IEEE TRANSACTIONS ON INTELLIGENT VEHICLES, VOL. 7, NO. 3, SEPTEMBER 2022

Video Action Recognition for Lane-Change Classification and Prediction of Surrounding Vehicles

Mahdi Biparva, David Fernandez-Llorca, Senior Member, IEEE, Ruben Izquierdo Gonzalo, And John K. Tsotsos, Fellow, IEEE

SCH Univ.

Dept. of AI and Bigdata

Jaegyun Im

contents

1. INTRODUCTION

2. RELATED WORK

3. PROBLEM FORMULATION

4. VIDEO ACTIVITY RECOGNITION & PREDICTION

5. EXPERIMENTS & CONCLUSION

6. HOW TO APPLY

1. Introduction

General phenomenon : The driver uses only visual cues

Need to detect and respond early to increase safety and efficiency

Lane change recognition and prediction are video action recognition problems

Many approaches using the front camera are suggested

1. Introduction

Different Approaches

- Adaptive Cruise Control(ACC) ← Manual operation of the handle
- Traffic Jam Assist(TJA) & Traffic Jam Chauffeur(TJC)
- Highway Chauffeur(HHC)
- Highway Autopilot(HA)

The most advanced system

1. Introduction

Limitations exist

To fill in the gaps in the system, self-driving cars

use video-based behavioral recognition to predict lane changes.

1. Introduction

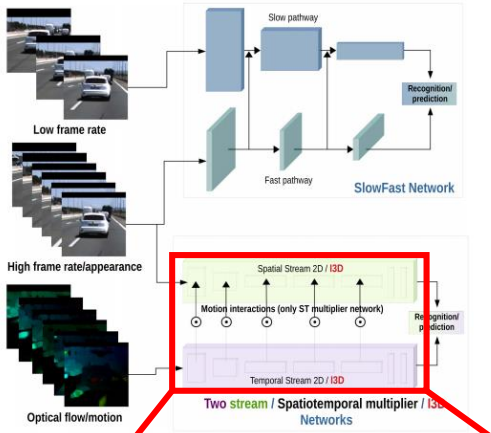
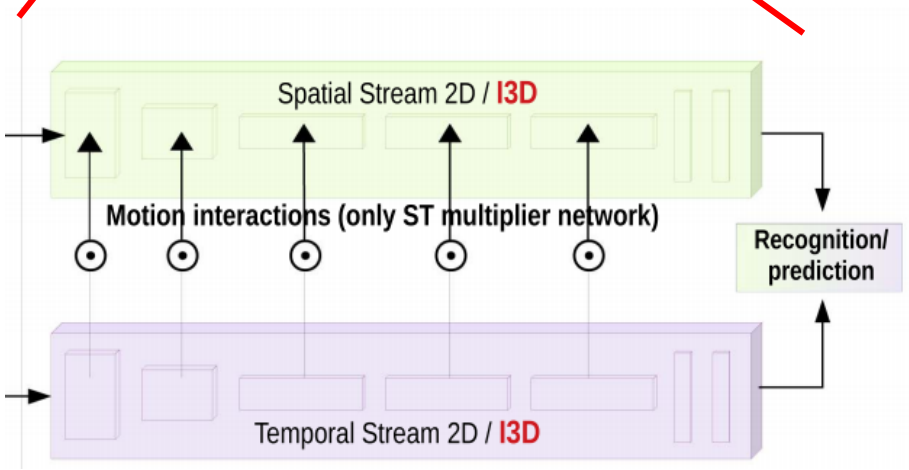
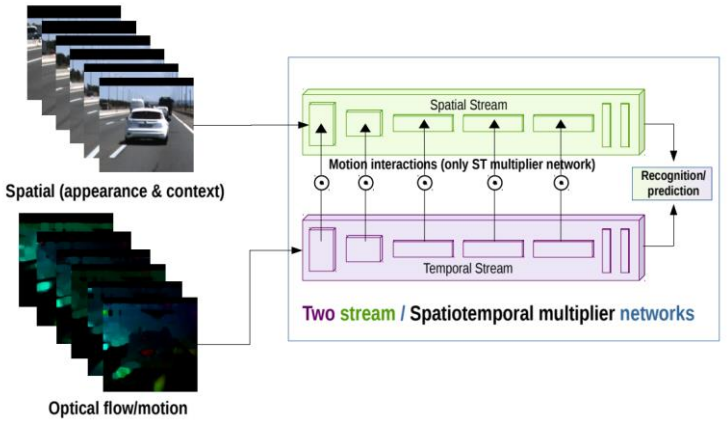


Fig. 1. Overview of the proposed video action recognition approaches for lane change recognition and prediction of surrounding vehicles, including Two-Stream Network, Two-Stream Inflated 3D ConvNet, Spatiotemporal Multiplier Network and SlowFast Network.





