# Paper Review

**Multi-Agent Reinforcement Learning Based Frame Sampling for Effective Untrimmed Video Recognition**
**Authors: Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, Shilei Wen**

**Artificial Intelligence - 2**
**Indian Institute of Technology, jodhpur**

**Reviewer:**

Umang Rajendra Barbhaya
M20CS017

Date: May 7th, 2021

**Abstract:** Current trends have seen increased enthusiasm with video recognition. Many researchers have tried building suitable agents using frame sampling technology to improve video feed recognition, but have degrading results to show for, especially in untrimmed videos. The cause has been identified due to the variation of frame-level salience. With the use of multiple parallel Markov decision processes, the authors are aiming to make a process which selects a frame/clip in every iteration with a gradually adjusting manner with initial sampling using deep reinforcement learning model, the authors of the reviewed paper, proposed a multi-agent reinforcement learning (MARL) framework. In this review, a study is done based on their experiments of the MARL framework. The results show that their policy network is able to generate the probability distribution over the action space along with a classification network for reward calculation and final video recognition.

**Keywords:** Reinforcement learning, video recognition, MARL, active learning, learning technique.

## Introduction

Recent trends with the usage of video surveillance, video search and more and more YouTube and online social media presence, their is an uprising in the usage of video recognition, as it partakes in the aforementioned usages. However, their is a fundamental difference between trimmed and untrimmed videos. untrimmed videos pose a more critical challenge since not all the frames consistently respond to the specified ground-truth label. Therefore, the authors Wu et al. [1] have come up with a study to find out which frames are the most informative and how they can be used to find information similar to a well trimmed video.

The current trends for the same results are based on the following techniques,

1. Extracting spatial features from the video frames, and then combining those features to perform recognition using a video descriptor. This is a two step solution.
2. The second technique is based on the 2D or 3D convolution solutions. In these solution, end-to-end video classification methods are used.

The researchers, however, faced problems as the results based on the aforementioned techniques have been, especially in untrimmed videos, poor. The cause has been identified due to the variation of frame-level salience.

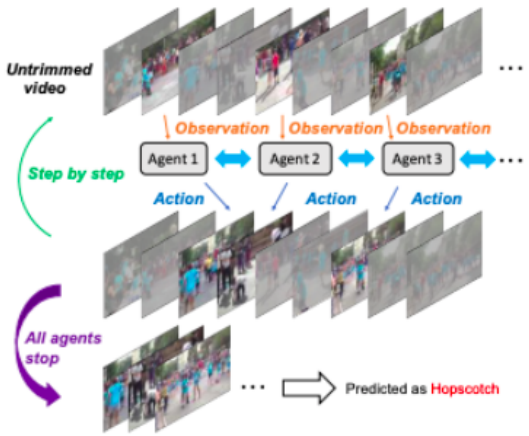


Figure 1: The painting process.

Therefore, with the use of multiple parallel Markov decision processes, each of which aims at picking out a frame/clip by gradually adjusting an initial sampling using deep reinforcement learning model, the authors of the reviewed paper, proposed a multi-agent reinforcement learning (MARL) framework, as shown in figure 1.

The researchers figured out that working with multiple N successive frames is like a brute force, making it an unnecessary computation, therefore there is a need for selecting the most informative frame. The process followed by them to make this agent is as follows,

1. Observing N scenes of the whole video.
2. Predict where the next useful frame will occur based on the past experiences and looking at the target video.
3. Using Markov Decision Process, to model a framework using reinforcement learning to automate the above process, and select multiple discriminating frames or video clips from an untrimmed video to improve the recognition performance.

Apart from the video recognition capabilities of RL, it has also been used in semantic level of video recognition, as it made an easy implementation for action detection and anticipation and uses the fast forward algorithm to remove the unnecessary complexity of the untrimmed video classification. In this process, RL is performing two tasks, frame skipping, and decision making.

