TWO STREAM NETWORK FOR VISION BASED VOILENCE DETECTION

Department of Computer Science and Engineering

National Institute of Technology Patna

Under the Supervision of:
Dr. Piyush Kumar
Assistant Professor
Department of CSE

Presented By:
Sneh Kumar (2006052)
Arjun Singh (2006067)
Paras Punjabi (2006069)

Outline

1. Introduction
    - Motivation
2. Problem Statement
3. Literature Survey
4. Methodology
    - Dataset
5. Result Discussion
6. Conclusion and Future Work

Introduction

Human Activity Recognition (HAR)

- Is a collection of human/object movements with a particular semantic meaning.
- To develop an automated system for the same.
- To identify all the objects, persons and the actions performed by them in the given sensor data/video data.

Types of Human Activity Recognition

- Vision Based HAR
- Sensor Based HAR

Motivation

- The two-stream model for violence detection combines visual and temporal features to detect violence from videos.

- The two-stream model combines the outputs of two streams to capture both visual and temporal violence in videos, resulting in better accuracy in violence detection.

Problem Statement

- The problem statement of two-stream violence detection is to develop a computer vision system that can accurately identify violent actions in video data. The challenge in violence detection is that it requires the system to not only recognize the visual features of violence, such as blood or weapons, but also to analyze the temporal dynamics of the actions and movements involve.

Literature Survey

Paper	Year	Dataset	Aim of Work	Method Used
Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM [2]	2021	RWF2000	To perform Vision Based Human Activity	Two Stream Network
FTCF: Full temporal cross fusion network for violence detection in videos [1]	2022	Real Life Violence Situations Dataset	To perform Vision Based Human Activity	Two Stream Networks using FTCF Blocks
Efficient violence detection using 3d convolutional neural networks[3]	2019	Movies Fight Detection Dataset	To perform Vision Based Human Activity	3D Convolutional Networks

Methodology

Dataset

❖ Real Life Violence Situations Dataset

- Total Classes: 2 (Violence and Non-Violence)
- Video-Type: MP4
- Total Files: 2000 (1000 (Violence) and 1000(Non-Violence))
- Average Video duration: 4s
- FPS: 30
- Link: Dataset

Dataset (Cont…)

❖ Movies Fight Detection Dataset

- Total Classes: 2 (Fights and noFights)
- Video-Type: MP4 and AVI
- Total Files: 201 (100 (Fights) and 101(noFights))
- Average Video duration: 4s
- FPS: 30
- Link: Dataset

Dataset (Cont…)

❖ UCF50

- Total Classes: 50
- Video-Type: AVI
- Total Files: 6650 (133 videos per class)
- Average Video duration: 4s
- FPS: 26
- Link: Dataset

