



# Iris Groen - Day 1 lecture +

## Part 1

→ Hubel & Wiesel (conts :))

→ orientation tuning

→ retinotopic organisation

fMRI ← where

MEG, EEG ← when

! Alexia part appears when you learn to read and  
it's always in the same place of the brain-

→ hierarchical theory of object recognition



## Part 2

How human vision influenced computer vision?

→ Gabor filters

→ computer vision meets deep learning

→ solve invariant object recognition

→ ImageNet & AlexNet

Lecun et al, Nature (2015)

O'Reilly (2020) (?)

DiCarlo & Cox (2007)

\* The Words I See  
Fei-Fei Li

\* Krizhevsky, Suts... (2012)

\* Fukushima (1980)  
Neocognitron

\* JeNet, LeCun

→ poloclub GitHub CNN explainer

\* Lindsay (2020)  
computational neural networks

\* Zeiler & Fergus (2014)

\* Guckl & von Gennet (2015)

## Part 3

- Do CNN networks predict brain activity? (features)
- HCNN layers (?)
- Grouping similarity
- Networks seem to mimic human brain
- Brain-Score
- Algo nauts
- Neuroconnectionist research cycle
- object recognition & scene processing

Lindsay (2020)

Jenkins et al. (2014)

Kriegeskorte (2014)

Kriegeskorte & Mur (2012)

King & Green et al (2018)

Doenig et al. (2022)

Green, Silson & Baker (2017)

## Part 4

How computer vision & human vision (still) different?

- retinal image vs cortical map.
- transformers are better than neural networks
- Neuroscience & computer vision work together, they feed each other
- global statistics extraction, feedback/recurrent connections, peripheral sampling, ecological training, etc. ? ? ?

\* Da Costa et. al. (2021)

\* Cheung, Wells ...

\* Müller, Scholte & Green (2020)

\* Geirhos et. al. (2018)

\* Geirhos et. al. (2019)

\* Papello, Moques et. al  
Simulating a Primary ...

→ An image worth 16x16 words

\* Tuli, Dasgupta, ... Are convolutional NNs more like com-vis.?

\* Connell et. al (2024)

# Andrew French - Day 1 Lecture 2

→ "no learning"?

\* Classical tracking

→ motion detection & tracking

→ individual vs. multiple targets

→ pixel-level motion

→ optic flow

→ background models } detect movement

not tracking ::

\* if we can model the motion, we can predict objects' future locations.

→ HCI, robotics, surveillance, medicine...

flow-field ↔ quiver plot

→ motion difference & background subtraction

two basic approaches  
in motion detection

motion detection

## - multiple object tracking -

\* Markov Chain Monte Carlo tracking

\* Metropolis-Hastings' algorithm

→ chain of predictions

→ Khan 2. et.al. (2004) An MCMC...

\* social extensions

→ behaviours

→ sharing motion information

## - tracking -

\* uncertainty

\* we want to be able to PREDICT where our target will be, and UPDATE our guess with a measurement.

\* Kalman Filter

\* Particle filters

→ real tracking is often multimodal

→ Isard & Blake (1998) Condensation

\* appearance model(?)

\* contour tracking

\* mixed-state condensation

→ transition probability

\* behavioural recognition ?

\* the curse of dimensionality

still relevant in the  
deep-learning era

# Michael Way - Day 2 Lecture 1

## Ego-centric Vision - Making Sense of the First-Person Perspective

\* Steve Mann, WearCam, WearComp  
applications in Personal Imaging

\* Wearable devices

\* Not limited to humans

↳ dogs, robots, shopping cards

\* Blue Sky Outcomes

↳ robotics to understand surroundings

↳ personal assistants

- remind you what you forgot
- help you fix things
- remind you a recipe's next step etc.

↳ healthcare for assisting people

↳ ethic & privacy is an issue!

↳ Hololens, Google Glass, etc.?