- Two-Stream Network: Combines spatial and temporal streams to recognize behavior
- Two-Stream Inflated 3D ConvNet: Extends traditional two-stream structures and transforms them into 3D to improve temporal and spatial classification
- Spatiotemporal Multiplier Networks: Combine spatial and temporal attributes to improve information transfer through interconnected residual connections
- SlowFast Networks: Capture temporal resolution by running two paths, with the slow path capturing spatial meaning and the fast path capturing temporal resolution

1. Introduction

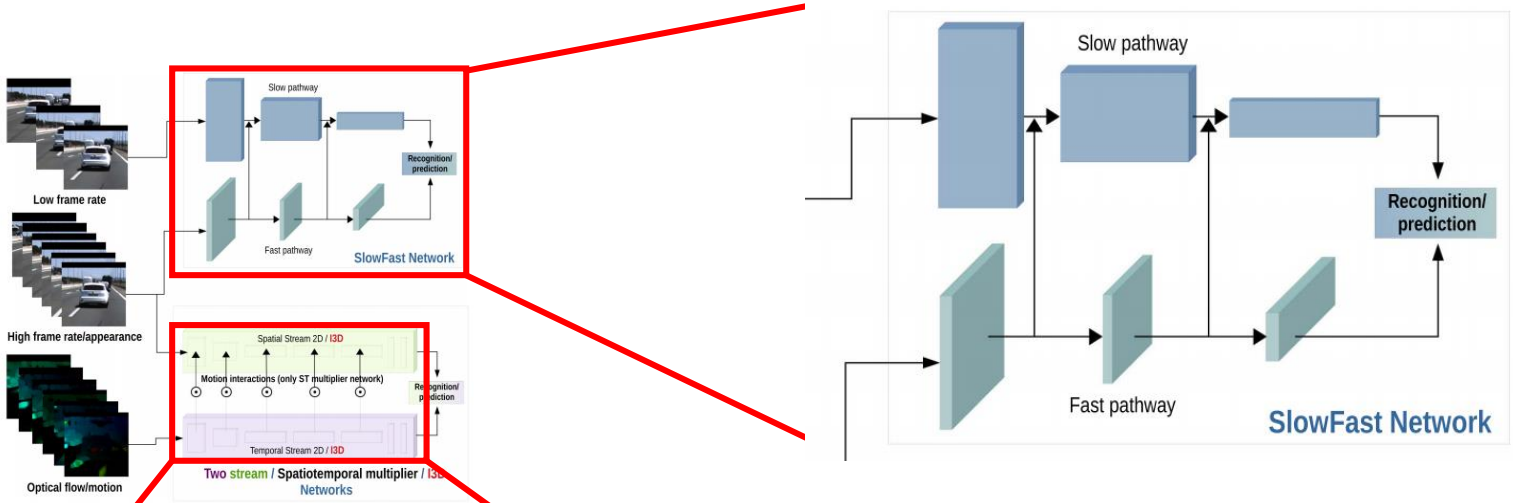
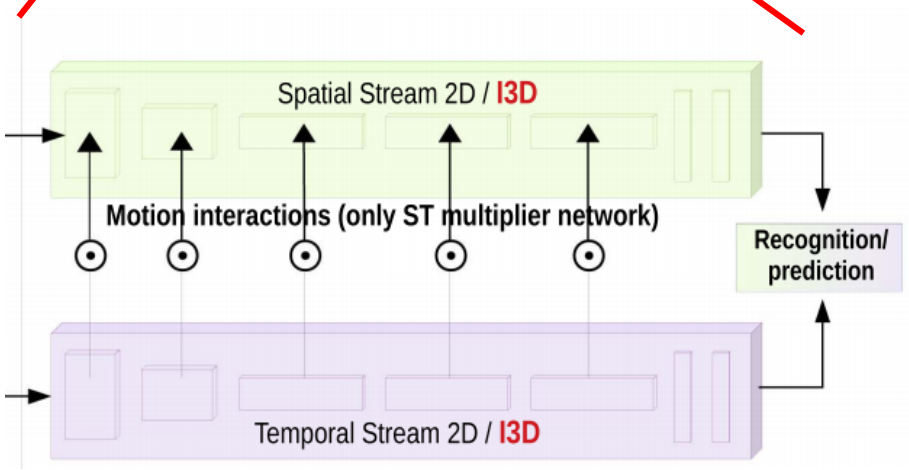


Fig. 1. Overview of the proposed video action recognition approaches for lane change recognition and prediction of surrounding vehicles, including Two-Stream Network, Two-Stream Inflated 3D ConvNet, Spatiotemporal Multiplier Network and SlowFast Network.



- Two-Stream Network: Combines spatial and temporal streams to recognize behavior
- Two-Stream Inflated 3D ConvNet: Extends traditional two-stream structures and transforms them into 3D to improve temporal and spatial classification
- Spatiotemporal Multiplier Networks: Combine spatial and temporal attributes to improve information transfer through interconnected residual connections
- SlowFast Networks: Capture temporal resolution by running two paths, with the slow path capturing spatial meaning and the fast path capturing temporal resolution

1. Introduction

Combine visual cues with temporal information

Predict lane changes for vehicles (and suggest new ways to do so)

Evaluated using the PREVENTION dataset, high prediction accuracy

2. RELATED WORK

Vehicle and lane marking detection and tracking is essential

Consider three levels of analysis

- Input variables
- Methodologies
- Datasets

2. RELATED WORK

- Input variables

Using physical variables : Define the vehicle's relative dynamic relationship with other vehicles and the environment

Includes lateral and longitudinal position (distance),
velocity, acceleration, time difference, heading angle and yaw rate

Most physical variables are obtained from cameras and distance sensors

Expect to measure accurately with onboard sensors

lateral and longitudinal acceleration, yaw angle or yaw rate = IMPRACTICAL

Expect V2V communication to solve the problem

2. RELATED WORK

- Input variables

  - A region of interest (ROI) is created for each vehicle detection
  - Contains information about the area around the vehicle
  - Appearance features extracted using GoogLeNet pre-trained on ImageNet
  - Benefits from not requiring an intermediate detection step

2. RELATED WORK

- Methodologies

Recognize vehicle lane change:
can be evaluated using trajectories estimated by behavioral models

Predicting trajectories of neighboring vehicles:
more accurate estimates when lane change intent recognition is available

Many previous studies do not consider vehicle-to-vehicle interactions

Other approaches:
based on the use of recurrent neural networks,
including regular LSTMs, LSTM encoder-decoders, and multi-mode architectures

Consequently, two-stream and intention-aware architectures have been proposed so far to perform lane change detection and prediction.

