NEUROMORPHIC VISION SENSING FOR CNN-BASED ACTION RECOGNITION

Aaron Chadha, Yin Bi, Alhabib Abbas, Yiannis Andreopoulos

Department of Electronic and Electrical Engineering
University College London, London, U.K.

Action recognition

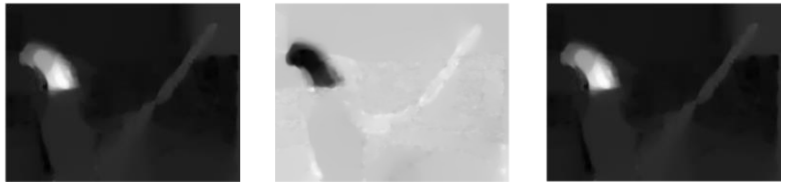
- Task: Classify video sequences based on their constituent action

Query: 

⇨ Recognition System ⇨

Action: *Archery*
Swimming
Running
Cycling
…

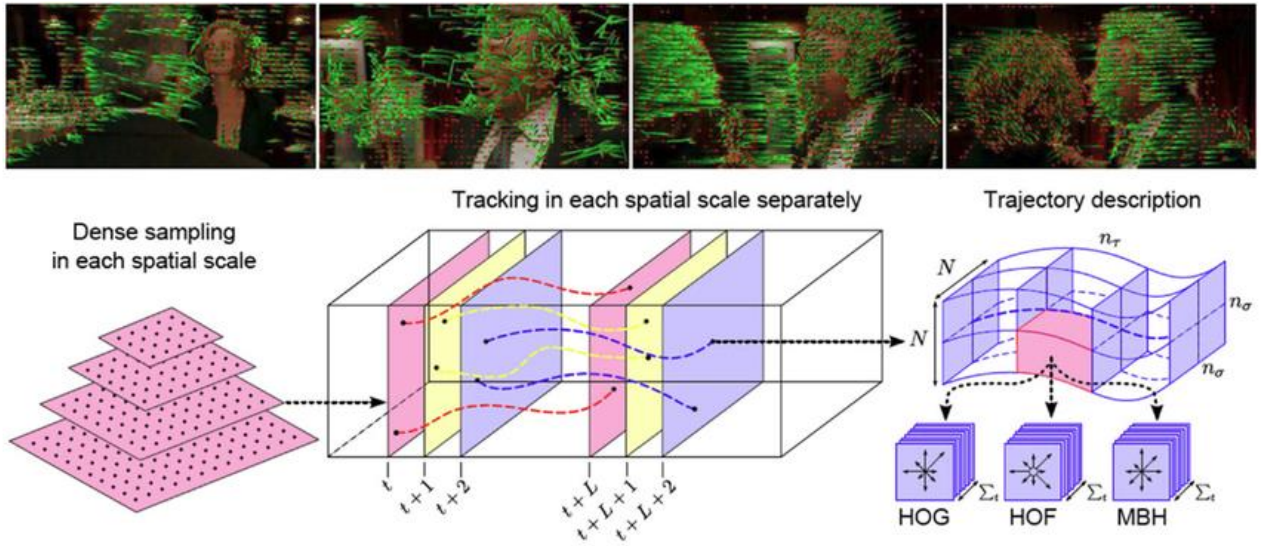
- *Additional modalities* are typically used to supplement RGB frames, such as optical flow:



time

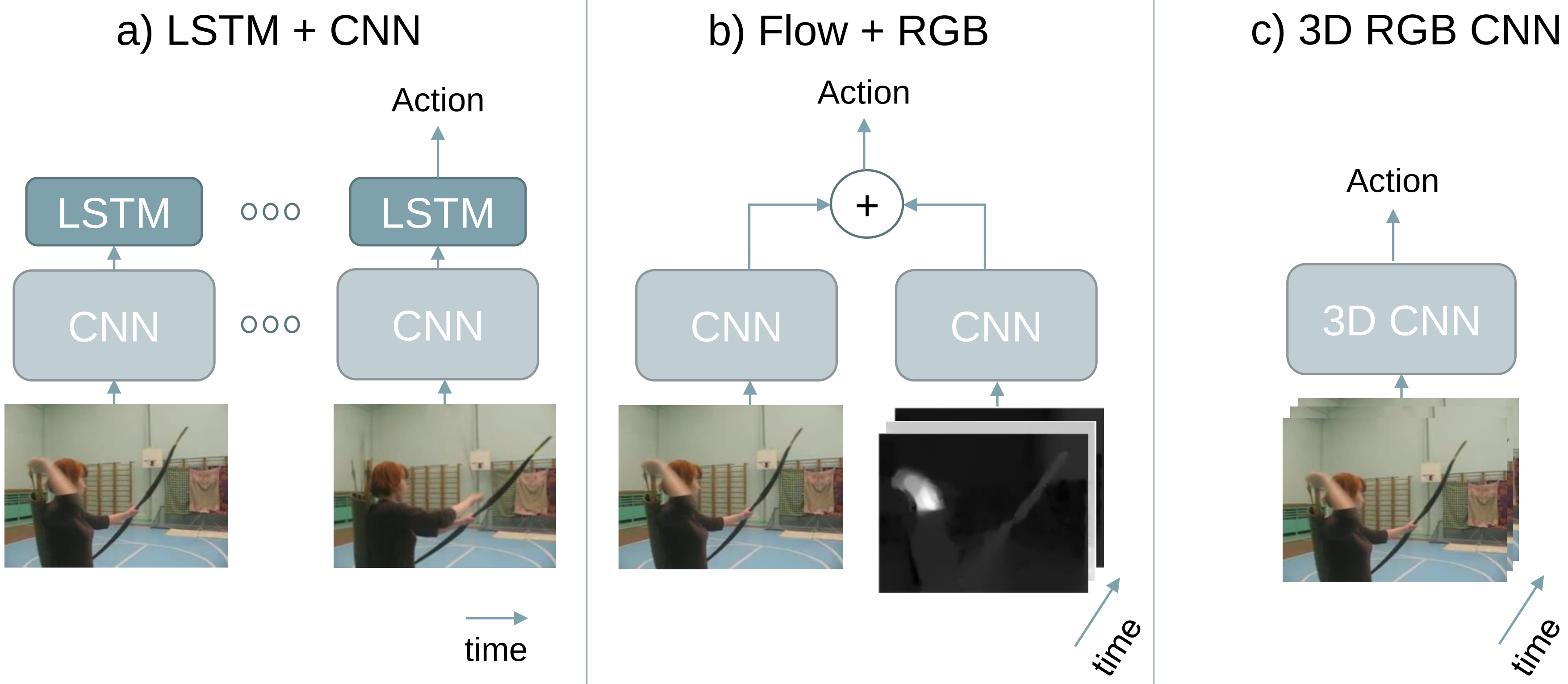
Background: Action recognition

- Before deep learning – dense trajectories using optical flow:



Background: Action Recognition

State-of-the-art deep learning methods:



[Donahue et al., 15] [Zisserman et al., 14] [Tran et al., 15]

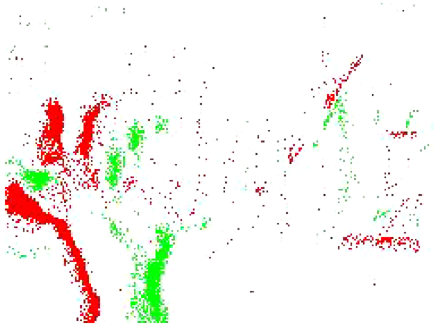
Active Pixel Sensing

- Motion vectors and optical flow both require active pixel sensing (APS) video
- APS video is cumbersome for multimodal frameworks due to:
  - Limited framerate
  - Calibration problems under irregular camera motion
  - Blurriness/distortion with varying illumination
  - High power requirements

Neuromorphic Vision Sensing

- Neuromorphic Vision Sensing (NVS) cameras emulate the photoreceptor-bipolar-ganglion cell information flow.
- Their output consists of asynchronous ON/OFF spike events
- The events are recorded as tuples indicating spatio-temporal position and polarity

