Bio-Inspired Artificial Intelligence

Lecture 12:
Neurally-Inspired Gesture- and Action Recognition

Knowledge Technology

http://www.informatik.uni-hamburg.de/WTM/

Gesture Recognition: Motivation

- Human gestures are:
  - Communicative: supports or could even replace speech
    - Sign language, Demonstration
  - Unconscious behaviour: accompany body expressions, e.g. when we tell a story or describe a scene
    - Gesticulation on the phone
- Babies learn gestures and discriminate between facial expressions before they learn language
- Hypothesis: Gestures part of language acquisition (co-evolution)

Gesture Types

- Hand gestures
  - Static postures
  - Dynamic movements
  - Sign language
- Arm gestures
  - Waving
- Head gestures
  - Nodding
- Whole-body gestures
  - Aircraft commands, e.g. show parking position for a plane

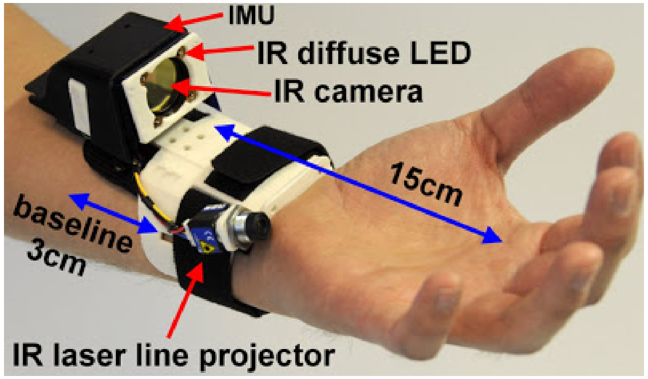
Gesture Taxonomy

- **Symbolic**, often culture-dependent
  - Victory sign
- **Deictic**
  - Pointing to an object
  - Holding an object
- **Iconic**
  - Hand movements produced along with speech
  - Narrative character, scene description
- **Pantomime**
  - Grasping gesture, tool-use
  - Making a fist

Sensors for Gesture Recognition

- Different input modalities

Wii Remote



Microsoft's Digits

Data Gloves

Nao robot

Data Acquisition for Gesture Recognition

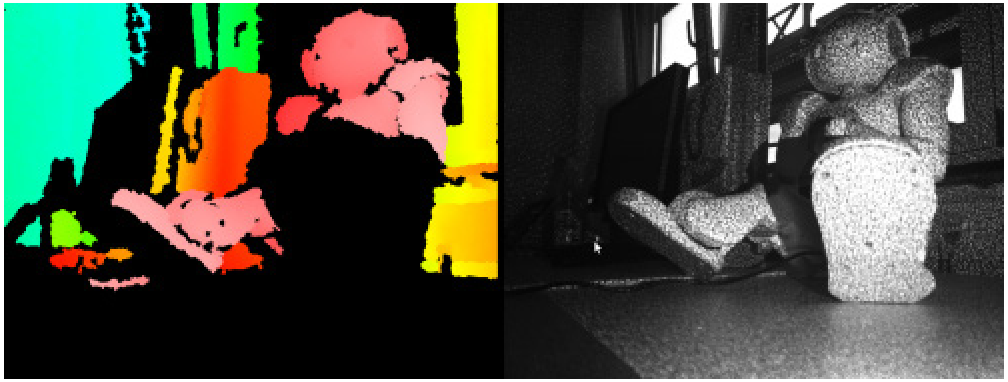
- Different devices to perform gesture recognition
- Mouse-based interfaces
  - - Difficult handling for infants and elder people due to cursor localization and possible neuro-motor diseases (Parkinson)
  - - No more prominent in research
- Data glove
  - + Sensors provide hand state and hand trajectory with high resolution
  - + Derivation of geometric hand models
  - - Technical equipment is expensive
  - - Complexity in processing sensor measurements
  - - Cables are uncomfortable for the users