2. RELATED WORK

- Datasets

NGSIM I-80 : Dataset captured from infrastructure using installed cameras

HIGH D, inD, INTERACTION : Cameras on drones

PKU: Collect using a vehicle equipped with 2D-LiDAR (Road lane markings x,
number of road lanes x,
information about the relative position of the autonomous vehicle x)

ApolloScape: Data acquired in an urban environment from a car traveling at 30 km/h
using four cameras and two laser scanners (Label information x)

PREVENTION: Includes data collected from three radars, seven cameras, and one LiDAR, covering
up to 50 meters around the autonomous vehicle (up to 200 meters in the forward area)

3. PROBLEM FORMULATION

Definition: Lane change prediction as a multi-classification problem

Goal: Determine whether a vehicle changes lanes left or right (LLC, RLC),
stays in a lane, in the context of observations up to a given time N

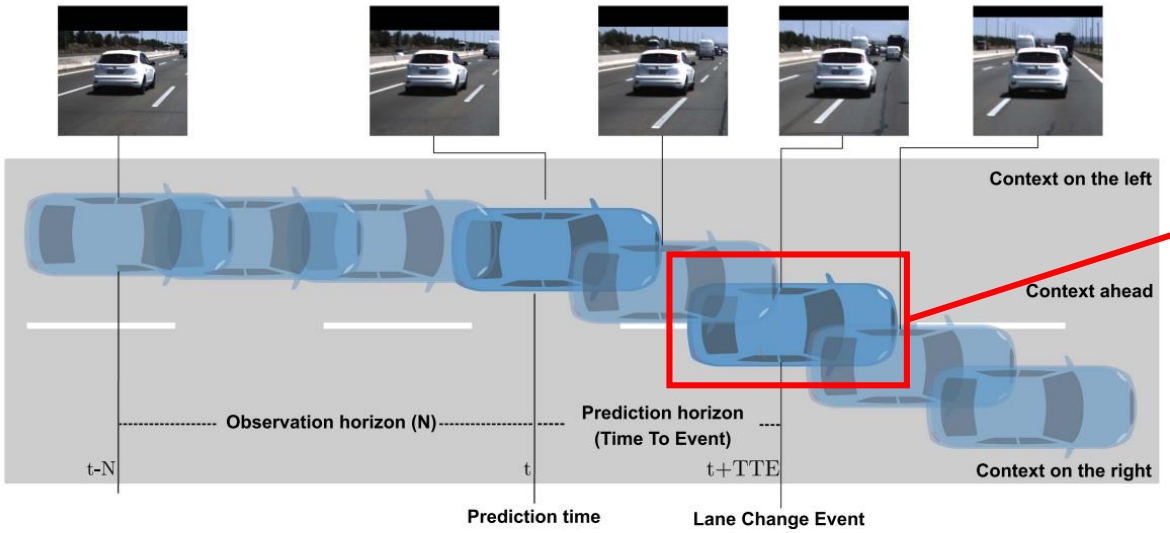


Fig. 2. Problem formulation: observation horizon (N), and time to event (TTE). The lane change event is labeled as the frame where the middle of the rear bumper is located just over the lane markings. This is the criterion established in PREVENTION dataset [15].

Lane Change Event : When the center of the rear bumper is directly over the lane markings

No Lane Change (NLC): Small lateral movements, lateral oscillations, or interrupted lane change maneuvers (including unsafe or aggressive behavior)

3. PROBLEM FORMULATION

Lane-change classification: when TTE = 0

- Observation horizon contains historical and current information to infer LLC and RLC classes

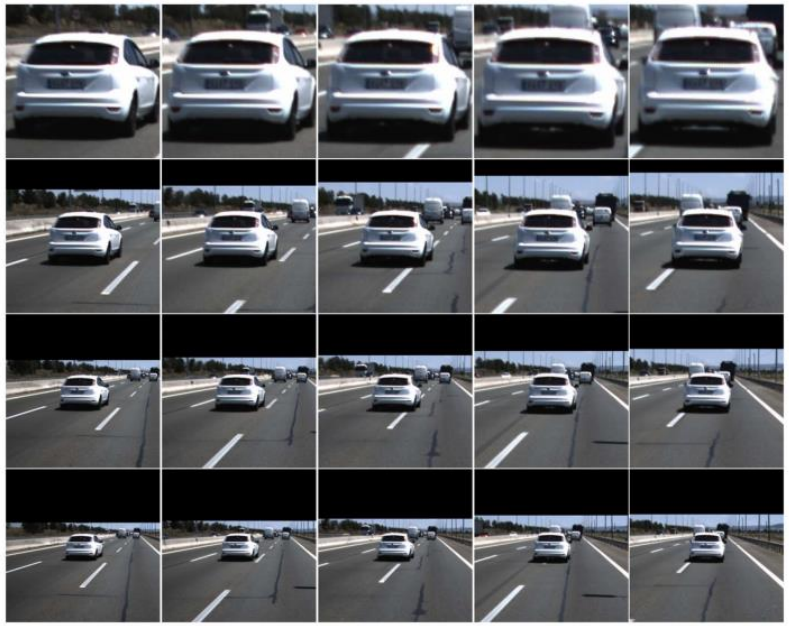
Lane-change prediction: when TTE > 0

- Observation horizon contains more or less information about actual lane change maneuvers for LLC and RLC classes

For very high TTE values the maneuver may not have started yet

Examine the impact of TTE or forecast horizon and number of observations (N) on the accuracy of lane change classification and prediction

3. PROBLEM FORMULATION



x4

x3

x2

x1

Fig. 3. ROI sizes. From upper row to lower row: x1, x2, x3 and x4. The vehicle is always centered. Zero-padding is applied when needed.

