (ST-CNN) has seen some advancements, comprehensive analysis remains relatively sparse [12], [13]. Among the notable contributions, a Spatio-Temporal Attention (STA) network that innovatively addresses the limitations of traditional 3D CNNs by differentiating the importance of video frames both spatially and temporally, significantly enhancing action recognition and detection in complex video sequences was introduced in [14]. This STA network demonstrates state-of-the-art performance on benchmark datasets such as UCF-101 and HMDB-51. On the other hand, the authors in [15] propose the Spatio-Temporal Progressive (STEP) action detector, a novel progressive learning framework that refines spatio-temporal proposals over multiple steps, effectively adapting to the dynamic nature of actions in videos and achieving impressive results on UCF101 and AVA datasets. The work in [16] explores a lightweight action recognition architecture that synergizes CNN, LSTM units, and a temporal-wise attention model to efficiently process RGB data for human action recognition, showcasing the potential of combining spatial and temporal analysis in a cohesive framework. Despite these innovative approaches, the exploration of ST-CNNs, particularly in terms of in-depth analysis and optimization for diverse applications, remains an area ripe for further research and development.

In this work, we will focus on analyzing the family of fast and advanced You Only Watch Once (YOWO) architectures in the context of testing on your own computer and in various datasets: HMDB-51, AVA, UCF-101.

The primary contribution of this study lies in our comprehensive analysis of the most advanced and efficient architectures for spatio-temporal convolutional neural networks (STCNNs), with a particular focus on enhancing the detection of actions in video streams. Our work meticulously evaluates both the performance and processing speed of leading-edge models, thereby shedding light on their potential to significantly advance the field of video analysis.

2. Materials and Methods

In this section, we describe the methods and datasets employed in our study to evaluate the performance of ST-CNNs for action detection. The methodology centers around the detailed analysis of the YOWO family of architectures, which are renowned for their capability to process video streams efficiently in real-time. We also outline the datasets used for benchmarking the models, providing insights into their structure, diversity, and applicability for various action recognition tasks. By combining innovative architectural enhancements with comprehensive dataset evaluations, this study aims to establish a robust framework for advancing the field of real-time video analysis.

The following subsections provide an in-depth discussion of the YOWO architecture and its variations, followed by a detailed description of the datasets utilized in our experiments.

2.1. Spatio-temporal convolutional neural networks

The You Only Watch Once (YOWO) architecture, described in Figure 1, represents an innovative approach to detecting actions in video streams, effectively combining the advantages of both 2D and 3D convolution neural networks for video data analysis. YOWO's work is based on extracting spatial features from the current frame using 2D-CNN, while 3D-CNN is used to model spatiotemporal features from a sequence of previous frames. This dual strategy allows the model not only to recognize objects and their configurations within the frame, but also to understand the dynamics of their changes over time, which is critical for accurate detection of actions.

To improve the performance and efficiency of feature processing, YOWO integrates advanced technologies such as Darknet-19 and ResNext-101, which serve as a powerful basis for the primary processing of visual information. The key point in the YOWO architecture is the use of a channel-wise at-

tention mechanism, which allows the model to focus on the most significant features, ignoring less important information. This is achieved by effectively combining features, where important channels are amplified and unimportant ones are suppressed, thereby improving the quality of final predictions.