Data Acquisition for Gesture Recognition

- Camera Settings:
  - + Most natural interface for Human-Machine-Interaction (no cables, enables communication at certain distance)
  - + New technology provides new ways for image processing (e.g. Kinect)
  - + Support Sign-Language
  - - Vast amount of images and gesture sequences needed for training and testing (similar as for speech recognition)
  - - Still complex computation hinders realtime application

Gesture Recognition with Kinect

- Drawbacks with cameras:
- In monocular camera-setup no depth information
- What is foreground, what is background?
- Difficult to extract face and hand as individual regions
- Hard to determine trajectories, e.g. command ‘Turn’
  - In 2D-plane no rotation visible



Gesture Recognition Focus

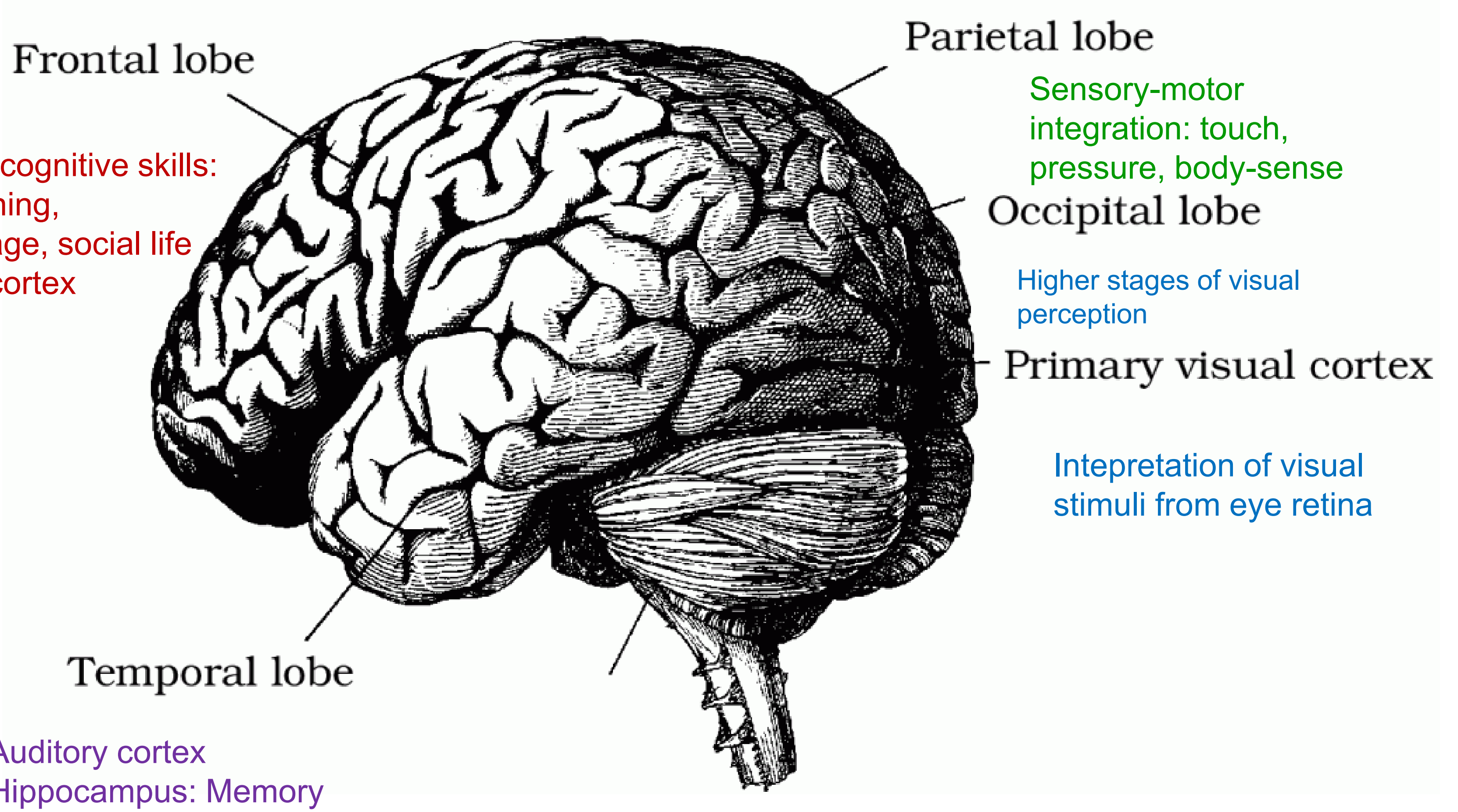
- Promising techniques for gesture recognition in general
  - E.g. Leap Motion for TV controlling
  - GestureTek for health assistance
- Usually very technical and constrained
  - for gloves, WiiMote: extra devices needed for the user
  - for Leap Motion: only small area of display can be controlled
  - for Kinect: skeleton tracker ends at users‘ wrist
    - No finger resolution, no hand model
    - Solution for KT: master student developed hand model
- Brain-inspired solutions for intelligent gesture recognition
- Important: Vision-based approaches to provide natural interface