Vehicles are always centered, with zero padding applied as needed

The prediction relies on visual cues that are computed from regions of interest (ROI) extracted from the contour labels provided in the PREVENTION dataset

The size of the ROI controls the amount of contextual information considered in the input data stream

x1 contains information primarily related to vehicle appearance

x4 includes a large amount of front and side contextual information

3. PROBLEM FORMULATION

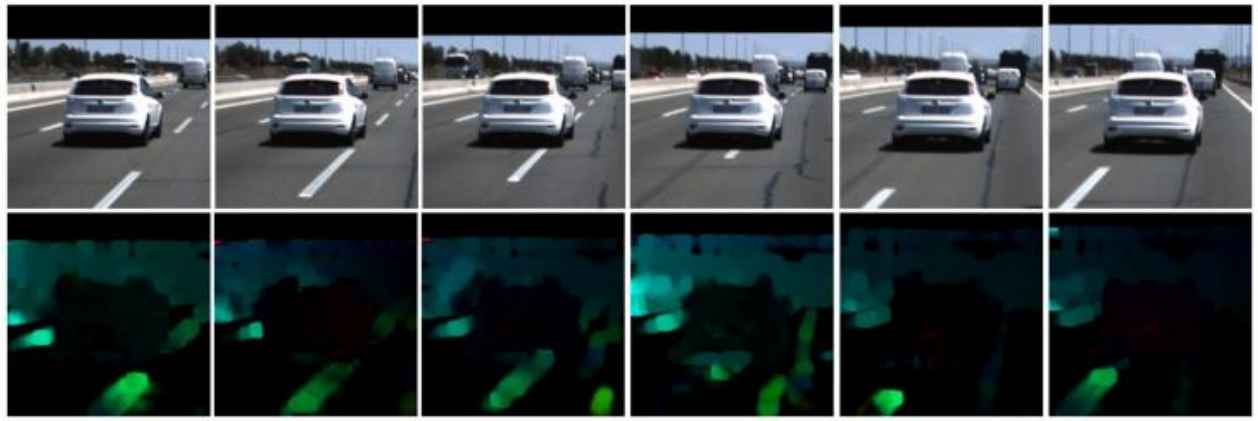


Fig. 4. Example of dense optical flow computation.

Optical flow is low in the region where the vehicle is, while it is more predominant around it

4. VIDEO ACTIVITY RECOGNITION & PREDICTION

Sequences of stacked images or regions of interest naturally decompose into spatial and temporal components

Spatial part in the form of a traditional area approach: conveying information about the vehicle itself and its surrounding context

All regions are created around the vehicle's outline to ensure the vehicle is always centered in the region of interest

4. VIDEO ACTIVITY RECOGNITION & PREDICTION

Consider four approaches to recognizing video activity

Disjoint Two-Stream Convolutional Networks

Two-Stream Inflated 3D Convolutional Networks

Spatiotemporal Multiplier Networks

SlowFast Networks

4. VIDEO ACTIVITY RECOGNITION & PREDICTION

Disjoint Two-Stream Convolutional Networks

A two-stream ConvNet architecture

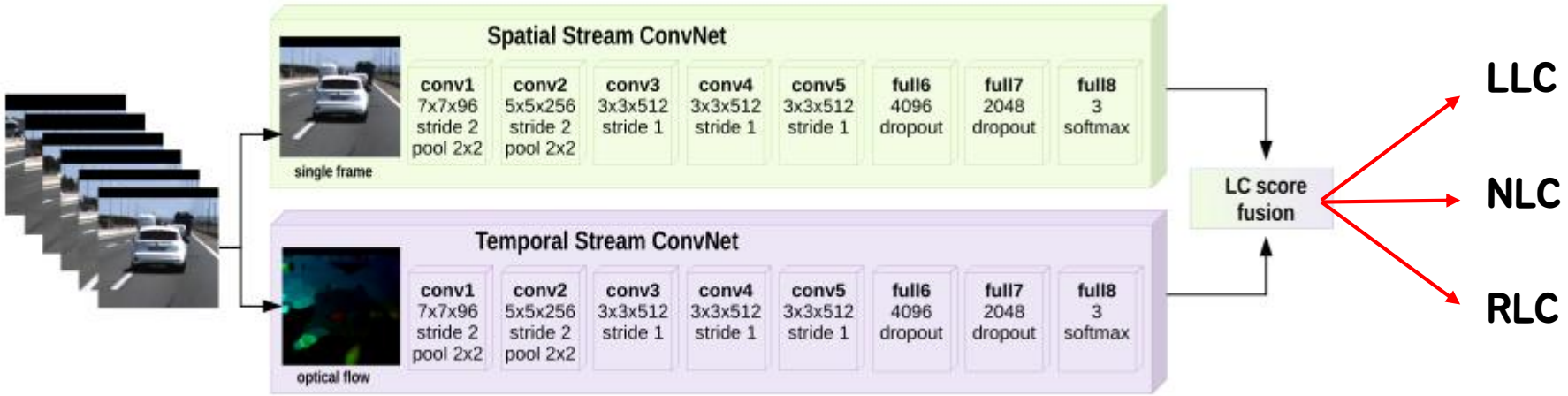
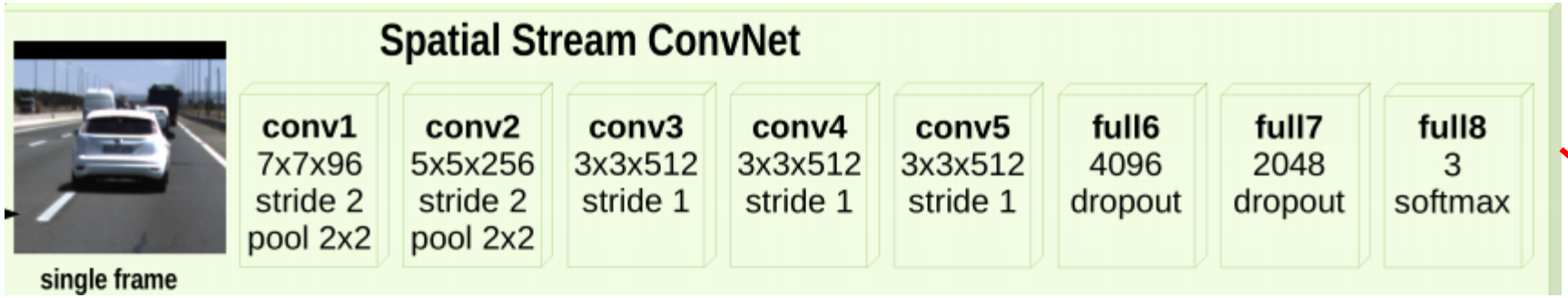


