Machine Learning for Computer Vision, a Case Study in Man Machine Interaction

Workshop ePicture this

TU Delft,

Dr. Ing. Boris Lenseigne
blenseigne@lrtechnologies.fr

Dr. Ing. Julia Cohen
jcohen@lrtechnologies.fr

Dr. Ing. Adrien Dorise
adorise@lrtechnologies.fr

Dr. Ing. Edouard Villain
evillain@lrtechnologies.fr

Organized by Penta projects:
2020005 Mantis Vision
2021004 Imagination

Summary

- Introduction
- Collecting data
- Experiments
- Results
- Challenges

Introduction

- Human Machine Interaction and assistive technologies → Enable Paraplegic users to use a computer
  - Consumer-grade hardware
  - Focus on vision based interaction → Face and gaze tracking
- **Objective: Control the mouse with face and eyes movements**
- Challenge: can we predict where the user is looking on the screen
  - Using the video stream from a webcam only (no light, no IR).
- A difficult problem with "traditional" image processing
  - Can an AI/Machine learning approach solve the problem ?
  - How should we do it ?

Nicole M. Bakker, Boris Lenseigne, Sander Schutte, Elsbeth B. M. Geukers, Pieter P. Jonker, Frans C. T. van der Helm, Huib J. Simonsz: *Accurate Gaze Direction Measurements With Free Head Movement for Strabismus Angle Estimation.* IEEE Trans. Biomed. Eng.60(11): 3028-3035 (2013)

Introduction

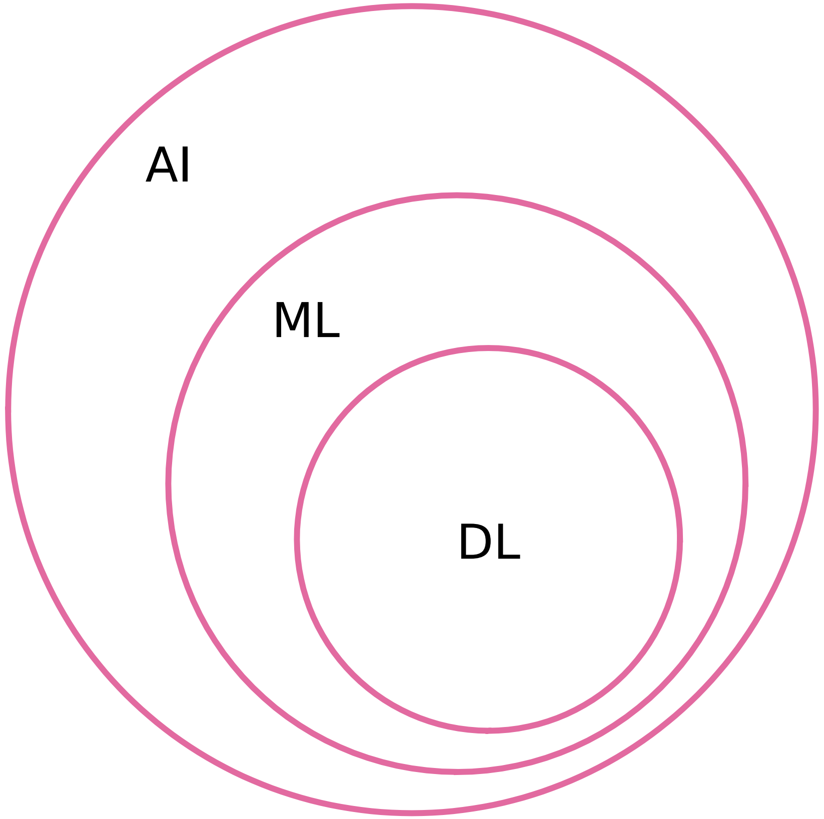
- Traditional approach to gaze tracking
  - Camera calibration
  - Eyes and face detection
  - 3D pose estimation
  - Eyes optical axis estimation
  - ...

- AI approach to gaze tracking
  - Gather data
  - Choose algorithm
  - Choose meta-parameters
  - Perform learning
  - Let the magic happen

- Expected benefits of AI
  - Task easily solved by humans
  - Unknown visual cues, difficult to model
    - AI → find regularities in the learning data

The computer must learn how it is used vs the user learns how to use the computer

What is Artificial Intelligence ?