Static Gestures vs. Dynamic Gestures

- Hand postures
- Example: Victory sign, OK (thumb up), depicting a cipher
- Need spatial-, shape-, and finger configuration information
- No temporal resolution needed
- Elastic Graph Matching, Multi-Layer Neural Network, Convolutional Neural Networks

- Hand movements
- Example: Turn-around, hand waving (hello, good bye), pointing
- Need both spatial- and temporal information
- Finger detection sometimes not necessary
- Recurrent Neural Networks, Hidden Markov Models, Self-Organizing Maps

Lobe Parcellation of the Brain



Higher cognitive skills:
Reasoning,
Language, social life
Motor cortex

Sensory-motor integration: touch, pressure, body-sense

Higher stages of visual perception

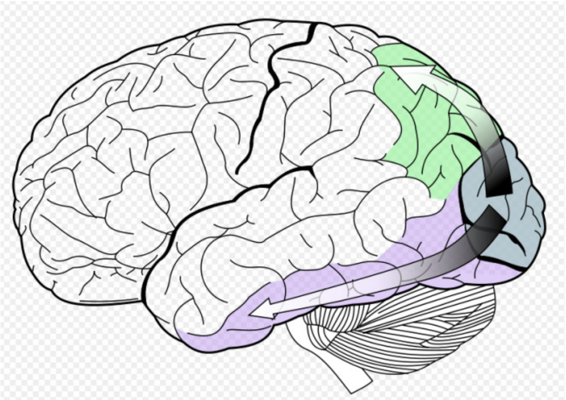
Intepretation of visual stimuli from eye retina