Fig. 5. Disjoint two-stream architecture for lane change classification and prediction.

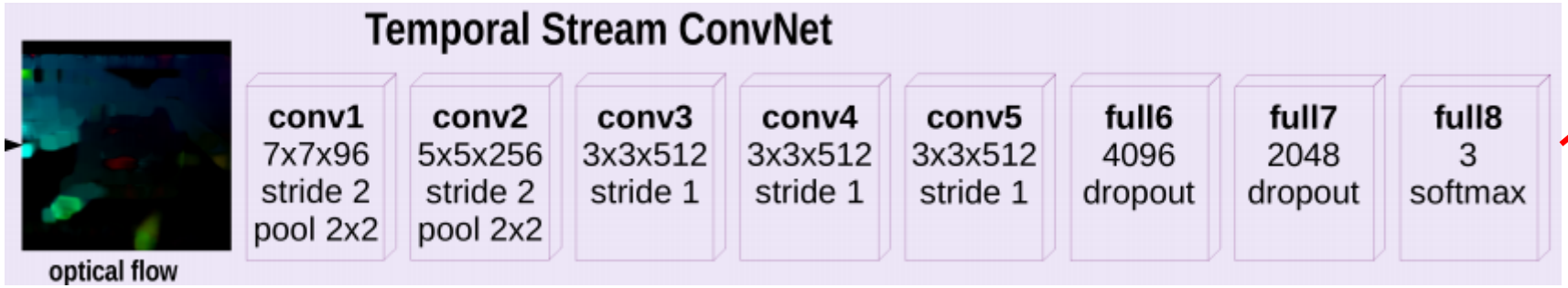
Includes 5 convolutional layers and 3 fully connected layers

4. VIDEO ACTIVITY RECOGNITION & PREDICTION

Disjoint Two-Stream Convolutional Networks



pre-trained using ImageNet



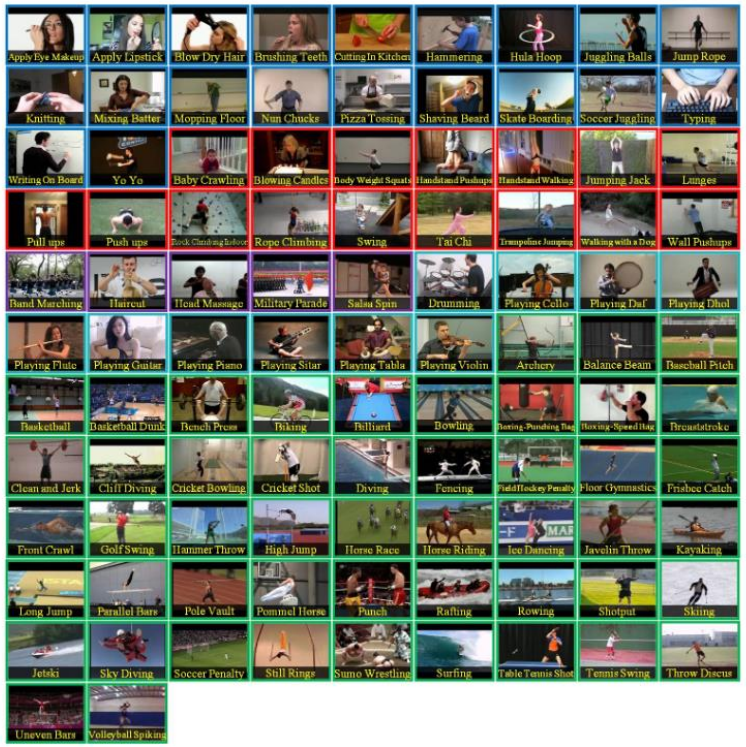
Hidden layers Using ReLU

multi-task learning using UCF-101 and HMDB-51

4. VIDEO ACTIVITY RECOGNITION & PREDICTION

Disjoint Two-Stream Convolutional Networks

UCF-101



HMDB-51

4. VIDEO ACTIVITY RECOGNITION & PREDICTION

Two-Stream Inflated 3D Convolutional Networks

A Natural Approach to Video Modeling: Using 3D Convolutional Neural Networks

Use spatiotemporal filters to create hierarchical representations of spatiotemporal data