Ego4D dataset

GTEA, BEOLD, GTEA Gaze, UT Ego,  
ADL, EGTEA Gaze +, EPIC-Kitchen-117  
Chades-Ego, EPIC-Kitchen-100

↳ combining every dataset

Ego4D

Past — Present — Future

- Episodic memory
- Hand object interaction
- Audio visual diarization
- Social interaction
- forecasting

### HD-EPIC

Digital Twin of kitchen videos

- narration, fine grained annot
- audio annotation, object detection
- object - fixture assignment

• Blind-Language: Llona 1.2, Gemini Pro

• Video-Language: Video Llona 2, LogVA, Gemini Pro

### Ego-Centric Tasks

• video understanding tasks  
↳ action recognition, VQA

- captioning, retrieval, grounding
- audio localization, activity & object recognition

→ EgoVis workshop

HierVL : Learning Hierarchical Video-Language Embeddings

- EgoVLP - previous study, not good (?)

- Assembly101 dataset (2022)

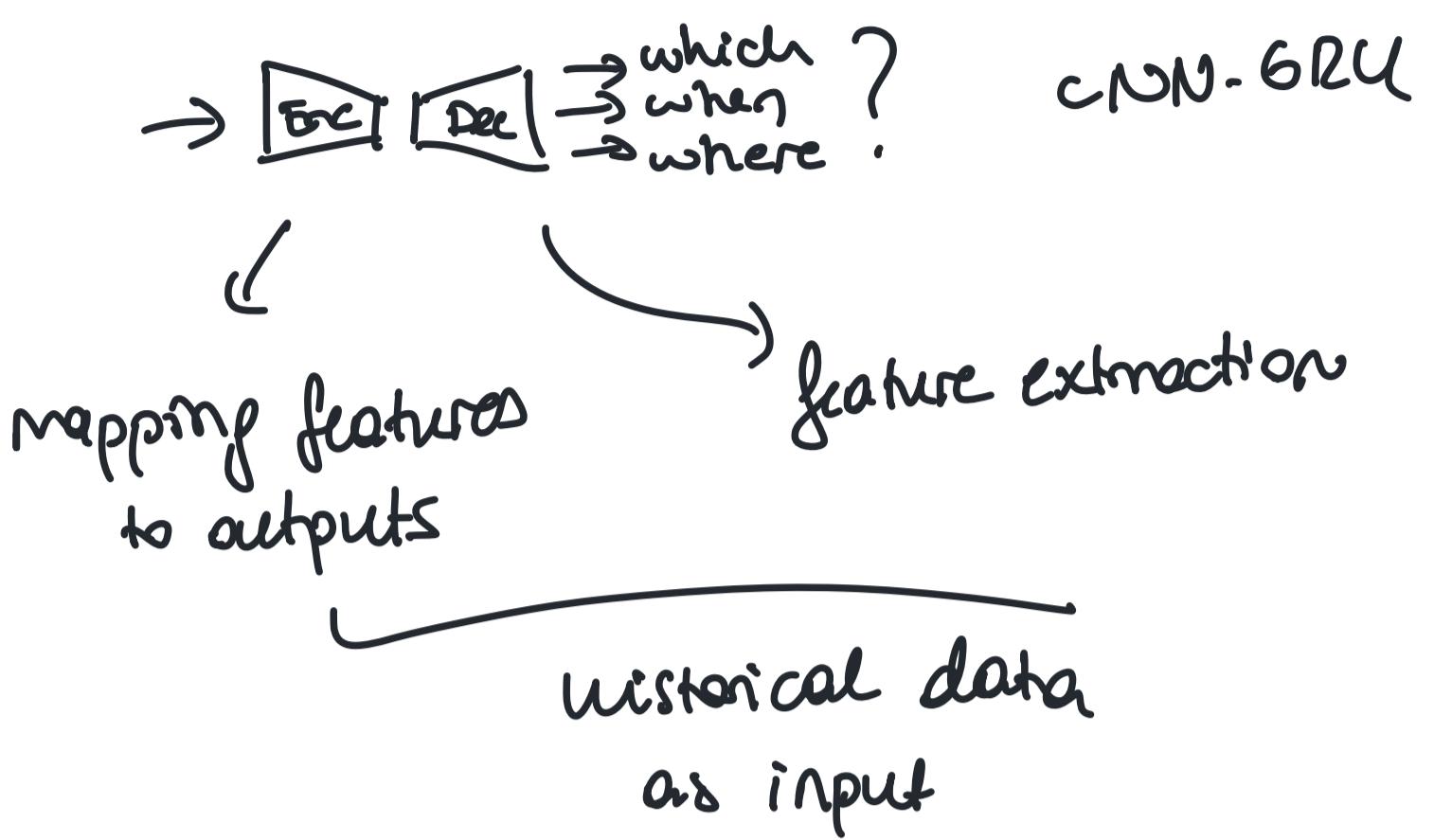
↳ assembly hands

↳ SVEgoNet : single view ego net

LEGO: Learning Egocentric Action Frame Generation via Visual Instruction Tuning

### \* CV + DL

which? → when? → where?  
trajectory tensor



→ used tensors instead of vectors  
to convey the uncertainty

\* FC, CNN, GRU, TAN

\* binary cross-entropy loss

Prof. Victor Sanchez - Warwick

Day 2 Lecture 2

CLIP: Contrastive Language-Image Pretraining for Video Analytics

\* Deepfakes, IAPR head of:

\* SIPLab,

• multimedia forensics & biometrics  
↳ platform provenance

Project: Multicamera Trajectory Forecasting tracking via person re-identification

Shyles, Guha, Sanchez (2021) Pattern Analysis...

Nootker, Sanchez (2025) WACV

### TinyFaces project

MTCNN + IFS

Leyva, Sanchez, Li (2018)

Leyva, Shen, Bahadur ... (2025)  
(Automatic Face & Gesture Recogn.)  
IEEE

### TinyFaces IFS detector

↳ Fisher Vector

\* IFS to FVs → reduces the # of operations

Kulshreshtha & Guha (2018)  
(ICIP)

## M-EVA dataset

Multiview extended video with activities

Qi, Tan, Yao, ... (2023) - Yolo5Face

Yu, Huang ... (2024) - YoloFacev2