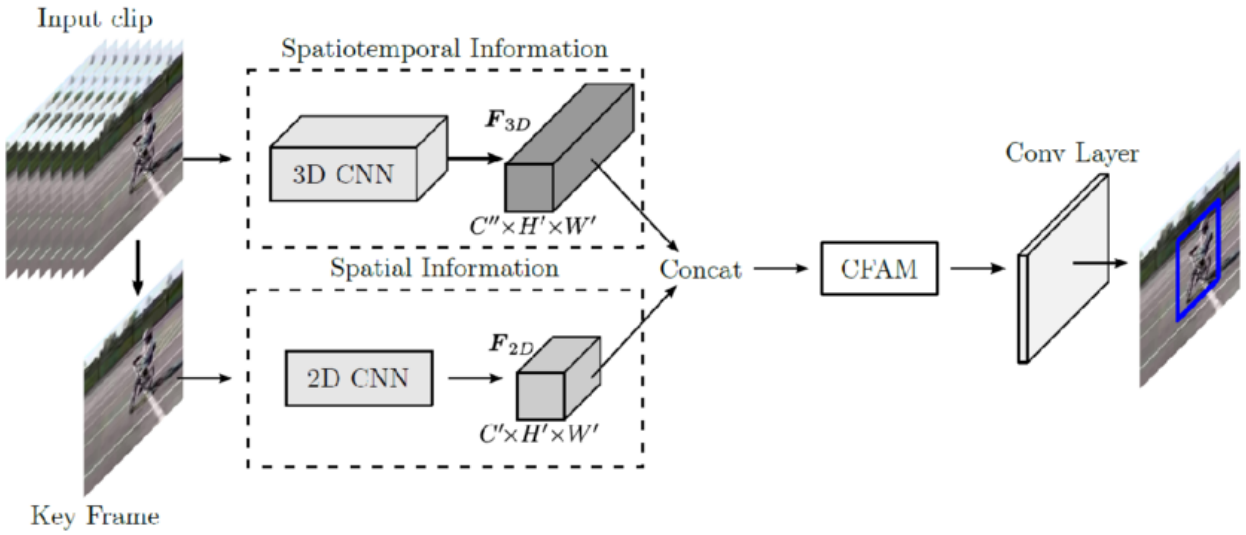

Figure 1 – YOWO architecture

In addition, the architecture uses the channel fusion technique, described in Figure 2, which provides a rough concatenation of features from different networks, despite their different number of channels. This is complemented by a small preprocessing using simple CNNs, which improves integration and interaction between heterogeneous data. An important feature of YOWO is also the use of Gram matrices to display inter-channel dependencies, which allows the model to more accurately determine which channels contribute to the overall information and how they interact with each other.

The use of the attention map and the learnable scalar parameter alpha further enhances the model's ability to adapt and integrate original and transformed features, providing a deeper understanding of video content. This combination of technologies and techniques makes the YOWO architecture an extremely powerful tool for detecting actions in real time, offering significant improvements in accuracy and speed compared to traditional approaches.

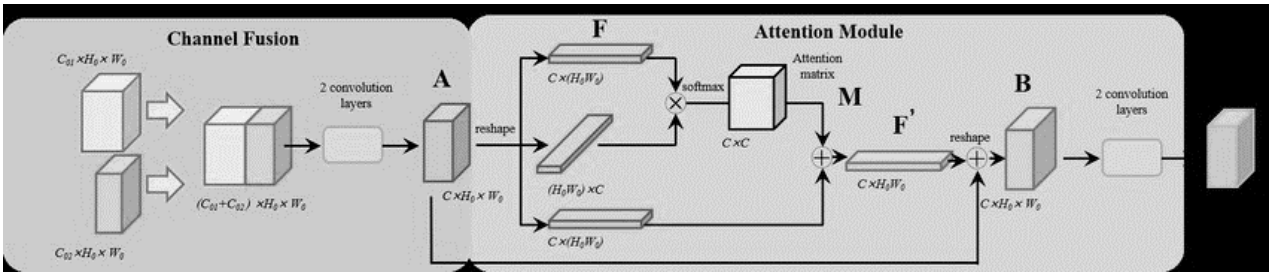

Figure 2 – CFAM mechanism

The further development of such an integrated approach combining the advantages of two-dimensional and three-dimensional convolutional neural networks to identify feature map and the use of Attention mechanisms that implement the selection of the most important features led to the development of the YOWO architecture and led to the emergence of YOWOv2, YOWO-Plus and YOWOF.

The YOWOv2 is a significant improvement over the previous YOWO model, demonstrating superior accuracy and processing speed due to several key architectural innovations. YOWOv2 implements a multi-level detection scheme that allows you to determine actions of various scales. This is achieved by implementing a simple and efficient 2D framework with the feature pyramid network, which improves the extraction of various levels of classification and regression features. As a result, the YOWOv2 family was created, including the YOWOv2-Tiny, YOWOv2-Medium and YOWOv2-Large variants, which provides flexibility in applying the model on platforms with different computing power.