3D Filter: Iterative copying with image video sequences
Bootstrap from pre-trained ImageNet models

The inputs to the model are short 16-frames sequences

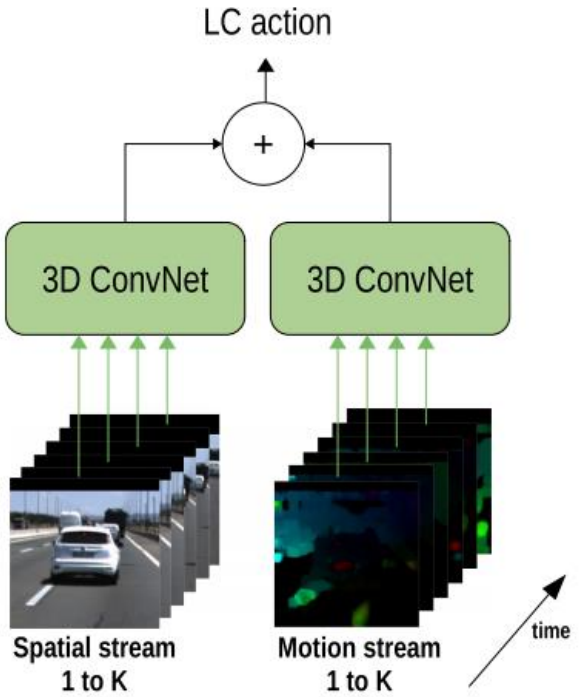


Fig. 6. Two-stream inflated 3D ConvNet for lane change classification and prediction.

4. VIDEO ACTIVITY RECOGNITION & PREDICTION

Spatiotemporal Multiplier Networks

Predicting vehicles that are not changing lanes but have their turn signals on

Using ResNets as a general architecture for spatial and temporal streams

$$\hat{x}^a_{l+1} = f(x^a_l) + \mathcal{F}\left(x^a_l \odot f(x^m_l), W^a_l\right)$$

x^a_l 와 x^m_l 은 각각 appearance 경로와 motion 경로의 L-번째 층의 입력이며,

W^a_l 는 변환에 사용되는 가중치를 나타냄

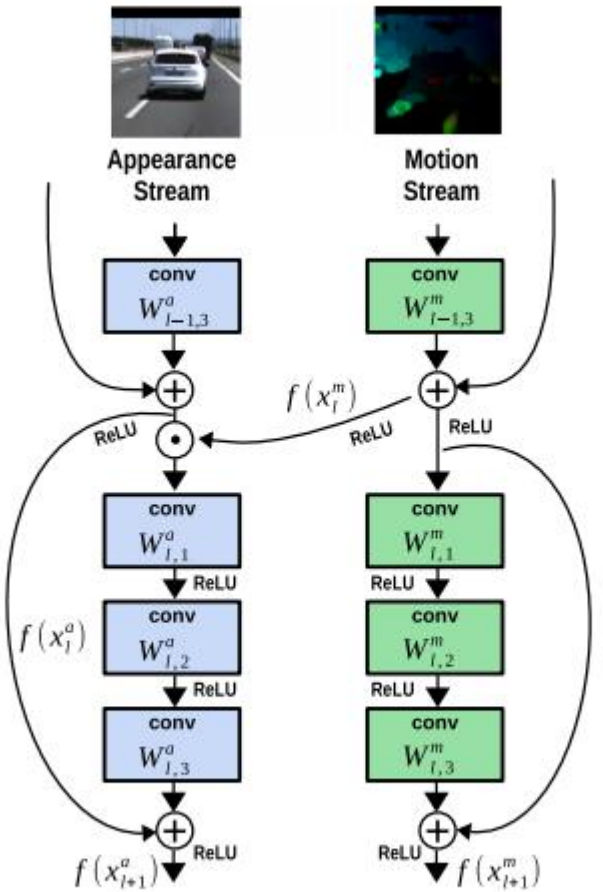


Fig. 7. Multiplicative residual gating from the motion stream to the appearance stream.

4. VIDEO ACTIVITY RECOGNITION & PREDICTION

SlowFast Networks

The most successful video motion recognition approaches

Can be considered a two-stream approach, but using the behavior path directly X

Slow streams operate with low frame rates, slow refreshes
-> Capture semantic information

Fast streams are high resolution, fast refresh rate
-> capture fast-moving action

Slow pathway is T = 16

Fast and Slow streams is α = 8.

