Computer Vision on the Edge

Akin Osman Kazakci
Maitre-Assistant
IHEIE, MINES ParisTech
akin.kazakci@mines-paristech.fr

Long story short

Ready for mobile and edge platforms:

- Object detection (2D) is now a widely disseminated application type, where the main issues are quality of data and speed&accuracy tradeoff.

Not mature enough:

- Video recognition (3D = 2D+time), on the other hand, is still an active research issue with several architectural and computational bottlenecks.

Video understanding is a key enabling technology that will enable applications in a wide range of domains such as autonomous vehicles, smart cities, home automation and interactive robotics.

Despite the tremendous potential, for many use-case scenarios, video understanding remains largely unfeasible due to important constraints

Cost

Existing deep learning architectures for video understanding requires very large computational resources to be used. Such models requires large computational resources for training and inference whose costs cannot be justified for all use cases. Also, using large models may incur significant energy footprint, which adds indirectly to model costs.

Latency & Bandwidth

A cloud-first approach is unsuitable for many real-life use-cases applications since sending the video data to the cloud, making the inference, and sending the inference results back to where the video data has originated increases latency. This is especially inacceptable for applications requiring real-time response (e.g; safety related applications).

Privacy

Despite the potential utilities video understanding has to offer, using camera data has important ethical and privacy implications.

Sending camera data to the cloud is inacceptable for most situations where images contain a person’s biometric footprint which would require consent and/or where the storage of such information is regulated.

Deep learning has a size problem*

The above constraints are forcing hardware and software ecosystems to move the video understanding from the cloud towards the edge. This is not surprising given that there are less than 100 million cloud servers, but there already exists 3 billion mobile devices and more than 150 billion embedded devices*

Deep learning has a size problem. https://heartbeat.fritz.ai/deep-learning-has-a-size-problem-ea601304cd8

On the hardware side…

Nvidia has launched several GPU enabled edge computing devices such as Jetson Nano and Jetson Xavier.

Google has built Coral to commercialise TPU (tensor processing unit) enabled developer kits.

Intel has launched Movidius Neural Computing stick, a deep learning accelerator hardware with the size of a USB key.

Huawei has launched GPU enabled Atlas computing devices, specifically tailored for computer vision on the edge.

On the software side…

The data science ecosystem is also moving swiftly towards smaller and more energy efficient models.

