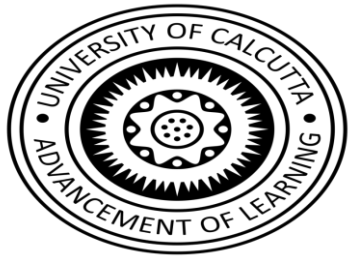


IEEE 18th India Council International Conference on Electrical, Electronics, and Computer Engineering

A Fusion Architecture for Human Activity Recognition

Paper ID: 1570747878

Sarosij Bose[*]*, Amlan Chakrabarti*[**]

[*]Department of Computer Science and Engineering, University of Calcutta.
[**]AK Chowdhury School of Information Technology, University of Calcutta.

Problem statement and Motivation

- **Problem**: Classifying human actions from entire videos.
- **Challenges:-**

1. Frame-by-frame based approaches weakly capture temporal information.
2. Vanilla 3D CNNs were too bulky and expensive.
3. Exclusive temporal data extraction from videos may not be always needed.
4. No defined relation can be captured between space and time.

- **Approach:-**

1. Adopt a simple late fusion based method between two different spatial streams.
2. Reduce space complexity and parameters to move towards a more online model.

Existing Drawbacks

- **Huge Computational Cost** A simple convolutional 2D net for classifying 101 classes has just ~5M parameters whereas the same architecture when inflated to a 3D structure results in ~33M parameters. It takes 3 to 4 days to train a 3DConvNet on UCF101 and about two months on Sports-1M, which makes extensive architecture search difficult and overfitting likely.

- **Capturing long context** models which predicted from a single frame were trained on the huge and diverse ImageNet dataset performed reasonably well by identifying the actions such as bending, falling etc. However for some complex/extended actions such as walking vs running or bending vs falling, more local temporal information became necessary since such differentiations are virtually impossible to make from a single frame.

- **Lack of suitable Datasets** The most popular and benchmark datasets have been UCF101 and HMDB-51 for quite some time. An architecture search on Sports1M can be extremely expensive. For UCF101, although the number of frames is comparable to ImageNet, the high spatial correlation among the videos makes the actual diversity in the training much lesser. Also, given the similar theme (sports) across both the datasets, Domain Generalization (DG) tasks also remain a problem. Other datasets such as THUMOS have limited classes and videos leading to overfitting. This has been solved to some extent lately with the introduction of some datasets like Kinetics, AVA etc. but other concerns remain.