Overview of Proposed Model

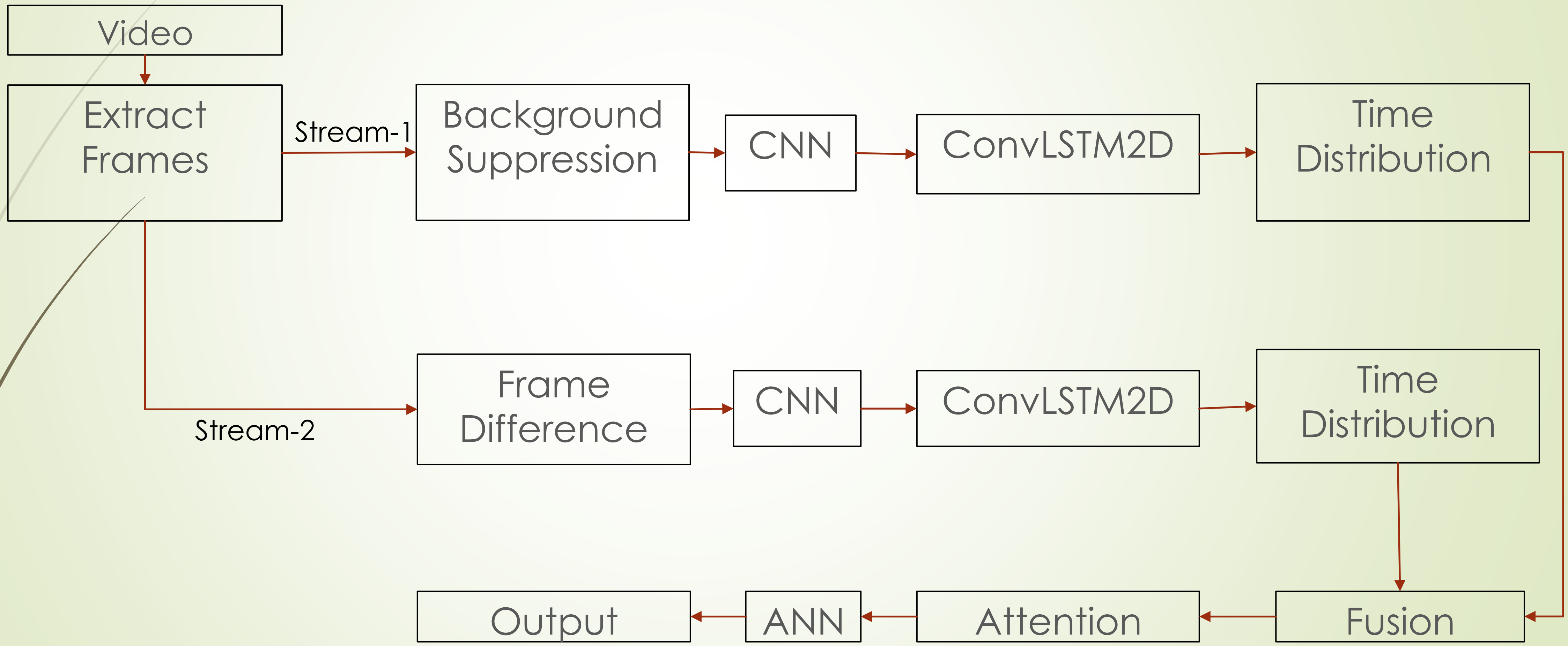

Fig. 1

Overview of Proposed Model (Cont...)

Frames Extraction

```
def frames_extraction(video_path:str):
    frames_list = []
    video_reader = cv.VideoCapture(video_path)
    video_frames_count = int(video_reader.get(cv.CAP_PROP_FRAME_COUNT))
    skip_frames_window = max(int(video_frames_count/SEQUENCE_LENGTH), 1)
    for frame_counter in range(SEQUENCE_LENGTH):
        video_reader.set(cv.CAP_PROP_POS_FRAMES, frame_counter * skip_frames_window)
        success, frame = video_reader.read()
        if not success:
            break
        resized_frame = cv.resize(frame, (IMAGE_HEIGHT, IMAGE_WIDTH))
        normalized_frame = resized_frame/255
        frames_list.append(normalized_frame)
    video_reader.release()
    return frames_list
```

Background Suppression

```
@tf.function
def background_suppression_image(img):
    m = tf.reduce_mean(img)
    f = img-m
    return f
```

Overview of Proposed Model (Cont...)

Convolutional Neural Network (CNN)

```
x = Conv2D(filters=16,kernel_size=(3,3),activation="relu")(x)
x = BatchNormalization(momentum=0.8)(x)
x = Dropout(0.1)(x)
```

ConvLSTM2D and Time Distribution

```
x = ConvLSTM2D(filters=8,kernel_size=(3,3),activation='tanh',data_format='channels_last',recurrent_dropout=0.2,return_sequences=True)(x)
x = MaxPooling3D(pool_size=(1,2,2),padding='same',data_format='channels_last')(x)
x = TimeDistributed(Dropout(0.2))(x)
```

Frame Difference

```
@tf.function
def calculate_frame_difference(video): # for a particular video
    out = []
    for i in range(SEQUENCE_LENGTH - 1):
        out.append(video[i+1] - video[i])
    out.append(out[-1])
    return tf.convert_to_tensor(out)
```

Overview of Proposed Model (Cont...)