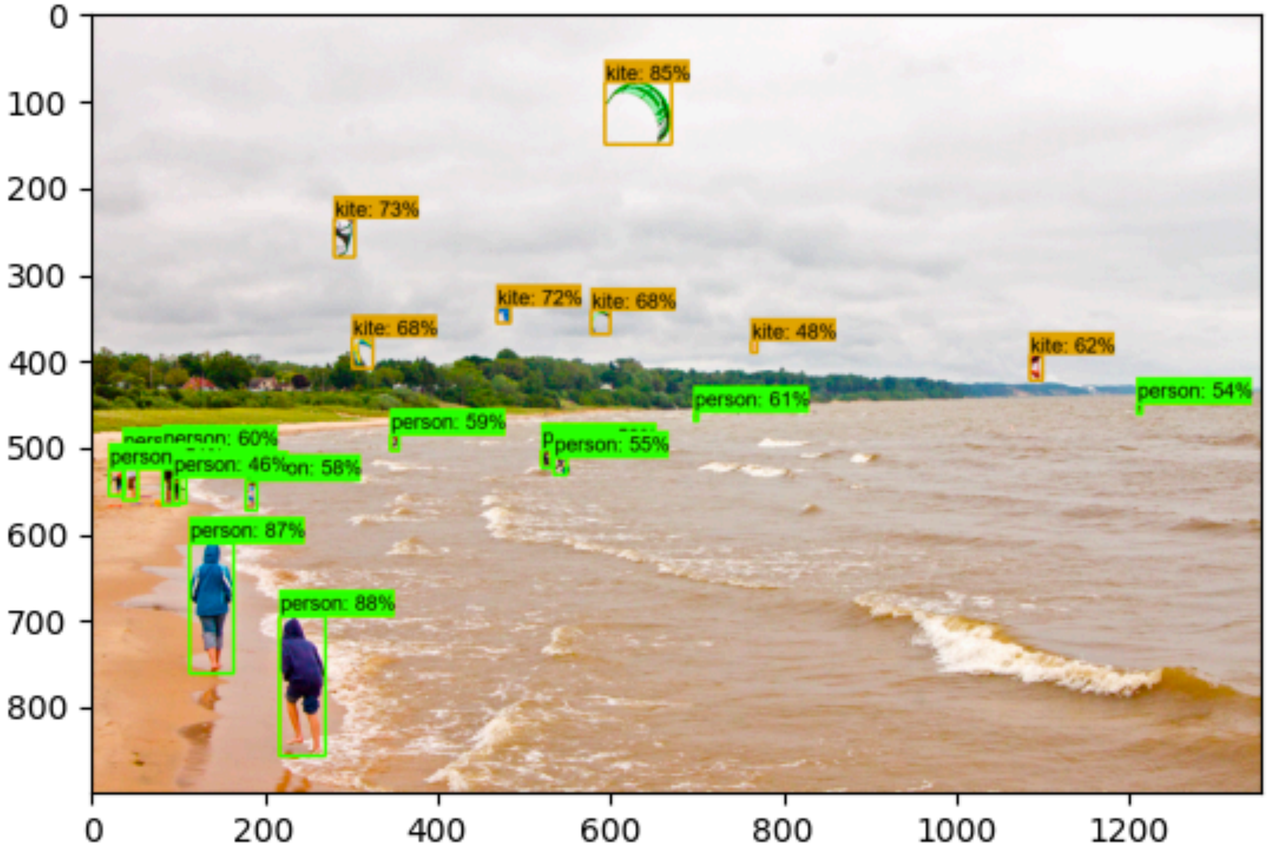
In 2019, TinyML (tiny machine learning) foundation has been built with support from both venture capital and IT companies such as Google and Qualcomm

Again in 2019, Google launched a framework for machine learning specific to web browsers and acquired MediaPipe, a company specialised in machine learning in mobile devices.

Very recently, Harvard University launched a professional certificate on EdX platform for machine learning on mobile devices.

Using tinyML techniques, MIT researchers reduced carbon footprint of training deep learning models by more than a thousand times

Object detection (2D)