Auditory cortex
Hippocampus: Memory

Image source: standford.edu, vista group

The Visual Cortex

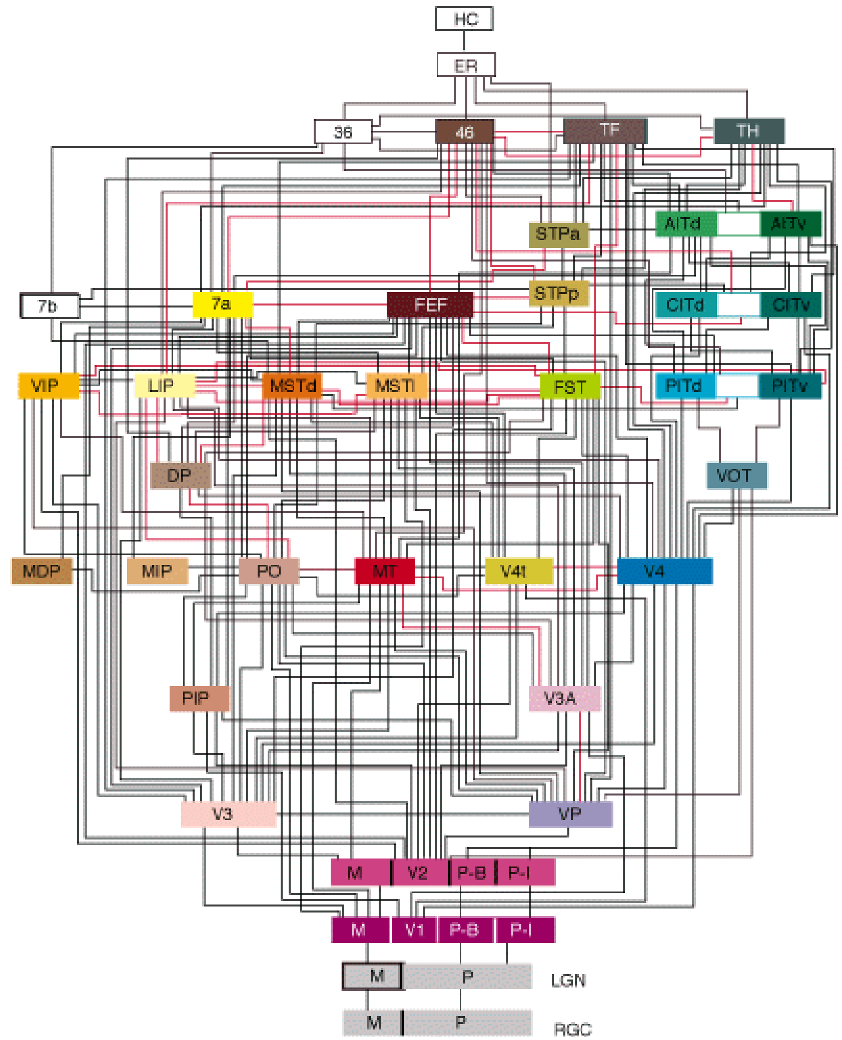
- Complex pathway of visual stimulus from retina to brain areas in visual cortex
- Two pathways in object recognition
  - Ventral 'what' and dorsal 'where', 'how'
  - Ventral codes for object properties, dorsal for spatial position



- Computational models reflect visual processing in feedforward fashion
  - Distinction between *simple* and *complex* cells
  - Inspiration for convolutional networks (LeCun) or Neocognitron (Fukushima)
- Feedback connections important in perception-to-action tasks e.g. reaching for an object → location update

Hierarchical organization of Visual Cortex

- Felleman&Essen, 1992
- Depict complex neural connectivity of visual cortex projecting to temporal lobe
- Computational model need not only feedforward, but also recurrent connections for feedback the information