The ratio of channels of the Slow stream with respect to the Fast one is 1/8

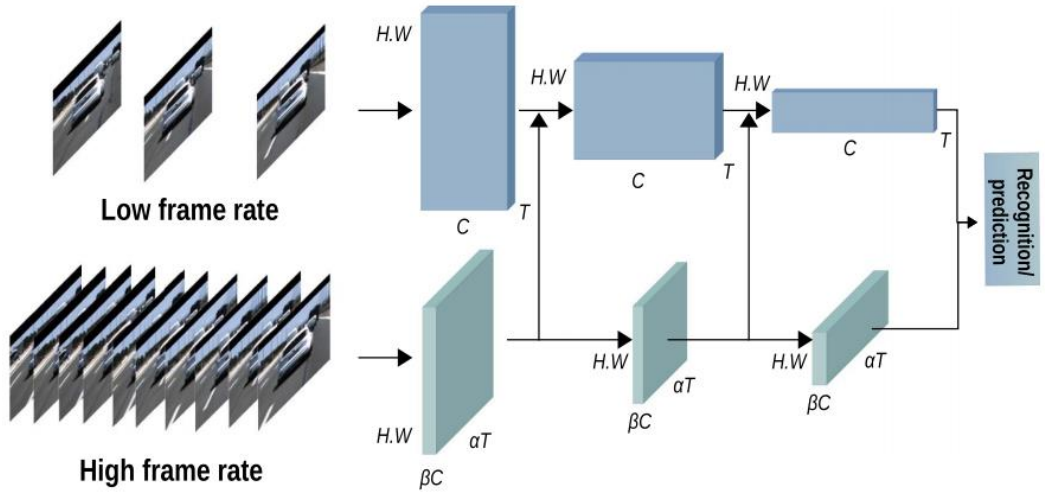


Fig. 8. SlowFast network for lane change recognition and prediction. The fast stream is lightweight by using a fraction $\beta = 1/8$ of channels.

The two pathways are fused by lateral connections

5. EXPERIMENTS

TABLE I
MAIN STATS OF THE DATASET. NLC/LLC/RLC: NO/LEFT/RIGHT LANE-CHANGE

	NLC	LLC	RLC
# of sequences	3110	342	438
avg. # of frames	50.9	96.8	80.1

The input size of both streams is 112 X 112

Training : 85%
Validation : 15%

Important : SlowFast Networks

TABLE II
LANE-CHANGE CLASSIFICATION ($TTE = 0$) ACCURACY (%)

		ROI size			
Method	Obs. Horizon	x1	x2	x3	x4
Baseline	20	83.41	83.25	**85.35**	84.06
	30	81.96	83.25	82.61	85.19
	40	81.80	82.45	81.32	81.80
Disjoint	20	83.22	86.18	86.26	87.43
	30	83.55	86.69	86.84	86.68
	40	84.97	87.69	**89.46**	88.79
I3D	20	82.45	**86.47**	85.99	85.67
	30	82.13	83.74	83.90	84.06
	40	82.13	83.09	81.80	82.29
ST	20	83.39	85.03	86.51	86.16
	30	84.38	84.70	85.36	84.73
	40	86.02	87.83	**90.30**	89.64
SF	20	88.89	89.69	**90.98**	89.37
	30	88.57	89.53	88.24	89.69
	40	86.96	89.05	89.53	90.34

Frames ([20 frames(2 seconds)], [30 frames(3seconds)], [40frames(4seconds)])

5. EXPERIMENTS & CONCLUSION

TABLE III
LANE-CHANGE PREDICTION ACCURACY (%). OBSERVATION HORIZON = 20 FRAMES (2 SECONDS)

		ROI size			
Method	TTE	x1	x2	x3	x4
Baseline	10	82.63	82.95	83.44	82.79
	20	82.00	81.67	82.79	**83.61**
Disjoint	10	84.05	84.54	85.20	85.36
	20	85.20	88.82	**91.02**	90.92
I3D	10	81.33	83.28	83.60	83.60
	20	81.01	81.67	**83.93**	83.61
ST	10	84.70	85.69	85.20	86.51
	20	86.84	90.30	91.45	**91.94**
SF	10	85.23	**88.96**	88.64	87.99
	20*	85.27	83.31	83.61	83.61

*Inconclusive results due to GPU memory limitations.

ROI size ⇧ ⟶ Performance ⇧

TABLE IV
SPATIOTEMPORAL MULTIPLIER NETWORK CONFUSION MATRIX, OH=20, TTE=20, X4

	Target class			
Output class	**NLC**	**LLC**	**RLC**	**Precision**
NLC	476	5	6	97.7%
LLC	8	33	11	63.5%
RLC	11	8	50	72.5%
Recall	96.2%	71.7%	74.6%	91.9%