Localisation + Classification
(See Yassine Kriouile's presentation**)**

Video (action) recognition

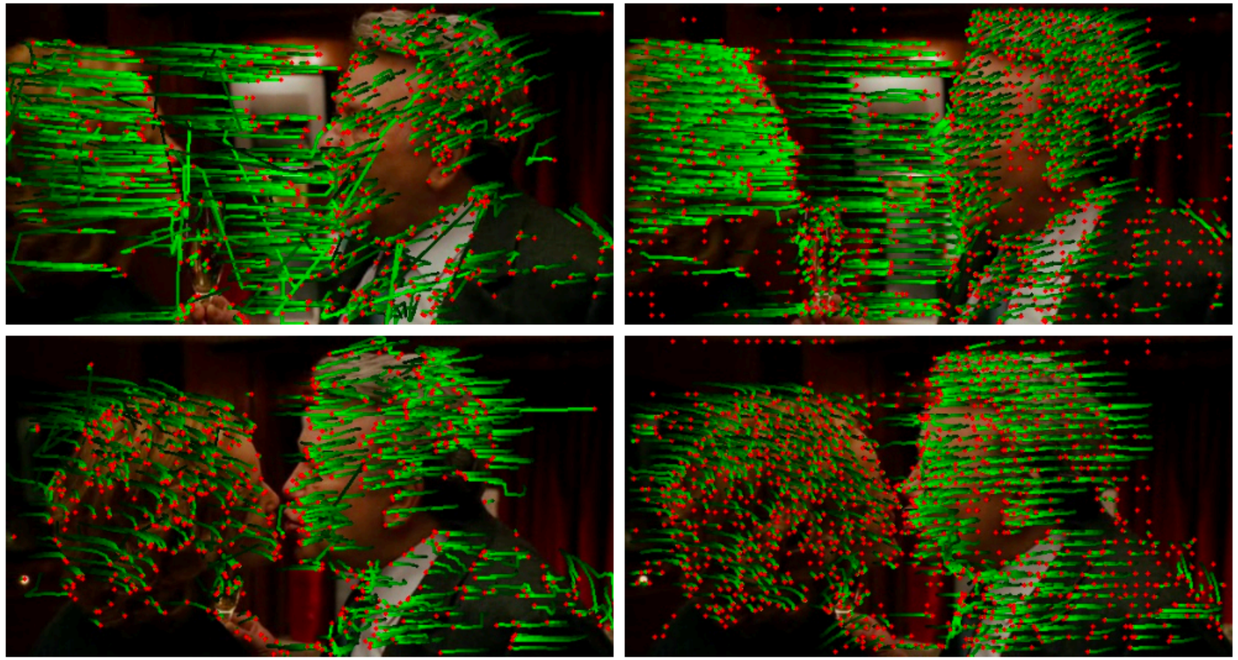
From spatial to spatio-temporal

Detecting the presence of, e.g. a person, on a single image is several orders of magnitude easier than understanding what does the person do (stand, talk, dance, run ...) in a sequence of images.

Optical flow features

"distribution of apparent velocities of movement”

Pre-deep learning era approaches such as the Improved Dense Trajectories still hard to beat