TinyFaces is still a challenge to be solved!

Recall: CLIP is proposed to be used in downstream tasks  
→ WebImageText (WIT)

## CLIP for segmentation

Lüddecke, Ecker (2022) → CUF

DACE: Distance-Aware Cross-Entropy Loss

HuggingFace: Spaces/Yiming-M / CLIP-EBC (?)

## Crowd Counting

→ rely on annotated 2D coordinates of individuals' head centres in images.

Ma, Sanchez, Guha (2022) → CLIP

Ma, Sanchez, Guha (2025) → ICME (CLIP-EBC)

→ Reform crowd counting as a classification task



they introduced this idea to CLIP.

\* zero-shot learning: can we facilitate?

↳ natural language supervision

\* contrastive learning

Radford, Kim, Hallacy ... learning transferable visual models from natural language supervision

\* symmetric crossentropy loss

Attention is all you need

Fine-tune of CLIP → ResNet50: attention pooling  
Vision Transformer

## From Paragraphs to Pixels: LLMs in Video Understanding

@andrewjohngilbert.github.io

## \* Video understanding

- ↳ long range temporal dependencies
- ↳ ambiguous/overlapping events
- :  
lots of challenges in this area.

YouTube: WhoDunnit? It is easy to miss  
stuff you're not  
looking for.

## \* Video Understanding with Large Language Models: A Survey, Yuanlong Tang et al.

## \* Unsupervised Learning of Visual Representations Using Videos, ICCV 2015

## \* Shuffle Learn ECCV (2016)

## \* Self-Supervised Video Rep CVPR 2017

## \* Arrow of Time, CVPR 2018 (?)

## \* ECCV 2018 - Tracking Inference by Colorizing Videos

## \* VideoMAE, NeurIPS 2022

## \* MOFO:, NeurIPS 2023 workshop

## \* CVPR 2023 - Learning Video ..

## \* Vid2Seq CVPR 2023

## \* VAST NeurIPS 2023, Sijian Chen

## \* FILS : Self-supervised Video Feature Prediction...

## \* Fine-Grained audible video description

## \* SWIN-BiT (?)

## \* AutoAD : CVF 2023

## → ? Donte AD ?

gap between pixels &amp; n1 reasoning

Surveillance

Healthcare

Autonomous Vehicles

Entertainment

Impact &  
ApplicationAI Storytelling

Art in Storytelling.