- Artificial Intelligence (AI)

  Techniques that enable a machine to reproduce traits of human intelligence.

- Machine Learning (ML)

  The set of techniques from AI that enable a machine to produce a result without explicit programming.

- Deep Learning (DL)

  The set of techniques from ML in which the model is an artificial neural network.

- Data for Machine learning

1) Learning data set
2) Validation data set (during learning)
3) Testing data set

State of the art

- Work on AI and face
  - Face tracking
    - A. Rabhi, A. Sadiq and A. Mouloudi, "Face tracking: State of the art," *2015*
  - Facial expression
    - Song Zhenjie, "Facial Expression Emotion Recognition Model Integrating Philosophy and Machine Learning Theory", *2021*
  - Smyle mouse
    - Commercial mouse control via head & gesture software
    - Windows OS-only
    - US patents

- Why develop a new project on AI and face ?

  - Open-source software and consumer-grade hardware
  - All HMI methods in a unique application
  - Precise/micro-movements: not solved

  - **Final goal : Computer learn how user can control his computer on the contrary of user learn how to use his computer**

Collecting data

Inputs

- Video stream from the webcam
- Facial key features extracted by a face mesh detector

Model

Targets

- Cursor position on the screen

What type of data can we acquire/produce

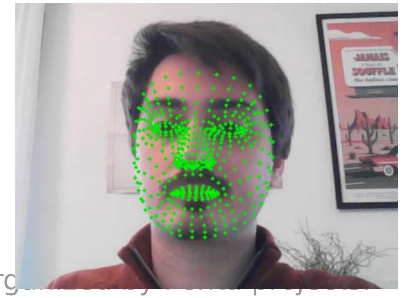
How to ensure clean data points?

How to synchronize inputs and outputs?

How many samples are required?

How to split between train/validation/test

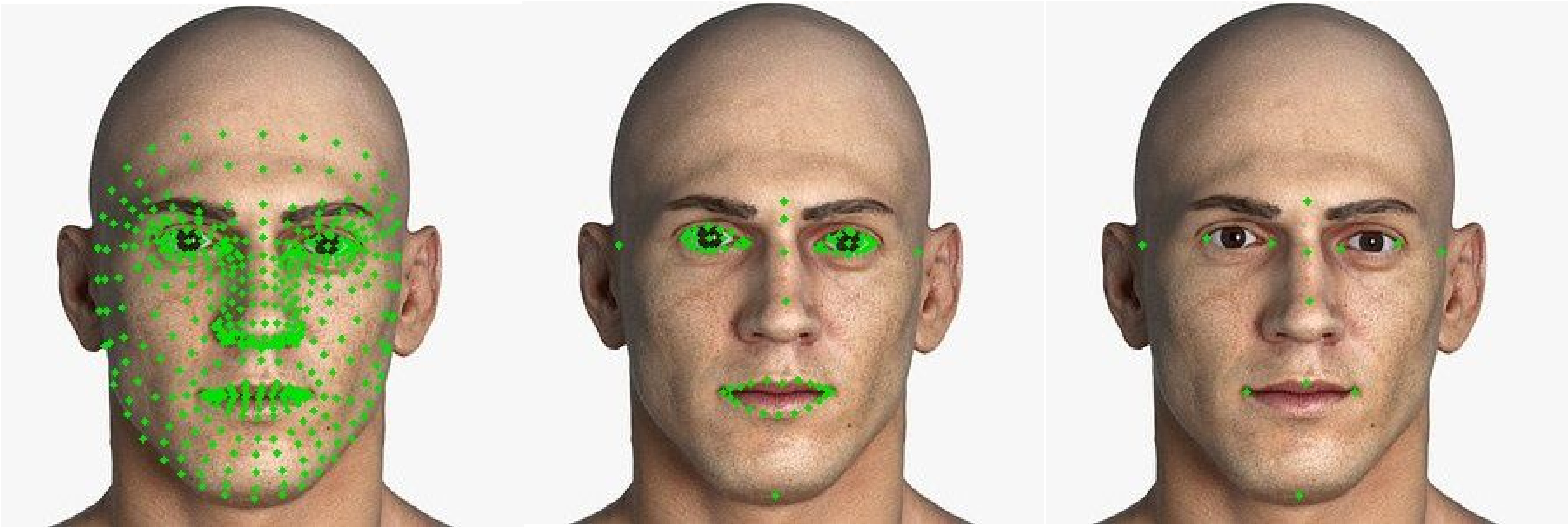
What scenario to follow?