(a) APS video



(b) NVS video

(c) NVS camera

APS vs NVS

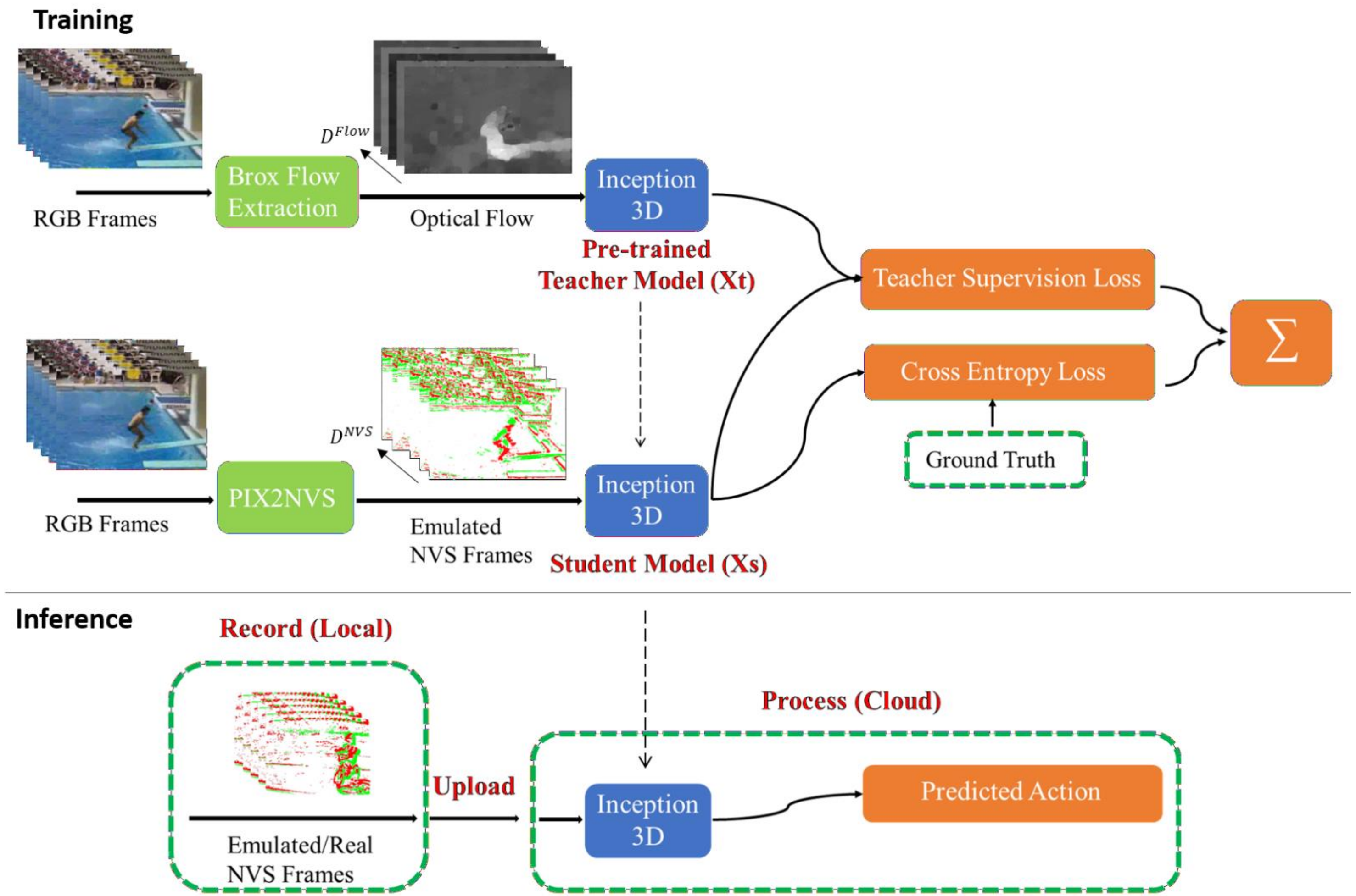
- Advantages of NVS over APS:
  - Much higher framerates (up to 2000 FPS)
  - Lower power consumption (on the order of 10mW)
  - More robust to distortions

- Disadvantages of NVS over APS:
  - NVS events are typically sparse and more difficult to train on
  - There is currently a scarcity of labelled NVS data for training compared to APS

Our Proposal

- We want to *reduce the acquisition and sensing complexity* in the multimodal framework
- We propose to replace the APS modalities with NVS frame representations
- To circumvent the disadvantages of NVS:
  - *Difficulty in training:* Train with supervision from optical flow data in a teacher-student framework
  - *Scarcity of real labelled data:* Embed an NVS emulator (PIX2NVS) into the learning framework for NVS emulation from APS video

Teacher-Student Framework



Teacher-Student Framework

- For a distribution of student NVS frame volumes $\mathbb{V}_s$, teacher flow volumes $\mathbb{V}_t$ and labels $\mathbb{Y}$:

$$L = \overbrace{-\beta\mathbb{E}_{(\boldsymbol{v}_s,y)\sim(\mathbb{V}_s,\mathbb{Y})}\sum_{k=1}^{K}1_{[k=y]}\log(p(\boldsymbol{v}_s)_k)}^{\text{Standard weighted cross entropy loss with labels}} - \underbrace{\alpha T^2\mathbb{E}_{(\boldsymbol{v}_s,\boldsymbol{v}_t)\sim(\mathbb{V}_s,\mathbb{V}_t)}\sum_{k=1}^{K}q(\boldsymbol{v}_t,T)_k\log(p(\boldsymbol{v}_s)_k)}_{\text{Teacher-student weighted cross entropy loss}}$$