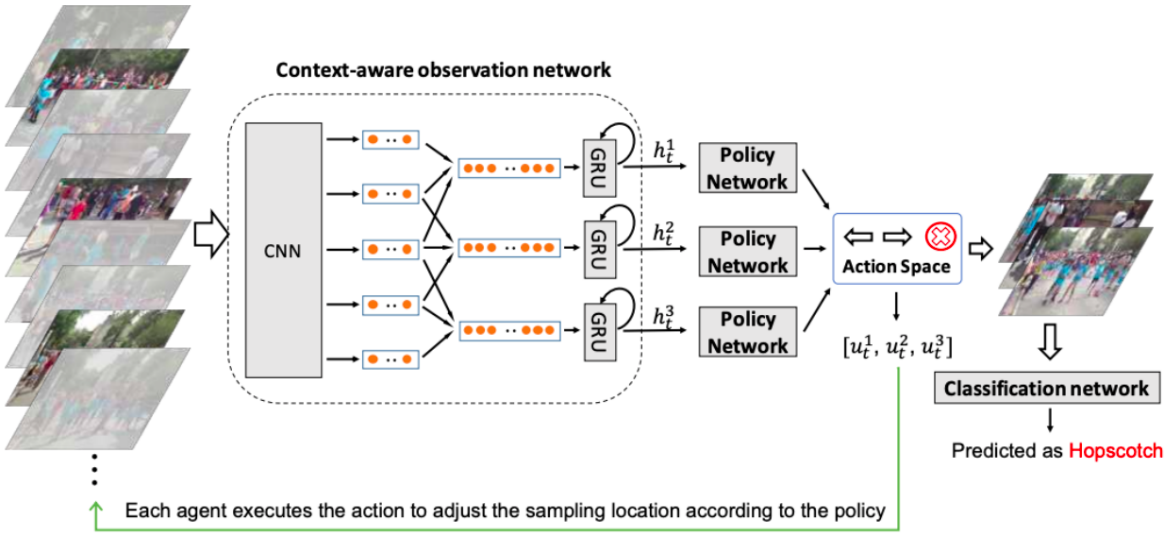


Figure 2:

Moving on to the architecture proposed by the researchers, as shown in figure 2 above, it is comprised of three different parts,

1. Context-Aware Observation network, this network would be used to encode the video feed. CNN based networks are being deployed, in order to gain information for a multi agent environment, and summarizes history using a recurrent neural network.
2. Policy Network. After every step $t$, the agent selects an action $a$ from the action space, according to the probabilistic distribution, generated by the policy network of the agent.

3. Classification Network, in this network, the frames are passed in an iterative manner, and the corresponding prediction logic is generated for all of these frames. For the simplicity in design, the classification network shares the parameters of layers before the last classifier layer with the observation network.

### Discussion

The researchers tried to verify their experiment, they tested their proposed framework on multiple datasets, such as ActivityNet v1.2 and v1.3, YouTube Birds and YouTube Cars, their results show that their proposed framework works better than the baseline 2D/3D convolution based neural networks.

The work proposed by the researchers closely resembles to the deep-learning based action recognition, including end-to-end convolutional classification networks and two-stage recognition solutions, however they have further improved the residual connection between concurrent streams and hence improved results. Also, the previously available techniques have been hand-crafted, and the researchers have proposed a learning-based strategy to improve recognition performance for untrimmed videos.

The work done by the researchers can be summed up in the following points,

1. They worked on the frame sampling rate, which was overlooked in the previously available solutions. They achieved their goal with the use of Markov Decision Process.
2. They implemented multi-agent reinforcement learning, which helped solve the problem of the sequential decisions to be made. Their agent was able to detect both historical and environmental information and then make decisions.
3. Their provided solution is versatile enough to be implemented with various solutions, previously made and in future as well.

### Conclusion

The researchers have proposed a learning-based strategy to improve recognition performance for untrimmed videos. The main usage of video surveillance, video search in recent trends is with YouTube and online social media presence, their is an uprising in the usage of video recognition, as it partakes in the aforementioned usages. However, their is a fundamental difference between trimmed and untrimmed videos. untrimmed videos pose a more critical challenge since not all the frames consistently respond to the specified ground-truth label. Therefore, the authors have come up with a study to find out which frames are the most informative and how they can be used to find information similar to a well trimmed video. Experimental results show that the agent can work with various types of solutions, and is very much efficient than the previously available solutions.

## References

[1] Wu, W., He, D., Tan, X., Chen, S., Wen, S. (2019). Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6222-6231).