YOWO-Plus is an advanced version of the original YOWO architecture, which has been optimized by YOWO from three points of view: backbone, label assignment, and loss function. One of the key innovations was the use of improved pre-trained weights for the YOLOv2 component, which were retrained to achieve better performance compared to the official version of YOLOv2. This made it possible to significantly increase the accuracy of detecting actions. In addition, the tag assignment system was optimized, which helped to improve the accuracy of recognizing actions. The introduction of IoU losses for box regression was another significant step towards more accurate definition of object boundaries, which, together with previous improvements, led to a noticeable increase in the overall performance of the YOWO-Plus system.

The YOWOF architecture unlike previous versions, YOWOF uses multiple encoders to process each frame of the video separately. This helps to identify various spatial objects, such as textures and shapes, that may indicate actions or activities. Next, a temporal encoder is introduced, which consists of several levels of convolutional LSTM modules (ConvLSTM). These modules help to capture time dependencies in video data by analyzing sequences of spatial objects from frames. ConvLSTM layers sequentially process the video stream, determining which actions occur in time. Thus, YOWOF combines spatial and temporal features for more accurate detection of actions in real-time video. This architecture sacrifices some of the temporal context that could be obtained by analyzing multiple frames in favor of faster processing time and reduced computational load.

2.2. Datasets

For action detection tasks, one of the following three datasets is usually used: UCF101, HMDB-51, AVA.

UCF-101 is one of the most well–known datasets for video analysis and action recognition. It contains 101 categories of activities that include a wide range of activities such as sports activities, musical instruments and daily activities, see Figure 0. The dataset consists of 13,320 video clips collected from YouTube, which makes it diverse and representative for training and testing action recognition algorithms. The video clips in UCF-101 represent real scenes with variable lighting conditions, backgrounds and points of view, which makes the dataset difficult for video processing algorithms.

HMDB-51 is a dataset consisting of 6,766 clips divided into 51 categories of actions. These activities cover various aspects of human activity, including gestures, physical activity, and human interaction with objects, see Figure 0. The dataset was compiled from various sources, including archival materials, films and publicly available video materials, which ensures its high variability and complexity. HMDB-51 is designed to test and evaluate algorithms for recognizing actions in the real world, considering the complexity and diversity of human actions.

AVA (Atomic Visual Actions) is a relatively new dataset designed for more detailed analysis of actions in video. It includes annotations for over 80,000 video clips extracted from films and describes actions at the level of individual frames with precise timestamps 4 and spatial localizations, see Figure 0. AVA focuses on “atomic” actions, i.e. basic, noncomposite actions that can be precisely defined and classified. This makes it especially useful for developing and testing algorithms capable of recognizing and interpreting complex, composite actions in a video.

Tests were performed on these three datasets, which are described in the next section.

2. Results and discussion

This section analyzes the performance of single-stage spatio-temporal CNNs, particularly the YOWO family, using FPS, frame-mAP, and video-mAP metrics. For the test, CUDA was used to increase computing power, with an NVIDIA GeForce GTX 1660 SUPER graphics card.

For the AVA dataset, our results indicate that the YOWO-Plus architecture achieves remarkable precision. Figure 3 summarizes the performance of YOWO-Plus, highlighting its frame-mAP, video-mAP, and FPS.