Org
2020005 Mantis Vision
2021004 Imagination

Facetracker

- Use of an existing ML-basedface tracker (Mediapipe FaceMesh detector)
- Possibility to select the number of (x,y,z) points to use
- Already a diffucult problem with traditionnal IP



478 points 69 points 13 points

Kartynnik, Y., Ablavatski, A., Grishchenko, I., & Grundmann, M. *Real-time facial surface geometry from monocular video on mobile GPUs*. 2019

Experiments

How to select a model, optimization method and hyperparameters?

Input = facial points

- Machine Learning
  - Tweedie, Bayesian-Regressor, SGD, SVM, K-NN, Decision-Tree, Random-Forest, AdaBoost, GBoost
- Fully Convolutional Neural Networks

Input = image

- Convolutional neural networks

How to define the sequence's length?

Input = image sequence

- LSTMs neural networks
- Conv LSTMs neural networks

Results: baseline

- Direct mapping between the face and mouse
  - No artificial intelligence

- Model: face tracker using the point between the eyes as mouse controller
  - No calibration
  - Direct mapping between the face movements and the mouse movements on the screen

- Mean Absolute Error on test set: 19.5% (of the size of the screen)
  - with large variations between the videos
  - with large error at the beginning of every new video, reducing with time

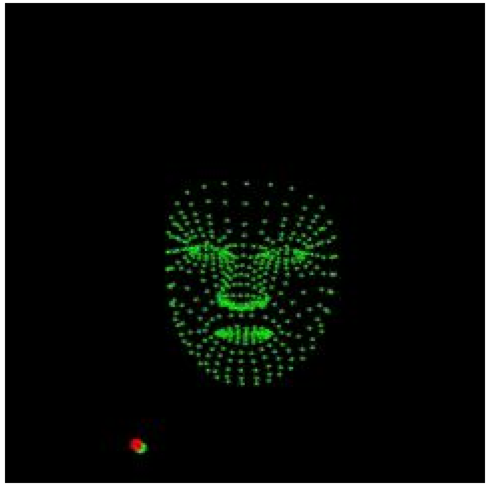
Results: Machine Learning

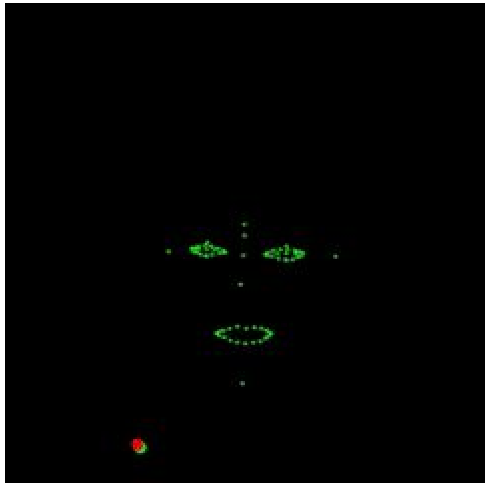
- Grid Search strategy performed on various methods for parameter research
  - Best performance was given by **Random Forest** and **AdaBoost**
- Prediction close to ground truth with one video
  - Face mesh features are relevant to predict cursor position
  - No noticeable difference when using a different number of features
- Precision drops when trying to generalise on multiple videos

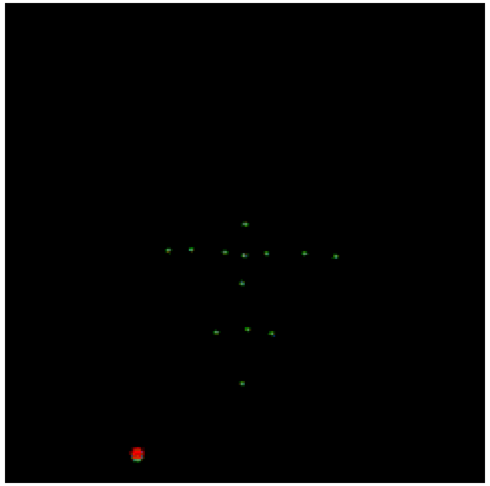
? *How to get the most important features?*

