Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM

Tasnimul Hasnat 190041113

March 14 2023

1 Introduction

In the field of computer vision, the recognition of human activity is a well-studied problem with several applications in robotics, human-computer interaction, surveillance, etc. Wide-scale video action detection has significantly improved recently thanks to the availability of large datasets, deep neural network topologies, video representation approaches, etc. On the other hand, several studies focused on specific action identification sub-tasks, such as the recognition of egocentric activity, the detection of anomalies, the action quality analysis (AQA), the localization of activity in space and time, etc. One such critical subset that is frequently employed in surveillance systems, internet video filtering, and public monitoring is violence detection. As digital media technologies like security cameras become more prevalent, it seems to be tougher to manually detect violence in video recordings.In order to overcome this issue, researchers have proposed numerous techniques that may automatically and without human participation identify violence in surveillance film. Violence detection is a component of the general action recognition task that focuses on identifying violent human actions like fights, robberies, riots, etc.

The focus of the majority of past research on violence identification [1]–[3] was on creating a number of descriptors that could precisely identify violent movements occurring in the video. The performance of these manually constructed features was eventually surpassed by a number of end-to-end trainable deep learning approaches that require little to no pre-processing [4]-[6]. Three well-known benchmark datasets—hockey, movies, and violent-flows—were used to assess the effectiveness of these methods. A recently proposed dataset called RWF-2000 is much bigger and more diverse. When applying these deep learning models in real-world practical applications, it is important to consider both compute economy and accuracy. In order to provide discriminative spatio-temporal features, we present a novel CNN-LSTM two-stream network that requires fewer parameters. In general action recognition tests, the environment or background knowledge may be employed as discriminative cues. For instance, seeing green grass in the background could be a good indicator that someone is playing golf. Contrarily, physical attributes like color, texture, and background information have little impact on aggressive behaviors and are largely eclipsed by body movements, posture, and interactions. We used background suppressed frames and frame difference as the inputs to our network in order to create discriminative features that may detect hostility.

We could summarize our main contributions as follows:

- ⋆ We put forward a new two-stream deep learning system that employs Separable Convolutional LSTM with pre-trained truncated MobileNet (SepConvLSTM).
- ⋆ We captured motion between frames using simple and quick input preprocessing techniques, and we highlighted moving objects in the frames while suppressing backdrops that aren't moving.
- ⋆ We employed SepConvLSTM, which reduces the number of parameters required by replacing the convolution process at each ConvLSTM gate with a depthwise separable convolution. We examined three fusion techniques for merging the output properties of two streams.
- ⋆ We analyze the performance of our models on three generally recognized benchmark datasets. The suggested model meets state-of-the-art performance on the other datasets while surpassing the prior best result for the RWF-2000 dataset. In terms of FLOPs and parameter requirements, our model performs adequately.

The remaining portions of the paper are structured as follows: Part 2 provides an overview of pertinent studies on violence detection. In Part 3, the recommended strategy is fully shown. Training techniques and experiments are presented in Section 4. Section 5 concludes our study and offers some ideas for further research.