Some Two-Stream Backbone Architectures

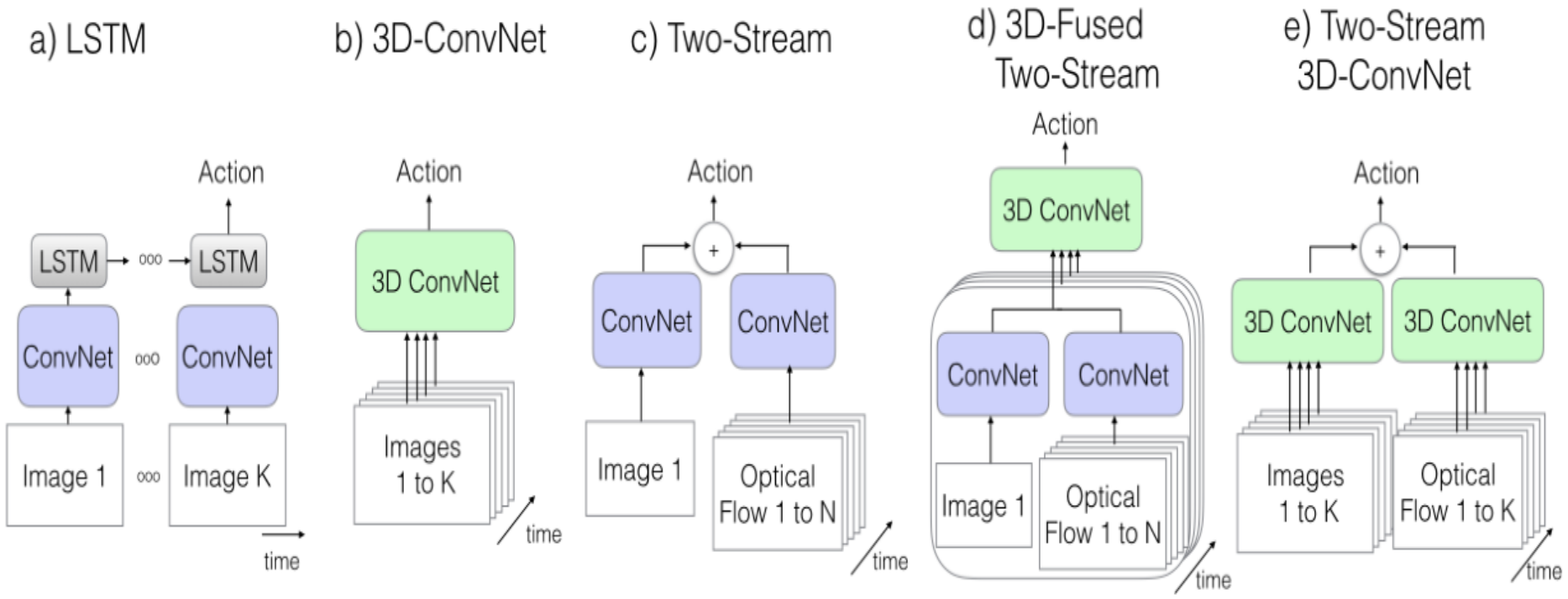
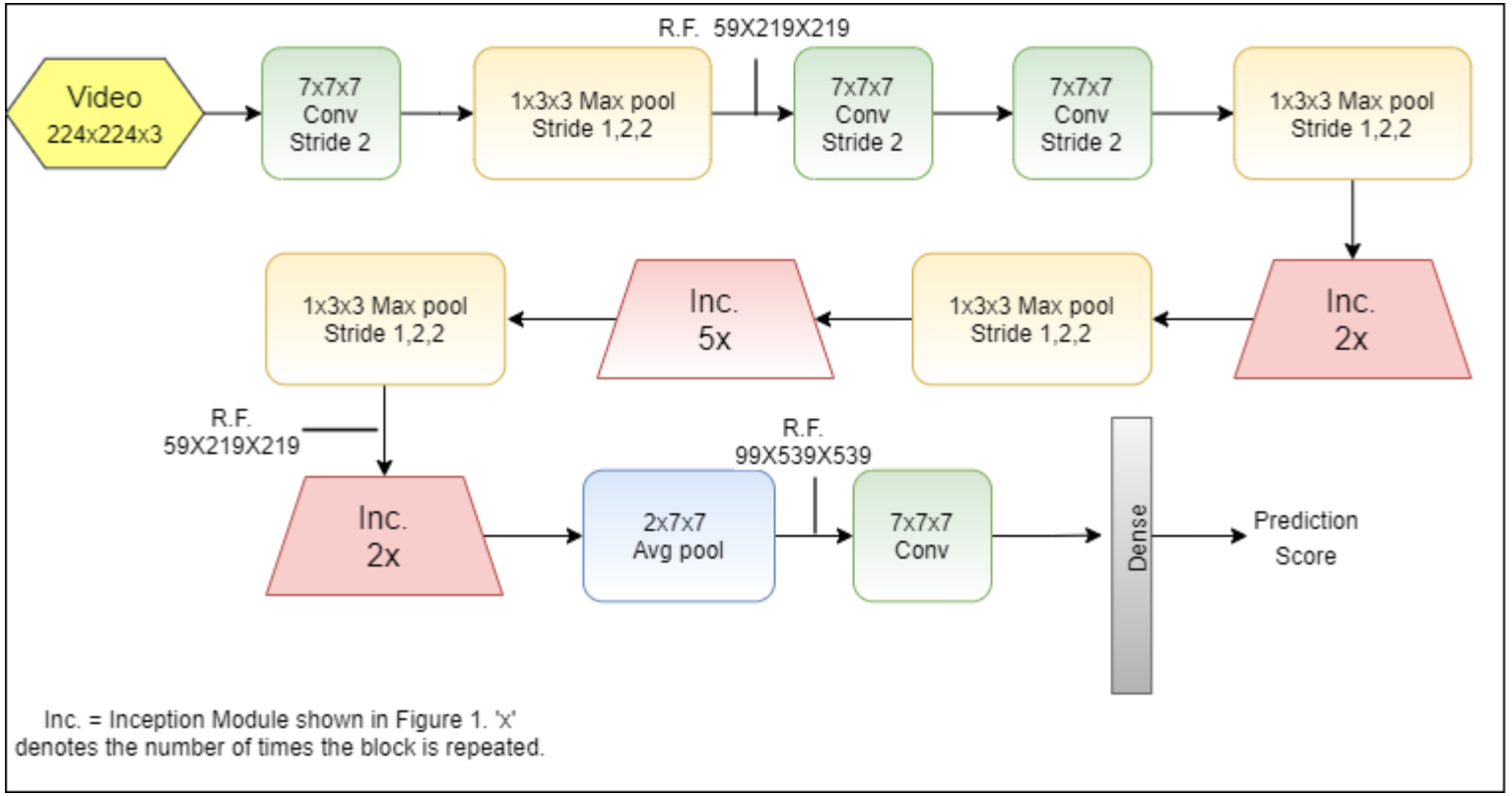