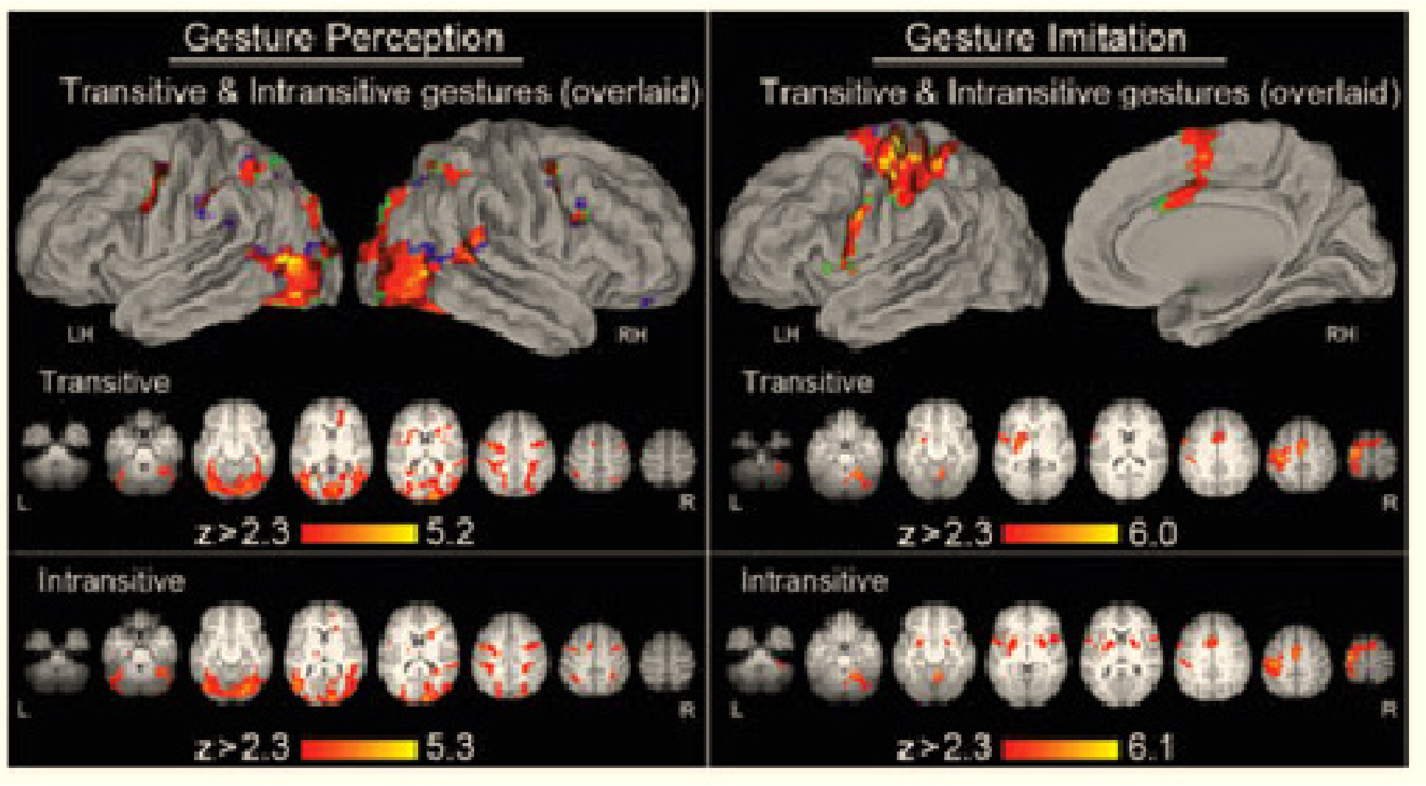


Activation for Gestures in the Brain

- Underlying neural processes for gestures still to be investigated
- Neuro-imaging techniques help localising brain regions involved in gesture recognition
- Differences in gesture recognition and ~ production
- Involvement of
  - Visual Cortex: V5, middle temporal (MT)
    - Finding implies motion-sensitivity
  - Frontal lobe: inferior frontal gyrus (IFG), suppl. motor area (SMA)
    - Finding implies motor-sensitivity