Fusion and Attention

```
a = Add()([x,y])
z = MultiHeadAttention(num_heads=16,key_dim=64,dropout=0.1)(a,a)
f = Flatten()(a)
x = Dense(64,activation="relu")(f)
x = Dropout(0.2)(x)
out = Dense(len(unique),activation="softmax")(x)
model = Model(inputs=inp,outputs=out)
model.compile(optimizer="adam",loss="categorical_crossentropy",metrics=["accuracy"])
model.fit(X_train,y_train,epochs=25,batch_size=1,validation_split=0.3,shuffle=True)
```

Evaluation Metrics

- The performance of the presented model is analyzed on the following four popular metrics:
  - Accuracy:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

  - Precision:

$$Precision = \frac{TP}{TP+FP}$$

  - Recall:

$$Recall = \frac{TP}{TP+FN}$$

  - F1-Score:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision+Recall}$$

Result Discussion

❖ Real Life Violence Situation Dataset

Accuracy: 82%

Classes	Precision	Recall	F1-Score
Non-Violence	0.82	0.80	0.81
Violence	0.82	0.84	0.83

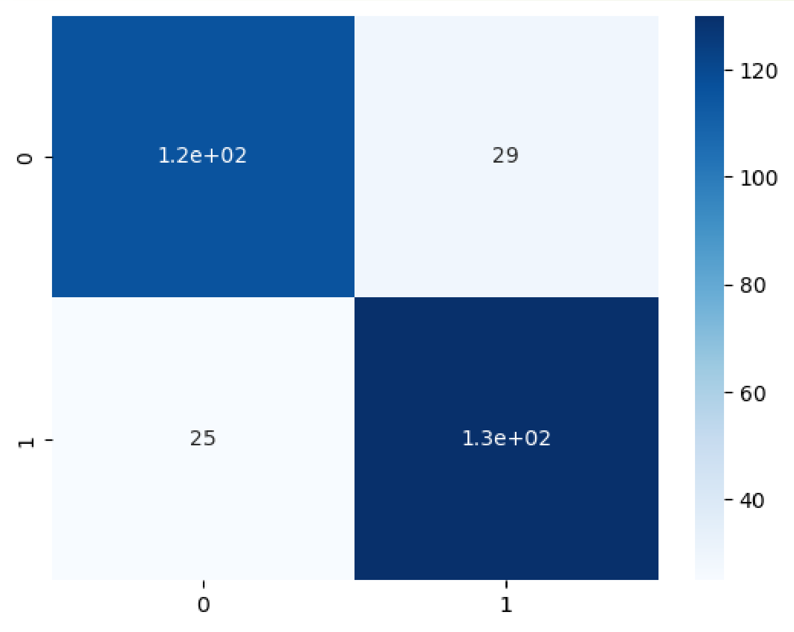

Fig. 2

Result Discussion (Cont…)

❖ Movies Fight Detection Dataset

Accuracy: 100%

Classes	Precision	Recall	F1-Score
Fights	1.00	1.00	1.00
Non-Fights	1.00	1.00	1.00

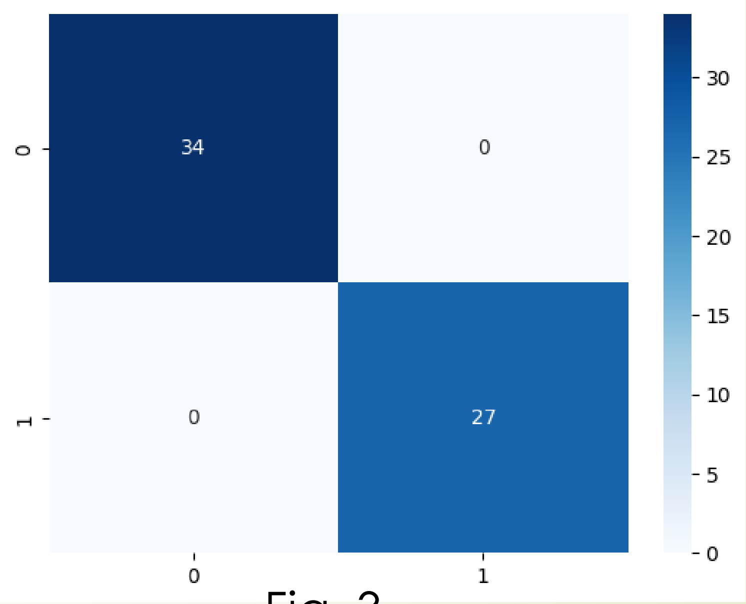

Fig. 3

Result Discussion (Cont…)