Fig 1. Evolution of various Architectures

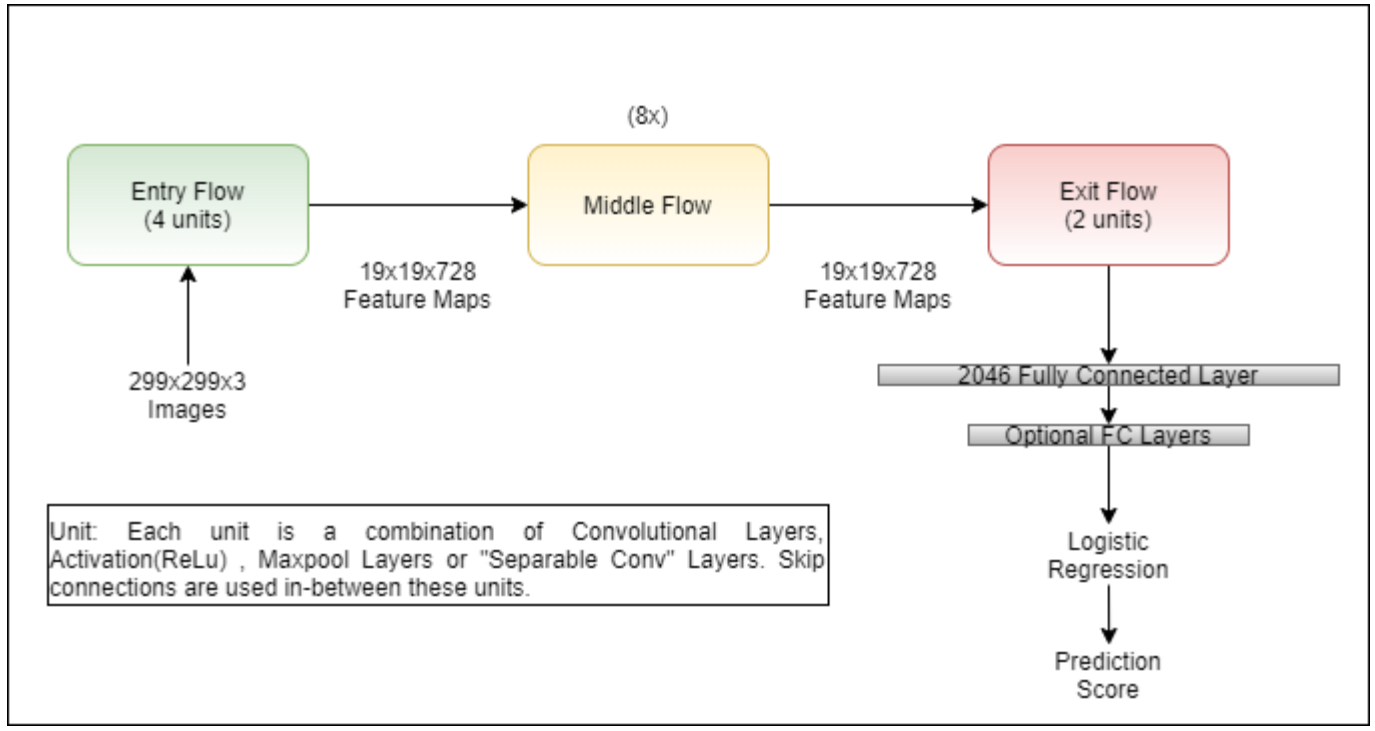
Implementation Details

- **SpatioTemporal Stream** This stream consists of the I3D Network, which is essentially just an inflated version of the Inception v1 architecture. Pre-processed videos each consisting of 79 frames in the numpy array (.npy) format is given as the input. Transfer Learning is used here based on the pre-trained weights of the Kinetics 600 dataset. The label space is also kept the same. The I3D network is shown in the diagram below.