Neural Substrates

- fMRI experiments by Villareal et al., 2008



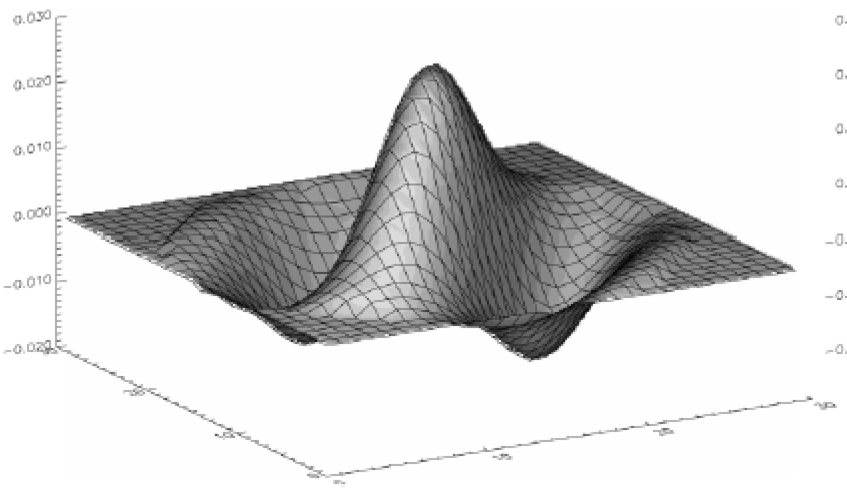
From Neuroscience to Computer Science

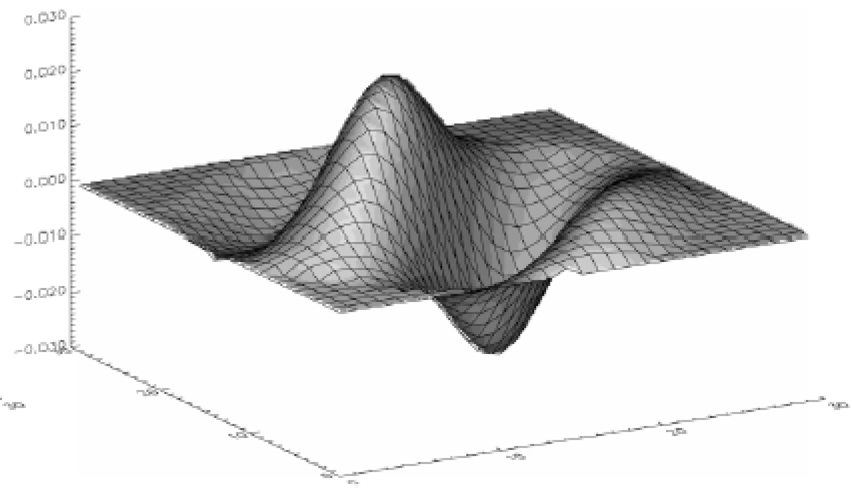
- Findings from neuroscience helps understanding processes
- Usually very complex to model (see Felleman&Essen)
- Abstraction needed
- Example:
  - V1: Edges
  - V2: colour
  - V3: depth
  - V4: shape
  - V5: motion

Layer	**Process**	**Represents**
S_1	Gabor filtering	simple cells in V1
C_1	Local pooling	complex cells in V1
S_2	Radial basis functions	V4 & posterior inferotemporal cortex
C_2	Global pooling	inferotemporal cortex

Response to visual stimuli

- Gabor wavelets as edge detectors
  - Functionality investigated in cats striate cortex
- Signal processing in frequency domain
- Gabor filter responses, wavelets
- Weighted series of sine and cosine functions





Elastic Graph Matching: Bio-inspired approach for static postures

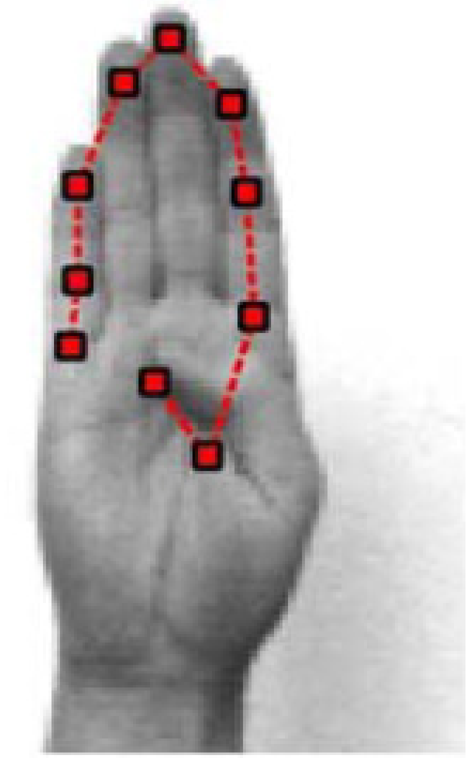
- Waves parameterized with different scales and orientation

