Poster ID:10 **SAVADOGO WENDOUNDI ABDOUL RASMANE, 2nd Year Master student. Cognitive Science Lab., The University of Aizu.**

Poster Session at Graduate School Information Fair: Factorized 3D Convolutional Neural Network for Violence Detection.

MOTIVATION

→Violence and terrorism incidents have become primary threats to world security and stability.
→Many Video surveillance are almost everywhere for people safety but due to human capabilities limitation, it is unrealistic, impossible to manually guard these videos and capture every violent scene in real time.

ARCHITECTURE

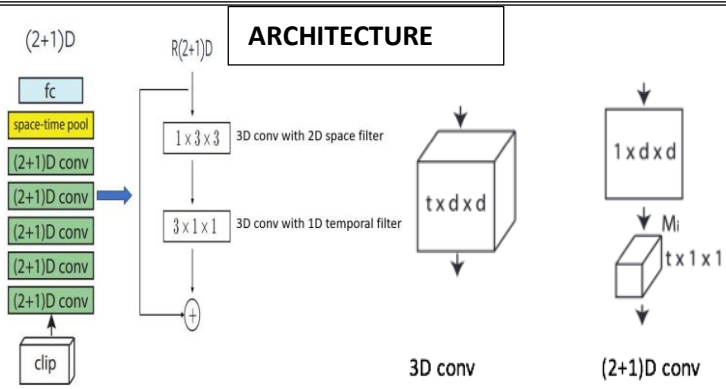

Fig. 2: 3D vs (2+1)D

Fig. 1: (2+1) D t = temporal depth, d = width, d = height

CONTRIBUTION: The factorized 3D CNN called **R(2+1)D is introduced for this violence detection task,** the model is based on the ResNet architecture.

→Benefit from the **residual aspect of the network.**
→**It doubles the number of nonlinearities** in the network due to the additional ReLU between the 2D and 1D convolution in each block. Increasing the number of nonlinearities increase the complexity of functions that can be represented.
→ **the optimization become easier** therefore the training error become lower compared to a 3D CNN of the same capacity.

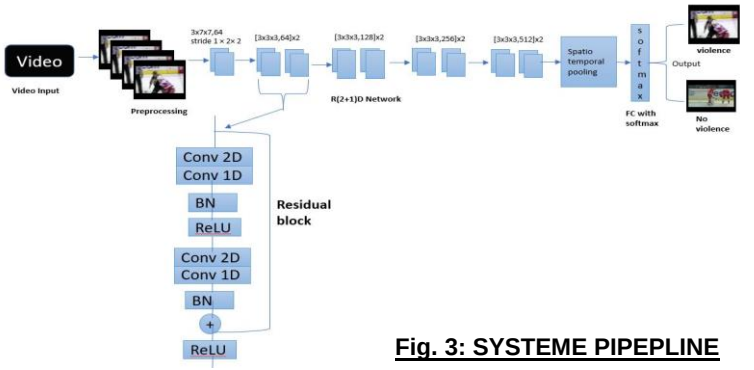

Fig. 3: SYSTEME PIPEPLINE

Dataset: For our work, we are using 3 videos datasets and each dataset has two labels: violence and no violence:
Real word fight dataset: 2000 videos
Hockey Fight Detection dataset: 1000 videos.
Crowd violence dataset: 246 videos.

RESULTS: At the early stage of the training, the best results are given bellow. Before being feed into the network, all the frames are resized to 128x171 and crop to 112. The dataset is shuffled, batch size is 100, clip length is 16. Learning rate is 0.001. The model is saved after each 50 epochs and tested after each 20 epochs.

Datasets	**Hockey Fight**	**Crowd Violence**	**Real World Fighting**
Training	98.28%	82.69%	80.47%
Validation	94.38%	87.50%	76.56%
Testing	96%	84%	74.75%

Table 1: training results.

Fig. 4: first line = Hockey Fight dataset; second line = Crowd violence dataset.

Fig.5: hockey dataset training results.

References: A Closer Look at Spatiotemporal Convolutions for Action Recognition Du Tran, Heng Wang, Lorenzo Torresani1, Jamie Ray, Yann LeCun, Manohar Paluri, Facebook Research, Dartmouth College.