Implementation Details

- **Spatial Stream** The spatial stream uses the Xception Architecture as the backbone which is pre-trained on the Imagenet dataset. The Xception network uses “separable convolutions” which is much more efficient than traditional convolutions like VGG and ResNet. It also takes in higher resolution images of 299x299. 5 center-cropped equally spaced frames are fed into this network and the highest (Top-1) prediction accuracy as well the corresponding label is then retrieved. The Xception architecture is shown below.



Proposed Model Architecture

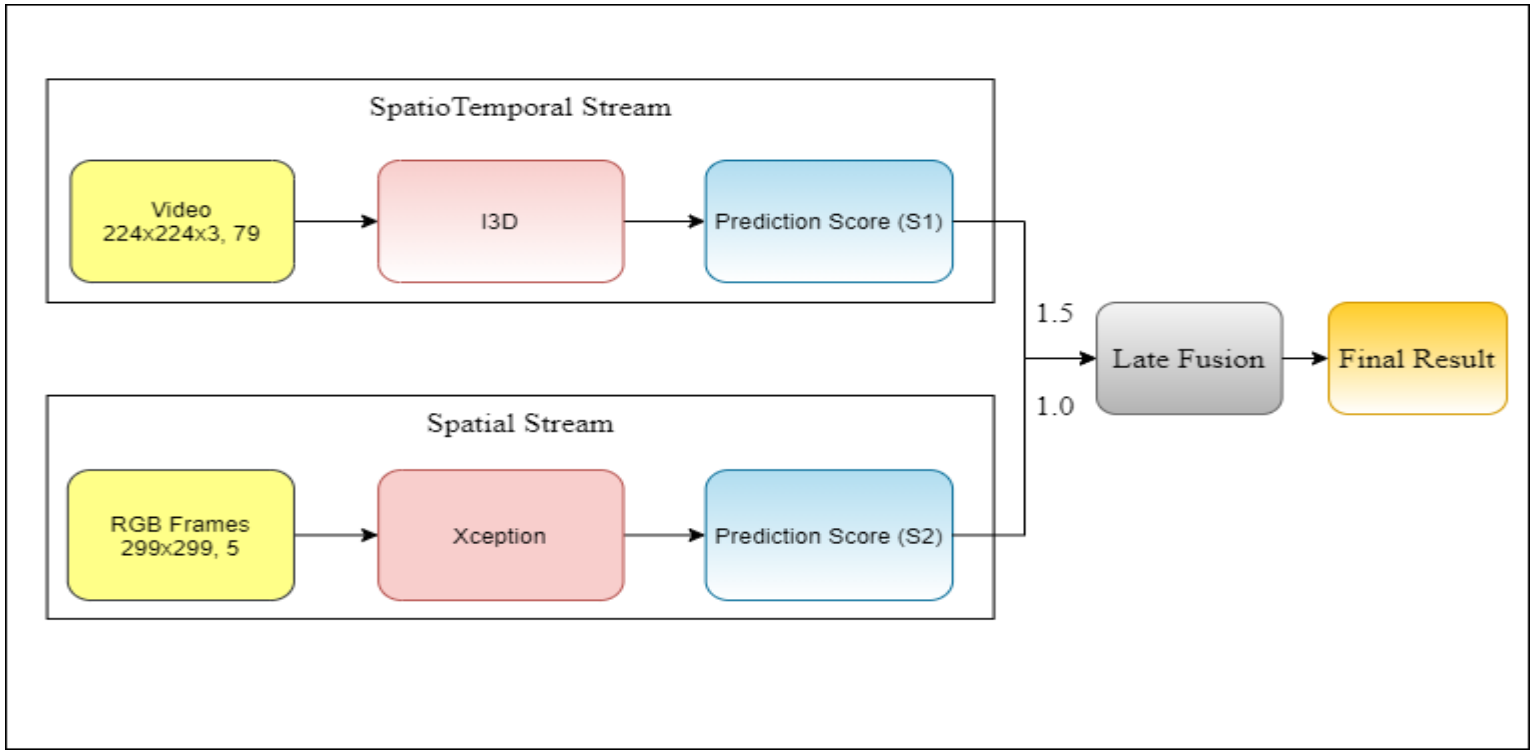


Fig 2. Model Architecture

The architecture primarily consists of the softmax score fusion between I3D and Xception. I3D is a 3D CNN model pre-trained on the Kinetics-600 dataset and Xception is another 2D CNN based model pre-trained on the ImageNet dataset. Final score is calculated based on the averaged late fusion between the individual streams.