## \* Visual Agents in Software

- ↳ unreal engine
- or Game Engines mostly

} Visual Interface  
Agent.

## \* MAGMA : Multimodal Agentic Foundation

Tang et al. 2025

## \* UI-TADS, 2025

## Audio-Visual Fusion (?)

## \* Romanovna... "OWL..." CVF 2023

## \* Egocentric audio-visual object localization, CVF 2023

## \* Self-supervised moving vehicle ...

## \* Epic-fusion : Audio visual temporal binding for egocentric action recognition.

→ Dataset: Youtube8MTHUMOS

Local (?)  
Learnable Getting  
Cross Attention

## \* Multi-Resolution Audio-Visual Feature Fusion...

## \* DEL: Dense Event Localisation...? 2025

→ Action Quality Assessment

→ Scale to this.

ECCV 2022

with temporal  
padding transformer

## \* learning to score Olympic Events

# Oisin Mac Aodha - U. of Edinburgh

## Day 2 Lecture 6

### Representation learning

- \* bad vs good features
- Wei Koll et.al. Concept Bottleneck Models, ICML 2020

\* success depends on data representation

↳ what is good representation?

- compact (minimal)
- explanatory (sufficient)
- disentangled (independent factors)
- interpretable
- make subsequent problem easier

Eli Cole, Caltech (slide credits)

\* Transfer learning (ImageNet)

- Bengio et al. Representation Learning: A Review and Perspectives

credit: Justin Johnson eecs498

### Self-Supervised learning

"Most of human and animal learning ... " ~Yann LeCun  
(On true AI)

\* How to learn features from the unlabelled data?

\* Papers with Code: Self Supervised Image Classification

\* Effective methods for learning representations.

- Gidaris et.al. Unsupervised Repr. Learning... ICLR 201

- Weng & Kim, Self-Supervised learning, NeurIPS 2021

Self-prediction

SimCLR (2018)

- Chen et.al. A Simple Framework for Contrastive learning ...

• Github: google-research/simclr

@sthalles/github-afew-words-on-representation-learning

\* Once trained, the representation can be transferred to other tasks.

### Multi-modal Contrastive learning

- Radford et.al. Learning Transferable Visual Models from NL Supervision ICML 2021

\* zero-shot classification

\* there are limitations

↳ can require large batch size

↳ relies on "good" negative selection

↳ space of plausible augmentations can be dataset specific and must be defined in advance

\* masked language models

↳ BERT masked language models

\* vision transformers (ViT)

- Dosovitskiy et.al. An Image is worth 16x16 Words. ICLR

- He et.al. Masked Autoencoders Are Scalable Vision Learners CVPR 2022

- Zhang et.al. 2016, 2017

- Doersch et.al 2015

- Noroozi & Favaro, 2016 ?

- Noroozi et.al 2017

## Recent advances

- \* non-contrastive Siamese networks
- \* BYOL, SimSiam
- \* DINO
- Emerging Properties in Self-supervised Vision Transformers, ICCV 2021

\* DINOv2 builds on DINO & iBOT

- DINOv2 : learning Robust Visual Features...



## Limits of SSL

\* most SSL pretrained with ImageNet

- Cole et al. When Does Contrastive Visual Representation Learning Work? CVPR'22

\* dataset size matters

- ↳ amount of unlabelled data  
for pretraining
- ↳ amount of labelled data  
for supervised learning

\* iNAT

- Von Henn et al Benchmarking... (?)
- Zong, Andher... Self Supervised ...

Vision