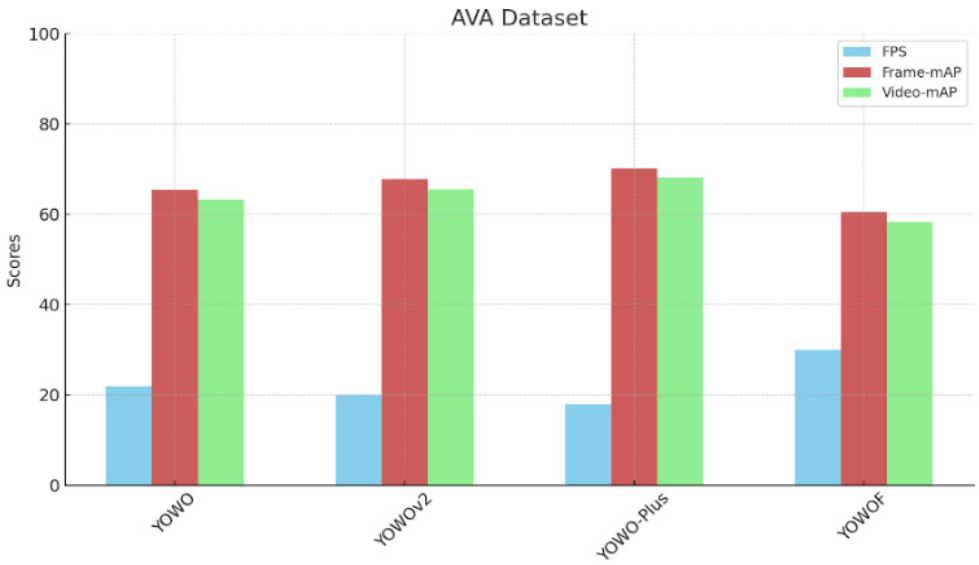

Figure 3 – Comparison of FPS, frame-mAP, video-mAP of the YOWO family of algorithms on the AVA dataset

Turning our attention to the HMDB-51 dataset, we observe a similar trend of performance improvement with YOWO-Plus. The algorithm demonstrates its robustness across a variety of action scenarios, maintaining high precision while ensuring real-time processing. The corresponding diagram in the Figure 4, details the algorithm's performance, highlighting its superior frame-mAP and video-mAP compared to other single-stage architectures.

Furthermore, when evaluating the UCF-101 dataset, the YOWO-Plus and YOWOFrame models show a notable enhancement in processing speed and accuracy. The YOWOFrame stands out for its exceptional speed, operating at a frame rate significantly faster than other models without a considerable sacrifice in accuracy. The results for this dataset are summarized in a Figure 5.

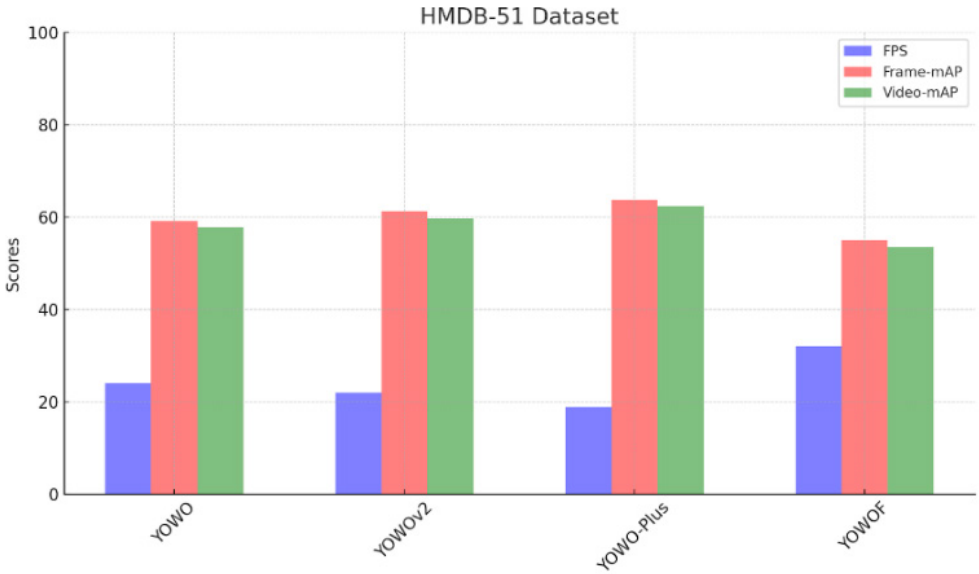

Figure 4 – Comparison of FPS, frame-mAP, video-mAP of the YOWO family of algorithms on the HMDB-51 dataset