Datasets used and Storage costs

SL No.	Dataset	Classes	Total elements	Source
1.	UCF-101	101	13, 000	YouTube
2.	Kinetics-600	600	500, 000	YouTube
3.	ImageNet	730-1000	1 mill. +	Manual coll.

Table 3

SL No.	Model Name	Storage
1.	Two Stream	65 GB.
2.	I3D	27 GB.
3.	**Ours**	**6 GB.**

Table 4

Table 3 features the datasets used in our research for Transfer Learning. In **Table 4**, we provide a **4.5x** improvement over the nearest comparable architecture. Most models which feature a temporal branch often demand pre-extracted optical flow data of the magnitude of 1TB to 2TB. Such offline solutions make it infeasible for edge deployment.

Results

SL No.	True Label	Predicted Label	Prediction %	Top 1/Top-5
1.	v_Bowling_g22_c04	Bowling	99.9 (Highest)	Yes/yes
2.	V _ CricketBowling_g02_ c01	Playing squash or racketball	71.2	No/yes
3.	v_BabyCrawling_g18_ C06	Crawling baby	98.1	Yes/yes
4.	v_HammerThrow_g23_ C05	Hammer throw	99.5	Yes/yes
5.	v_BrushingTeeth_g17_ C02	Brushing Teeth	97.6	Yes/yes

Table 1

SL No.	Model Name	Parameters	Top-1 Accuracy (RGB)
1.	LSTM	9 Million	68.2%*
2.	3D CNN	79 Million	65.4%
3.	Two Stream	12 Million	86.9%*^
4.	C3D	73 Million	82.3%
5.	Res3D	33 Million	85.8%
6.	T3D	25 Million**	71.4%
7.	I3D	25 Million	88.8%*^^
8.	**Ours**	**31 Million**	**87.5%**

Table 2

Obtained results based on our fusion model. Table 1 shows the prediction accuracies obtained on a few sample videos from UCF-101. Table 2 shows the parameters and Top-1 accuracies of our model compared to some other architectures.

Advantages and Applications

Edge deployment friendly Our model requires only 6 GB of secondary storage and is therefore not bulky. The advantage of using such a model which uses less secondary storage is that it can be readily deployed into various real-time systems and edge-devices where resources are constrained and storage space on device is very limited. The use of transfer learning also means that productionizing the model is easy and maintainable.

Numerous Applications Video understanding is probably the biggest application of SpatioTemporal fusion. Human action recognition, scene understanding, real-time detection and several other applications exist. In other areas such as CT scan diagnosis and Medical Imaging, it is useful to observe changes in patterns in the infected area over a certain period of time (abnormality detection) or Surgical workflow modelling and monitoring. Other areas include robotics (autonomous driving, 3D mapping in drones) and manufacturing (Quality control).

Future Direction

- Possible next approaches are broadly summarized below:-

1. Resolving the missing frames issue which may lead to mis classification of videos.
2. Improvement of the model for better video understanding. This can be done keeping in mind multiple approaches: Localization instead of brute-force classification, improving the temporal component using other architectures (self-attention, transformers etc).
3. If possible, find a specific use case (domain adaptation) in which this model can be applied. For Ex: HAR in an indoor environment, security applications etc.
4. Move towards more Online models for detection(body, objects etc).

References

- https://blog.qure.ai/notes/deep-learning-for-videos-action-recognition-review
- https://openaccess.thecvf.com/content_cvpr_2017/papers/Carreira_Quo_Vadis_Action_CVPR_2017_paper.pdf
- https://paperswithcode.com/task/3d-human-action-recognition
- https://openaccess.thecvf.com/content_cvpr_2017/papers/Chollet_Xception_Deep_Learning_CVPR_2017_paper.pdf

* Please see the paper for all the references.

Thank You!

For more information, please read our paper: “A Fusion Architecture for Human Activity Recognition”.

Code is available at: https://github.com/sarosijbose/A-Fusion-architecture-for-Human-Activity-Recognition