❖ UCF50

Accuracy: 78%

Classes	Precision	Recall	F1-Score
Baseball Pitch	0.82	0.90	0.86
Basketball	0.76	0.66	0.70
Bench Press	0.98	0.98	0.98
Biking	0.79	0.64	0.71
Playing Guitar	1.00	0.80	0.89
Walking With Dog	0.35	0.48	0.41
TaiChi	0.71	0.79	0.75
Swing	0.43	0.53	0.48
HorseRace	0.77	0.89	0.83
Punch	0.89	0.85	0.87

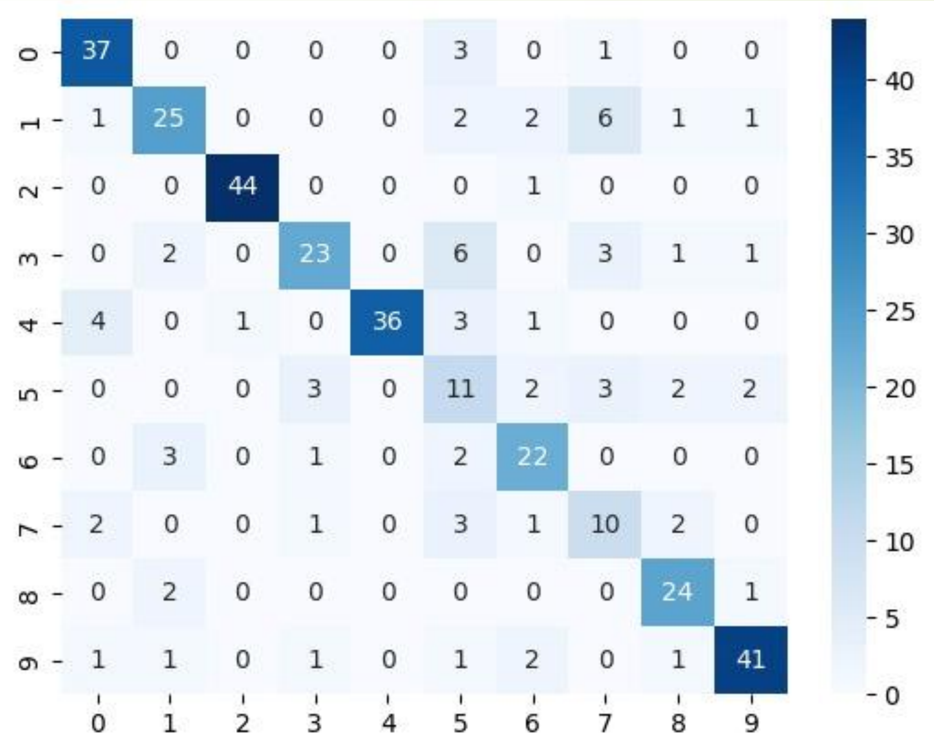

Fig. 4

Comparison with Existing Model

Paper	Dataset	Accuracy	Our Accuracy
FTCF: Full temporal cross fusion network for violence detection in videos[1]	Real Life Violence Situation Dataset	98.5%	82%
Efficient violence detection using 3d convolutional neural networks[3]	Movies Fight Detection Dataset	100%	100%
Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM [2]	RWF2000	89.75%	73%

Conclusion

- The paper proposes a two-stream network architecture that utilizes both spatial and temporal information to accurately classify violent and non-violent scenes.

Future Work

- To test the effectiveness of our model , we will create our custom dataset with labeled examples that are representative of the real-world data, ensuring that the model is trained on a diverse set of data and can generalize well to new and unseen examples.

References

1. Tan Zhenhua, Xia Zhenche, Wang Pengfei, Ding Chang and Zhai Weichao, “FTCF: Full temporal cross fusion network for violence detection in videos”.
2. Zahidul Islam , Mohammad Rukonuzzaman, Raiyan Ahmed , Md. Hasanul Kabir , and Moshiur Farazi , “Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM”.
3. J. Li, X. Jiang, T. Sun, and K. Xu, “Efficient violence detection using 3d convolutional neural networks,” in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019, pp. 1–8.
4. K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Advances in neural information processing systems, 2014, pp. 568–576.

THANK YOU