- Accuracy: Without teacher - 71.0%; With teacher - 77.0%

Results

- Two stream accuracy vs state-of-the-art:

Method	Σ GFLOPs	UCF-101	HMDB-51
inc. optical flow			
Two-Stream [70]	150	88.0	59.4
3D Conv Fusion [71]	153	92.5	65.4
Action-VLAD [72]	-	92.7	66.9
ST-ResNet [153]	-	93.4	66.4
Two-Stream I3D [84]	648	97.8	80.9
no optical flow			
EMV-CNN [81]	150	86.4	-
CoViAR[82]	110	90.4	59.1
C3D [66]	385	82.3	51.6
Res3D [154]	193	85.8	54.9
I3D (RGB only)[84]	324	95.1	74.3
LTC (RGB only) [155]	308	82.4	-
Proposed, NVS (emulated)-RGB CNN	84	89.0	62.0

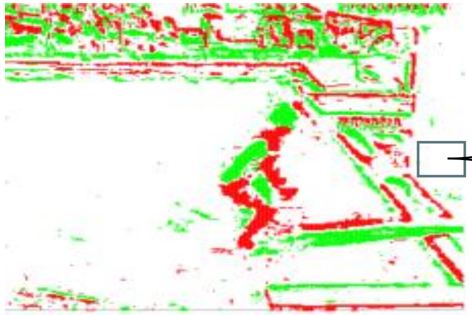
- **Note:** To minimize the APS bottleneck we infer on a single shot of 8 RGB frames at maximum motion activity

Results

- We present an efficient multimodal framework for NVS-based action recognition
- Training with optical flow supervision improves accuracy by 6% on a single shot of 8 frames
- We achieve 89.8% on UCF-101 with less than 100 theoretical GFLOPs for CNN processing
- However, accuracy is reported on emulated NVS events; we want performance to generalize better to real NVS events

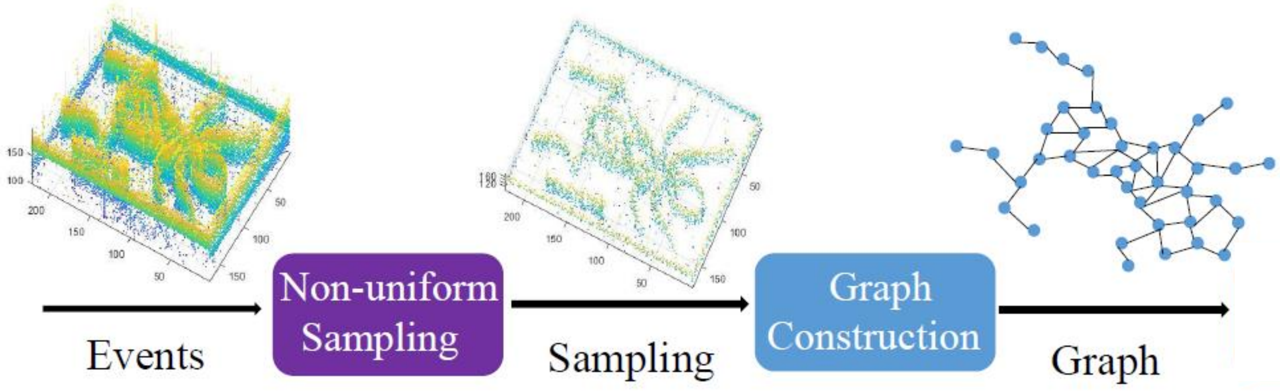
Further Work: Graph-based Object Classification

- An NVS frame and its pixel value:

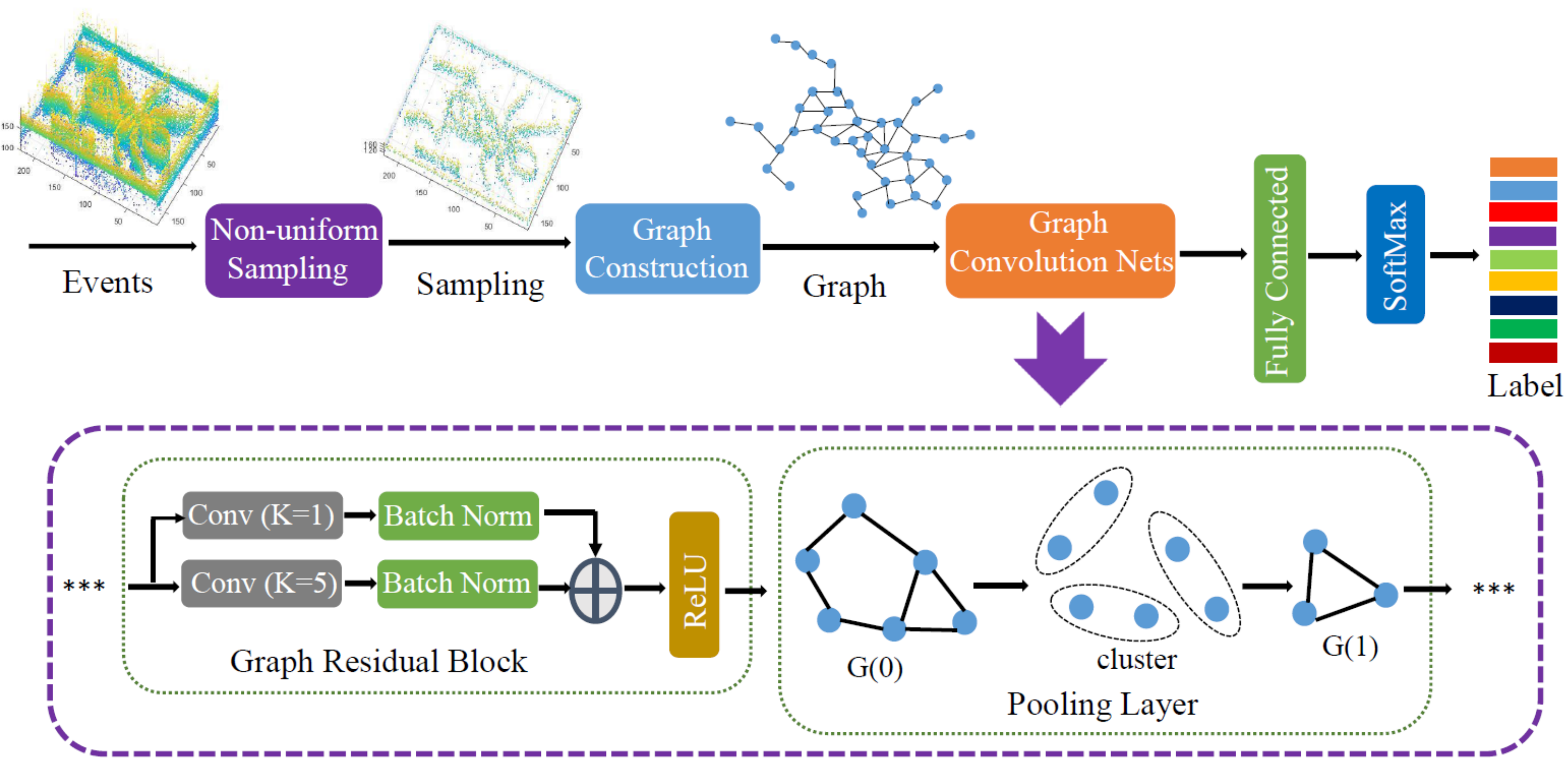


0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

- Compact graph representation:



Further Work: Graph-based Object Classification



Further Work: Graph-based Object Classification

- Top-1 accuracy of our CNNs w.r.t. the state of the art & other graph convolution networks on object classification:

Model	N-MNIST	MNIST-DVS	N-Caltech101	CIFAR10-DVS	N-CARS	ASL-DVS
H-First [46]	0.712	0.595	0.054	0.077	0.561	-
HOTS [29]	0.808	0.803	0.210	0.271	0.624	-
Gabor-SNN [30, 42]	0.837	0.824	0.196	0.245	0.789	-
HATS [56]	**0.991**	0.984	0.642	0.524	0.902	-
GIN [62]	0.754	0.719	0.476	0.423	0.846	0.514
ChebConv [17]	0.949	0.935	0.524	0.452	0.855	0.317
GCN [27]	0.781	0.737	0.530	0.418	0.827	0.811
MoNet [37]	0.965	0.976	0.571	0.476	0.854	0.867
G-CNNs (this work)	0.985	0.974	0.630	0.515	0.902	0.875
RG-CNNs (this work)	0.990	**0.986**	**0.657**	**0.540**	**0.925**	**0.901**