⇩

Interrupted lane change maneuvers

5. EXPERIMENTS & CONCLUSION

Limitations

Other models mini-batch 128

SlowFast model batch size 8

If limit x -> the slow-fast model would have provided better results

6. HOW TO APPLY

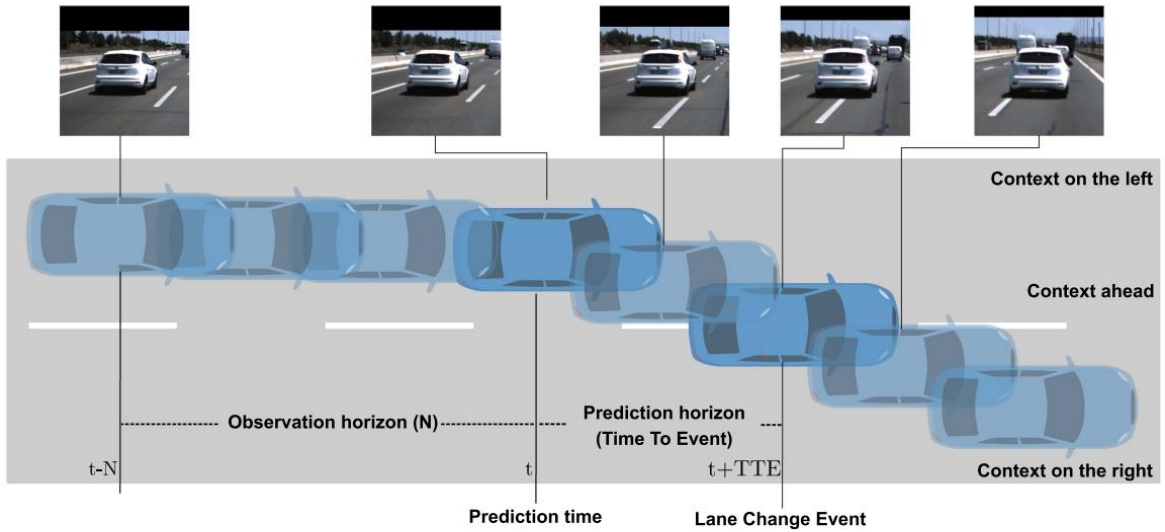
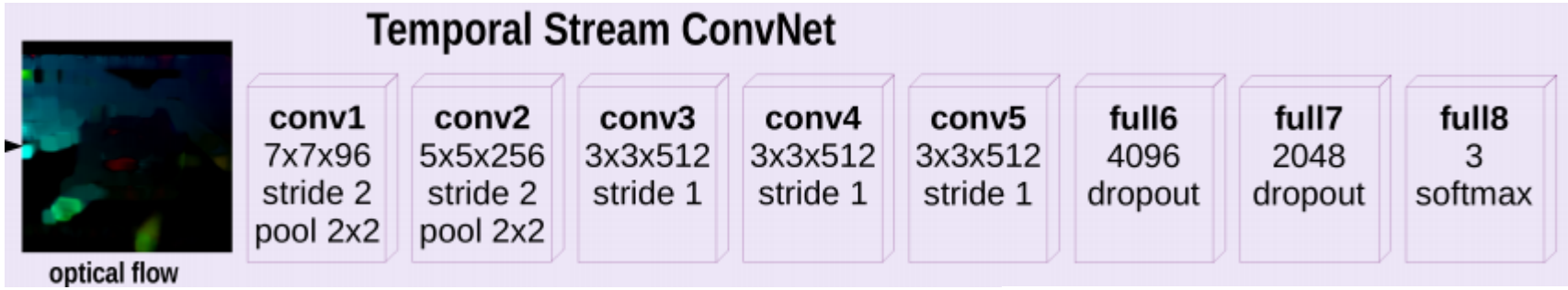


Fig. 2. Problem formulation: observation horizon (N), and time to event (TTE). The lane change event is labeled as the frame where the middle of the rear bumper is located just over the lane markings. This is the criterion established in PREVENTION dataset [15].

vehicle's rear bumper is in the center of the lane,

it's considered a lane change

6. HOW TO APPLY



Extract using optical flow

Separately Training Appearance & motion

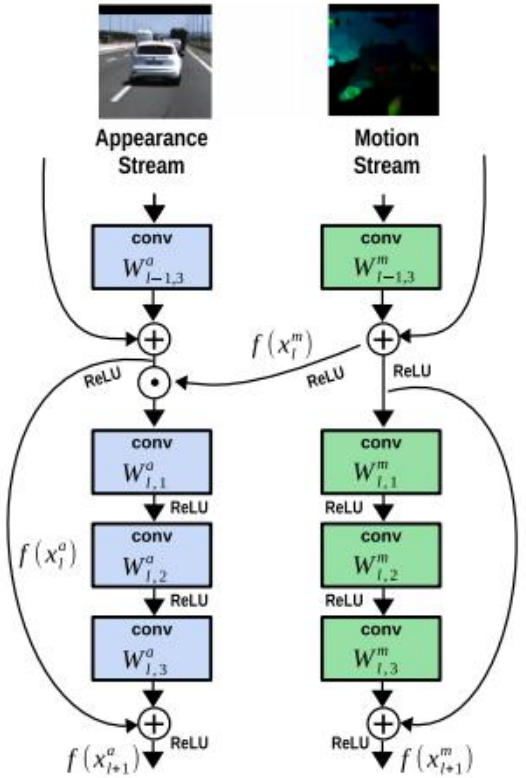


Fig. 7. Multiplicative residual gating from the motion stream to the appearance stream.

6. HOW TO APPLY

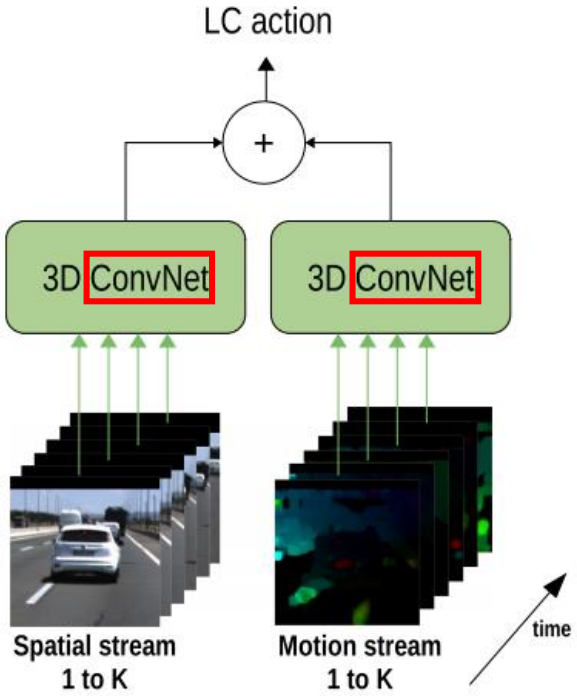


Fig. 6. Two-stream inflated 3D ConvNet for lane change classification and prediction.

3D RNN model

3D LSTM ..etc