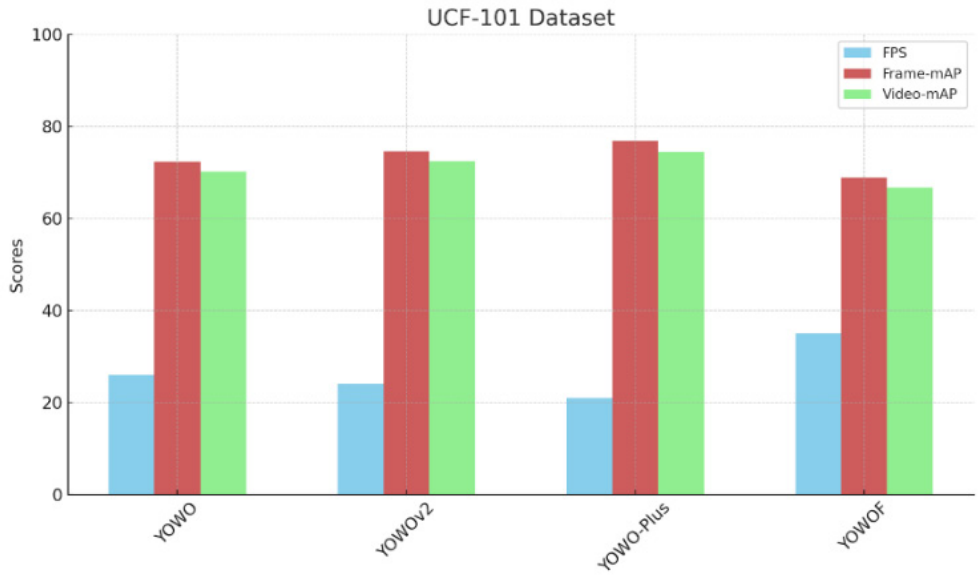

Figure 5 – Comparison of FPS, frame-mAP, video-mAP of the YOWO family of algorithms on the UCF-101 dataset

Tables 1–3 reveal a trade-off in the YOWO family between speed and accuracy. YOWO-Plus offers the best accuracy but slower FPS, while YOWO-Frame excels in real-time detection. Future work could focus on improving this balance and expanding tests to diverse datasets.

Table 1 – The results of testing YOWO algorithms on the AVA dataset

Model	FPS	Frame-mAP	Video-mAP
YOWO	22	65.4%	63.2%
YOWOv2	20	67.8%	63.5%
YOWO-Plus	18	70.2%	68.1%
YOWOF	30	60.5%	58.3%

Table 2 – The results of testing YOWO algorithms on the HMDB-51 dataset

Model	FPS	Frame-mAP	Video-mAP
YOWO	24	59.1%	57.8%
YOWOv2	22	61.3%	59.7%
YOWO-Plus	19	63.7%	62.4%
YOWOF	32	55.0%	53.5%

Table 3 – The results of testing YOWO algorithms on the UCF-101 dataset

Model	FPS	Frame-mAP	Video-mAP
YOWO	26	72.3%	70.1%
YOWOv2	24	74.6%	72.4%
YOWO-Plus	21	76.8%	74.5%
YOWOF	35	68.9%	66.7%

During our research into the practical application of the YOWO algorithm, we conducted a live demonstration to test its real-time action detection capabilities. Using an IP camera, we broadcast video footage in real time over the real-time streaming Protocol (RTSP). The video stream was captured and processed using the OpenCV library. Once the video stream was captured, it was transferred to a pre-trained YOWO model running on a local computer. The model processed incoming frames in real time, detecting and classifying actions as they occurred. The output of the YOWO model included both action labels and their corresponding bounding boxes inside video frames, demonstrating its ability to identify and localize multiple actions simultaneously. Results are shown in Figure 6.