! *Multimodal output not common in ML*

Training and testing on same video: **Error 2%**



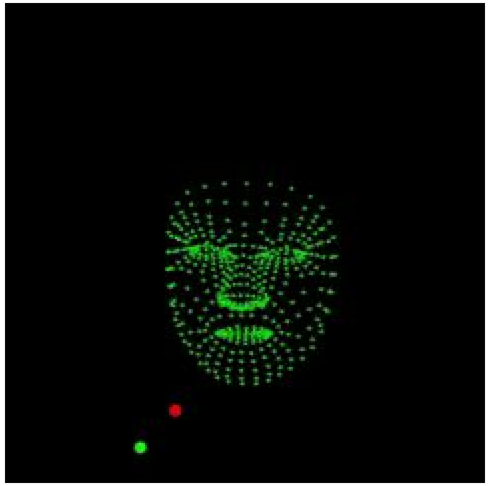




Generalisation

! *It can be long to train on 30.000+ samples*

Training on 18 videos: **Error 18%**

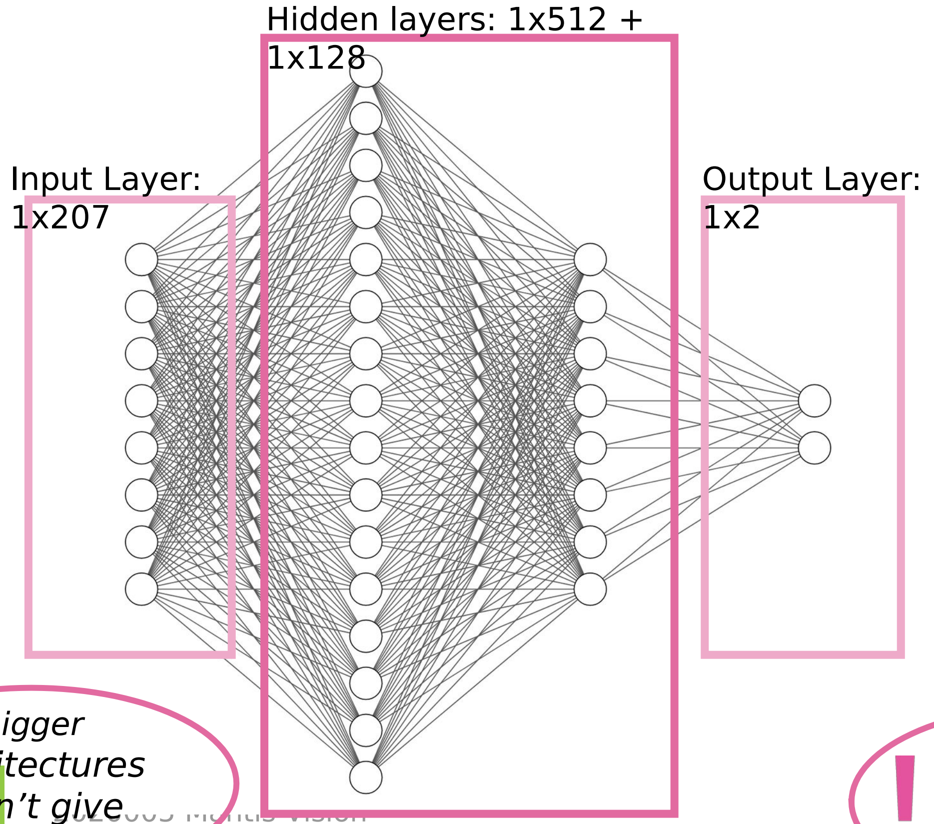


Target Prediction

Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001)

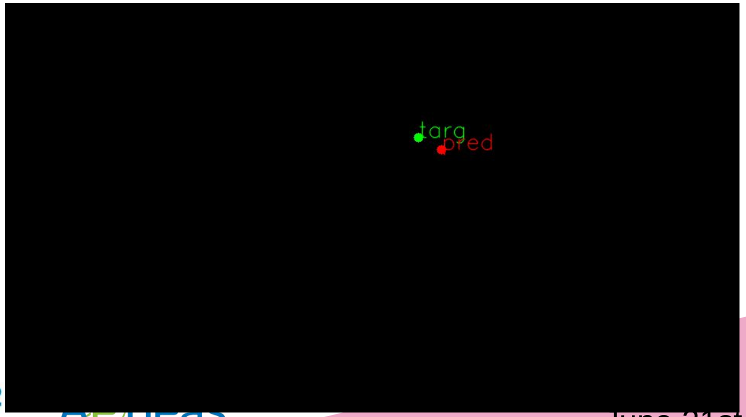
Results: Fully Connected Network (FCN)