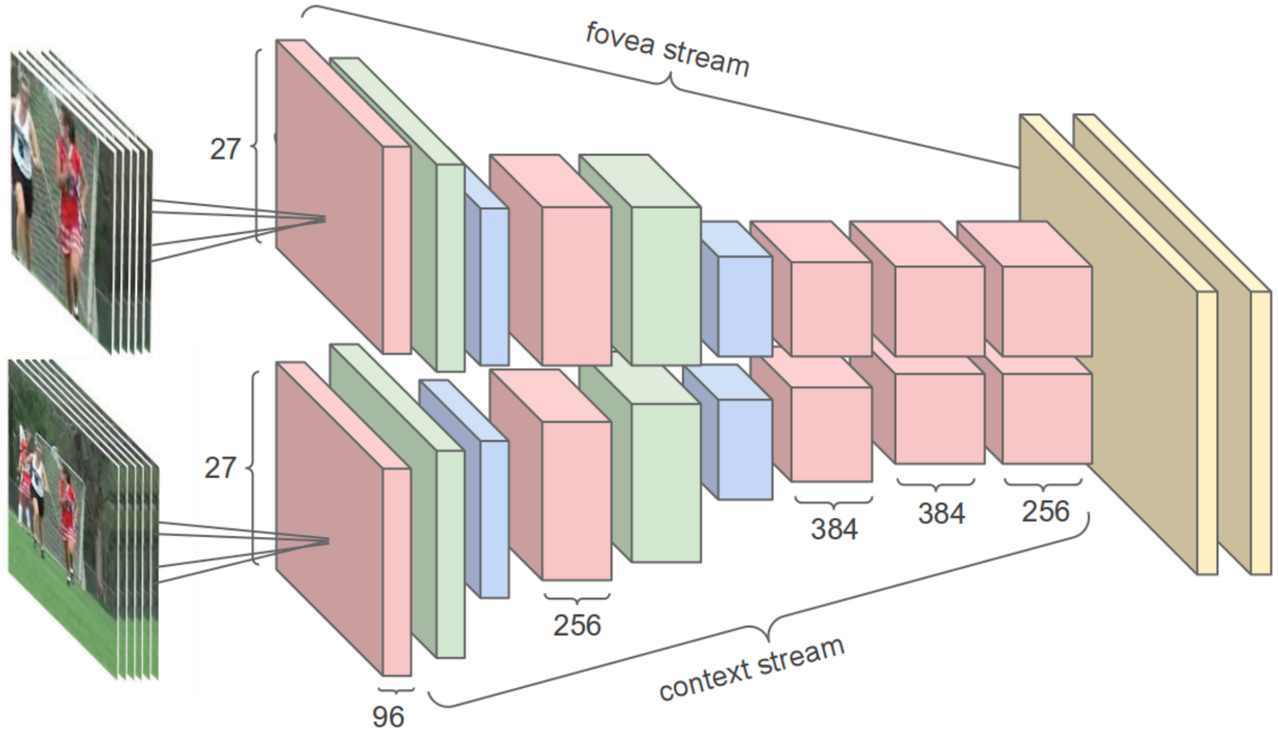


KLT Dense trajectories

Figure 1. A comparison of the KLT tracker and dense trajectories. Red dots indicate the point positions in the current frame. Dense trajectories are more robust to irregular abrupt motions, in particular at shot boundaries (second row), and capture more accurately complex motion patterns.

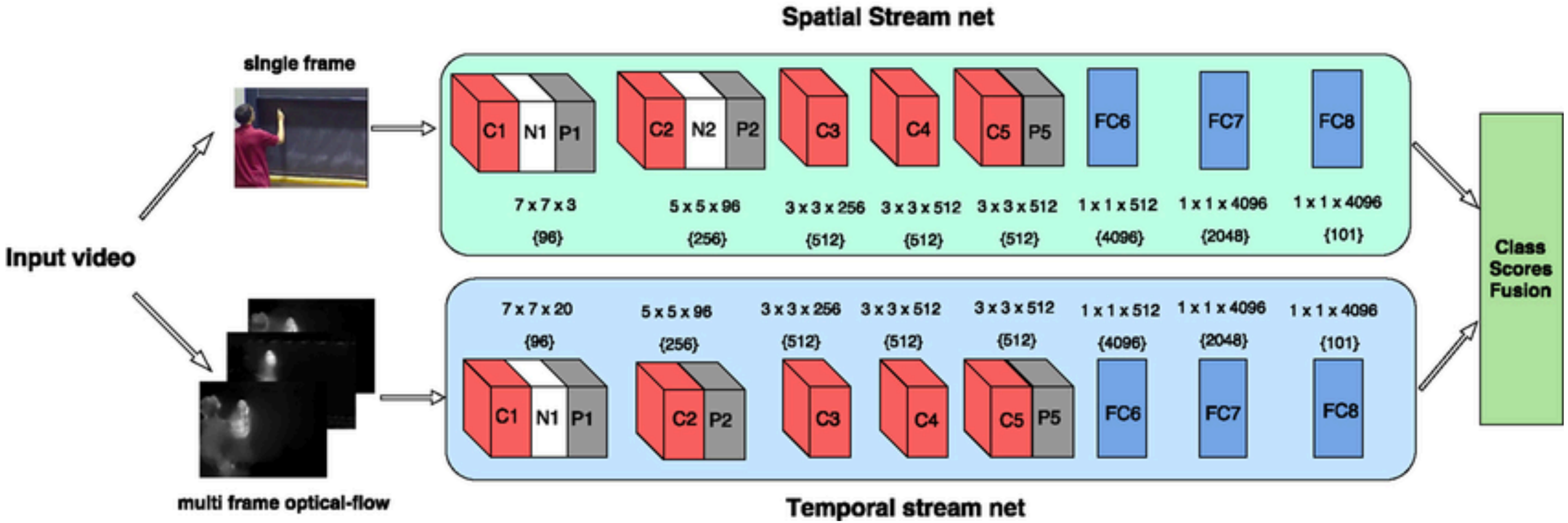
Action Recognition by Dense Trajectories, Wang et al. 2011)