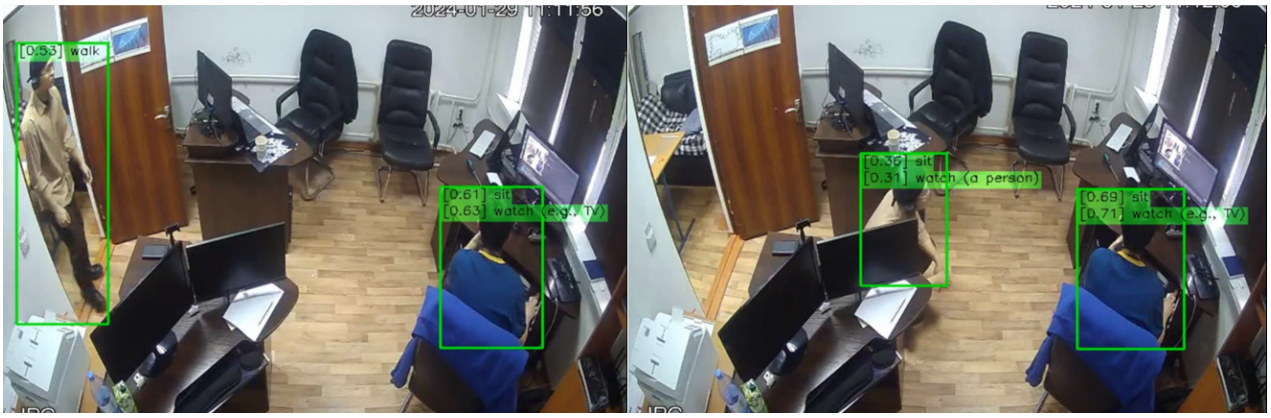

Figure 6 – Demonstration of the inference of the YOWO

In summary, this analysis provides practical insights for selecting suitable models in real-time video analysis, advancing the field of action detection technology.

3. Conclusions

In conclusion, our study in the field of spatio-temporal convolutional neural networks using the YOWO family of architectures, provided a key insight into action detection. This study aimed to demonstrate the various trade-offs associated with the four distinct YOWO architectures when applied to action detection tasks. Our careful assessments across diverse datasets such as AVA, HMDB-51, and UCF-101 have demonstrated the robustness and adaptability of the YOWO framework, particularly spotlighting the advanced performance of the YOWO-Plus and YOWO-Frame variations. We used 3 metrics to benchmark these architectures, namely FPS, frame-mAP and video-mAP.

YOWO-Plus emerged as the top performer across all evaluated datasets, albeit with the lowest frames per second (FPS), indicating a trade-off between accuracy and speed. In contrast, YOWO-Frame delivered quicker performance, achieving relatively high scores in both frame-mAP and video-mAP metrics, proving to be a more suitable option for live action-detection tasks. These results significantly contribute to the field of video analysis, providing a robust framework for real-time action detection and laying the groundwork for future technological developments. Our empirical evaluations, including live tests conducted via an IP camera with RTSP protocol support facilitated by the OpenCV library, validate the practical applicability of YOWO models for real-time deployment. The proficiency of YOWO models in processing live video feeds efficiently, coupled with their notable accuracy in action detection, stresses their potential for an array of real-world applications, from surveillance systems to interactive media.

Acknowledgments

The authors would like to thank all individuals and institutions who provided support and assistance throughout the research process.

Funding

This research received no external funding.

Author Contributions

Conceptualization, B.M. and A.M.; Methodology, B.M.; Software, N.A.; Validation, B.M. and A.M.; Formal Analysis, A.M.; Investigation, B.M. and A.M.; Resources, B.M.; Data Curation, N.A.; Writing – Original Draft Preparation, N.A.; Writing – Review & Editing, B.M. and A.M.; Visualization, N.A.; Supervision, B.M.; Project Administration, B.M.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. V.-D. Hoang, D.-H. Hoang, and C.-L. Hieu, 'Action Recognition Based on Sequential 2D-CNN for Surveillance Systems', in *IECON 2018 – 44th Annual Conference of the IEEE Industrial Electronics Society*, Oct. 2018, pp. 3225–3230. doi: 10.1109/IECON.2018.8591338.

2. I. T. Toudjeu and J.-R. Tapamo, 'A 2D Convolutional Neural Network Approach for Human Action Recognition', in *2019 IEEE AFRICON*, Sep. 2019, pp. 1–5. doi: 10.1109/AFRICON46755.2019.9133840.

3. J. Arunnehru, G. Chamundeeswari, and S. P. Bharathi, 'Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos', *Procedia Comput. Sci.*, vol. 133, pp. 471–477, Jan. 2018, doi: 10.1016/j.procs.2018.07.059.