- Fully connected model and training from scratch with minimum amount of data
- For one user, 3'30'' of data is enough to approximatively fit FCN model but not enough to predict new data



Training set : one video (3'30'') -> Error 4%

Testing set : same user (2'30'') -> Error 13.8%



! *Bigger architectures didn't give better results*

! *Beware of the constant prediction local minimum*

Results: Deep Learning CNN (ResNet-like)

- Convolutional architecture using the first layers of ResNet18
  - Weights initialization with pretrained ResNet18 on ImageNet
  - Training set: 19 videos of 3 different users (30428 frames)
  - Testing set: 12 videos of the same 3 users (32423 frames)
- Mean Absolute Error on test set:
  - Input is the full image (224x224px): 13.4%
  - Input is the cropped face (224x224px): 11.8%

Pre-training, fine-tuning or training from scratch?



2020005 Mantis Vision
2021004 Imagination

Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun, "Deep Residual Learning for Image Recognition", *2015*

Results: Deep Learning CNN (VGG-like)

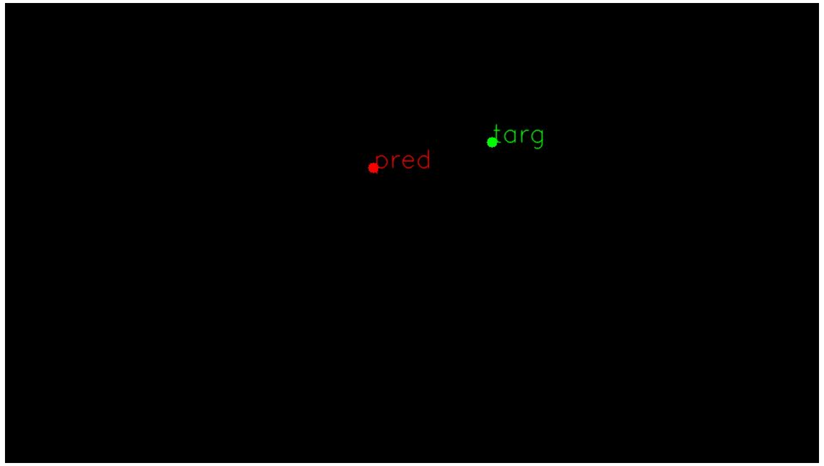
- Creating VGG-like model and training from scratch with minimum amount of data
  - Sequential model designed with 4 blocks of 2 convolutions and max pooling layers followed by 4 linear layers
    - Input : 224x224 - 3 channels images → Position (x, y) in the screen

Training set : one video (3'30'')
Error 0.8% (Mean Absolute Error)

? *Define minimum amount of data*

? *Multiuser scenario*

Testing set : same user (2'30'')
Error 11.5% (Mean Absolute Error)



- For one user, 3'30'' of data is enough to fit CNN model but not enough to predict for new data

Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *2015*

Results: Temporal models

- Long Short-Term Memory (LSTM) networks are used ad temporal models
  - The input is now a time sequence
  - Face mesh detector used as features
  - Combination of CNN and LSTM for image features is a work in progress
- First results do not show a sign of improvement compared to previous models
  - Training error = 5.2%
  - Test error = 18.4%
- Data collected were not designed for temporal models!
- Work in progress