DeepVideo (2014)



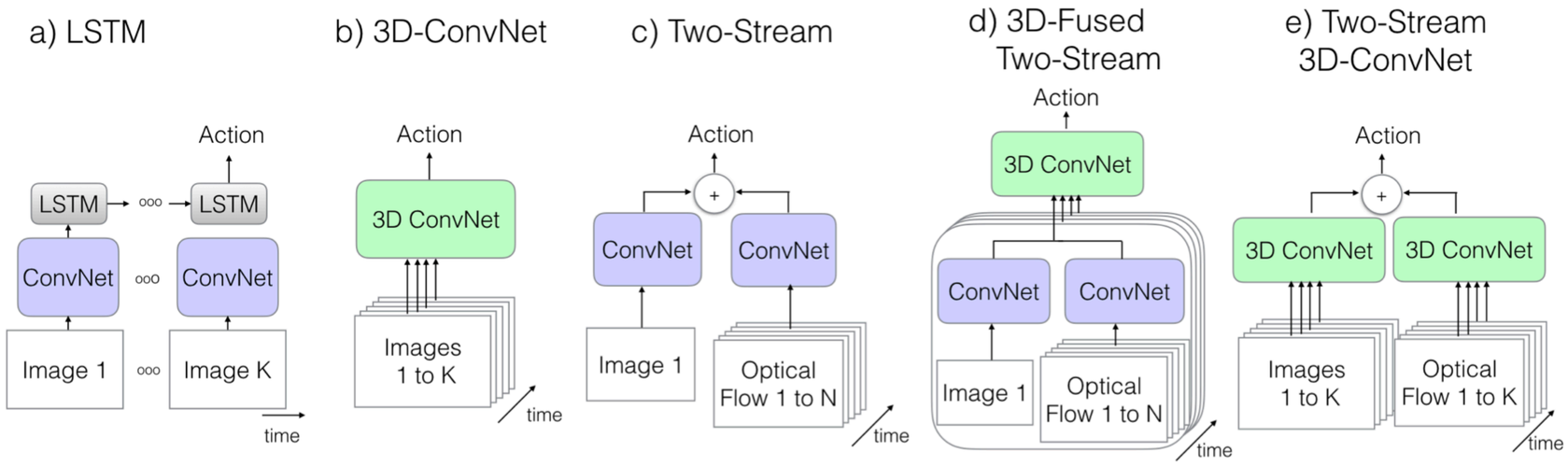
Early work by Karpathy et al (2014) following phenomenal success of convolutional neural networks. About 20% less accuracy compared to IDT model (https://cs.stanford.edu/people/karpathy/deepvideo/)

Two stream networks (Simonyan and Zisserman 2014)



Temporal learning is achieved through a stream of optical flow features. Accuracy on par with IDT model.

I3D networks (Carreira and Zeisseman, 2017)



Major contribution: Inflated 2D ImageNet features to 3D, effectively eliminating the computational burden of training from scratch the 3D computational filters.

- End to end trainable -

- %97 accuracy on UCF101 and %80 on HMDB51 -

3D CNN era for video understanding

- Several other architectures have followed these seminal works
- SlowFast networks (Feichtenhofer et al. 2019)
- Temporal Shift networks (Lin et al. 2019)
- Motion squeeze (Kwon et al. 2020)
- …

The increase in accuracy came, however, at the expense of efficiency

- The number of model parameters (hence, the required memory size), and the associated computational cost for training and using these models, increased significantly.
- To give an example, storing optical flow images for Kinetics400 requires ~4TB memory.
- When training such models, the read/write operations becomes the bottleneck during training a two-stream architecture, instead of computational power (e.g. GPU) - wasting computational resources and slowing down experimental cycles.
- 3D CNN methods, on the other hand, remains hard to train and to deploy effectively.