4. R. Hou, C. Chen, and M. Shah, 'An End-to-end 3D Convolutional Neural Network for Action Detection and Segmentation in Videos', Nov. 30, 2017, *arXiv*: arXiv:1712.01111. doi: 10.48550/arXiv.1712.01111.

5. L. Yang, I. O. Ertugrul, J. F. Cohn, Z. Hammal, D. Jiang, and H. Sahli, 'FACS3D-Net: 3D Convolution based Spatiotemporal Representation for Action Unit Detection', in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sep. 2019, pp. 538–544. doi: 10.1109/ACII.2019.8925514.

6. S. Ji, W. Xu, M. Yang, and K. Yu, '3D Convolutional Neural Networks for Human Action Recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

7. K. Xu, F. Ye, Q. Zhong, and D. Xie, 'Topology-aware Convolutional Neural Network for Efficient Skeleton-based Action Recognition', Dec. 09, 2021, *arXiv*: arXiv:2112.04178. doi: 10.48550/arXiv.2112.04178.

8. J. Tan, X. Zhao, X. Shi, B. Kang, and L. Wang, 'PointTAD: Multi-Label Temporal Action Detection with Learnable Query Points', Mar. 21, 2023, *arXiv*: arXiv:2210.11035. doi: 10.48550/arXiv.2210.11035.

9. J. R. Medel and A. Savakis, 'Anomaly Detection in Video Using Predictive Convolutional Long Short-Term Memory Networks', Dec. 15, 2016, *arXiv*: arXiv:1612.00390. doi: 10.48550/arXiv.1612.00390.

10. H. Zhao and R. P. Wildes, 'Review of Video Predictive Understanding: Early Action Recognition and Future Action Prediction', Jul. 16, 2021, *arXiv*: arXiv:2107.05140. doi: 10.48550/arXiv.2107.05140.

11. J. Luo *et al.*, 'A Tracking-Based Two-Stage Framework for Spatio-Temporal Action Detection', *Electronics*, vol. 13, no. 3, p. 479, Jan. 2024, doi: 10.3390/electronics13030479.

12. J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, 'Spatio-Temporal Attention Networks for Action Recognition and Detection', *IEEE Trans. Multimed.*, vol. 22, no. 11, pp. 2990–3001, Nov. 2020, doi: 10.1109/TMM.2020.2965434.

13. A. Ramaswamy, K. Seemakurthy, J. Gubbi, and B. Purushothaman, 'Spatio-temporal action detection and localization using a hierarchical LSTM', in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2020, pp. 3303–3312. doi: 10.1109/CVPRW50498.2020.00390.

14. Z. Yang, J. Gao, and R. Nevatia, 'Spatio-Temporal Action Detection with Cascade Proposal and Location Anticipation', Jul. 31, 2017, *arXiv*: arXiv:1708.00042. doi: 10.48550/arXiv.1708.00042.

15. X. Yang, X. Yang, M.-Y. Liu, F. Xiao, L. S. Davis, and J. Kautz, 'STEP: Spatio-Temporal Progressive Learning for Video Action Detection', in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 264–272. doi: 10.1109/CVPR.2019.00035.

16. L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, 'Human Action Recognition by Learning Spatio-Temporal Features With Deep Neural Networks', *IEEE Access*, vol. 6, pp. 17913–17922, 2018, doi: 10.1109/ACCESS.2018.2817253.

Information About Authors:

Nurtugan Azatbekuly is a master's student in the Computer Science Department at Al-Farabi Kazakh National University (Almaty, Kazakhstan, nurtugang17@gmail.com). His research interests focus on the analysis and development of computer vision algorithms. ORCID ID: 0009-0007-5843-8995.

Bazargul Matkerim is a PhD in the Computer Science department at Al-Farabi Kazakh National University (Almaty, Kazakhstan, bazargul.matkerim@gmail.com). Her research interests include parallel computing and applications of machine learning. ORCID ID: 0000-0002-5336-687X.

Aksultan Mukhanbet is a PhD student in the Computer Science department at Al-Farabi Kazakh National University (Almaty, Kazakhstan, mukhanbetaksultan0414@gmail.com). His research interests include machine learning and computer vision. ORCID ID: 0000-0003-4699-0436.