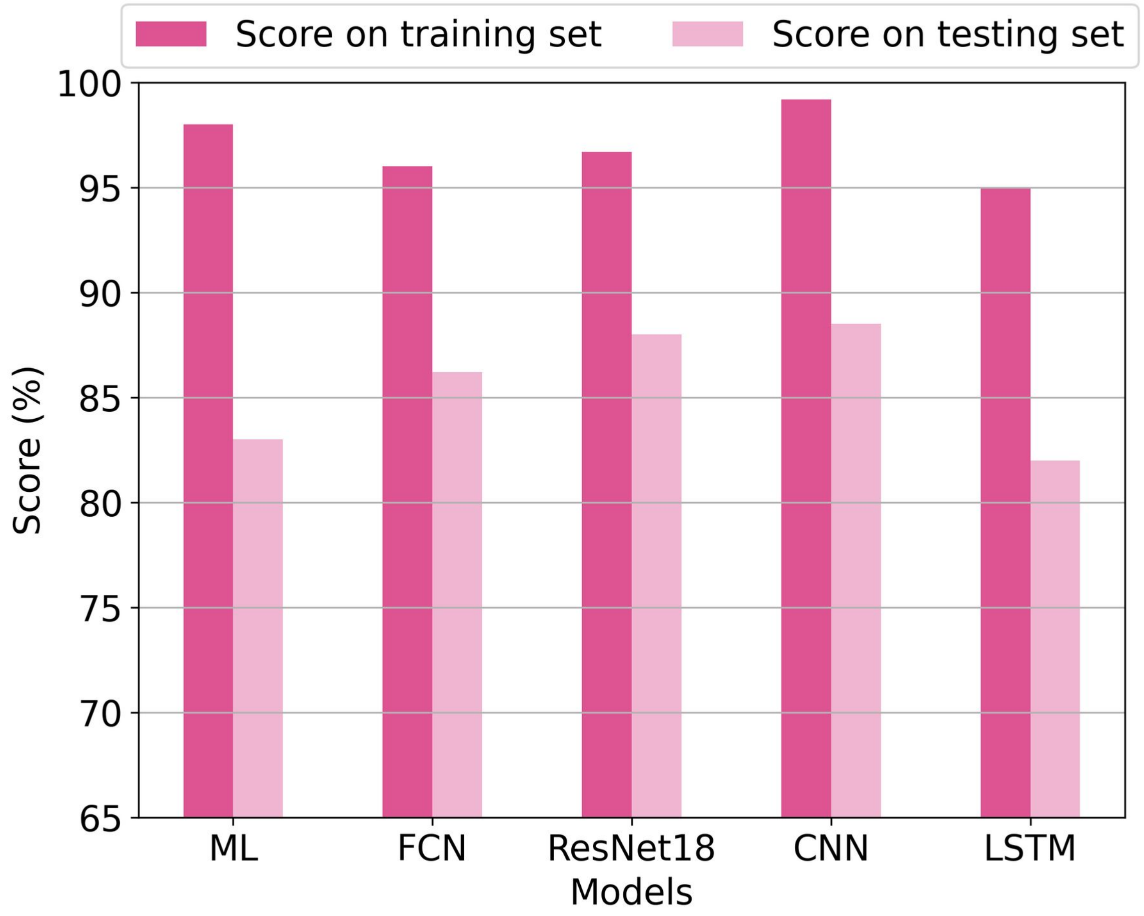
! *Training a temporal model is complex, and requires more data*

! *Thousands of epoch were needed*

Hochreiter Sepp and Schmidhuber Jürgen, "Long Short-Term Memory", *1997*

Results: Comparison

- All models are able to tracker the cursor on the training set
- All models have difficulties to generalise
- CNN outperforms other models
- Models based on the face mesh detector can find relevant features for cursor prediction
  - We believe that this method can help generalise across users



Challenges

- Dataset-related
  - How do we select the framerate, image resolution?
  - How do we normalize the acquired data?
  - Do we consider single samples or sequences?
  - Is our dataset balanced enough? Diverse?
  - Do we have enough data?
  - Can we model the relationship between images and mouse?
  - Does it depend on the configuration (relative position user vs. screen)

- Model related
  - How to select a model/architecture?
  - How to generalise across setup/user?
  - What learning rate to adopt to avoid local minimum and constant output?

Discussion

- Data type
  - Points extracted with face tracker do not seem enough to predict where the user is looking
    - Difficulty to fit model even on train set
    - Using images provides our best results
    - Whole images consume more memory
    - Necessity to use more complexes methods (deep learning approach)
    - Does cropping around user's eyes reduce memory consumption without performance decreasing?

- Data acquisition
  - Acquire more data : record colleagues during their work time → WIP

- Determine when the user is looking at the mouse
  - User can move the mouse without looking
  - User is sometimes looking after the mouse
  - User is sometimes looking before the mouse

- Use cases
  - In laboratory tests cases – Can we move the mouse without using it?
    - Not there yet !
  - Real cases – Can a paraplegic patient move the mouse?