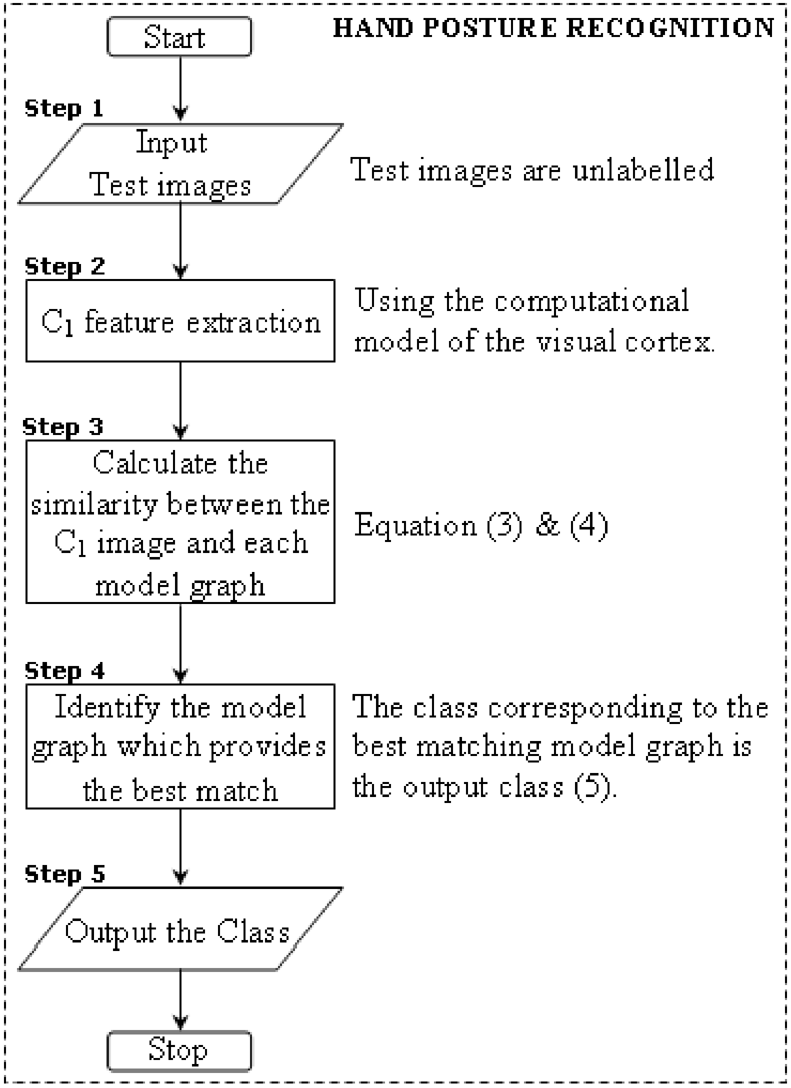
$$\psi_{\mathbf{k}}(\mathbf{x}) = \frac{\mathbf{k}^2}{\sigma^2}\exp\left(-\frac{\mathbf{k}^2\mathbf{x}^2}{2\sigma^2}\right)\left[\exp(i\mathbf{k}\mathbf{x}) - \exp\left(\frac{-\sigma^2}{2}\right)\right]$$

- Formula describes ‘Jets’, i.e. result is an n-dim. vector containing the different responses depending on μ and ν

$$\mathbf{k}_{\nu\mu} = k_\nu \begin{pmatrix} \cos\phi_\mu \\ \sin\phi_\mu \end{pmatrix} \quad \text{with} \quad k_\nu = k_{max}/f^\nu,\ \phi_\mu = \frac{\mu\pi}{D}$$

Elastic Graph Matching: Implementation





- Nodes are local image patches with 15 x 15 pixels and with four orientations (0°, 45°, 90°, -45°)

Elastic Graph Matching: Pro and Con

- + No image segmentation needed
  - One of the challenging tasks in gesture recognition
- + Successfull application even with complex backgrounds
- + Gabor wavelets insensitive to lightening conditions and scaling

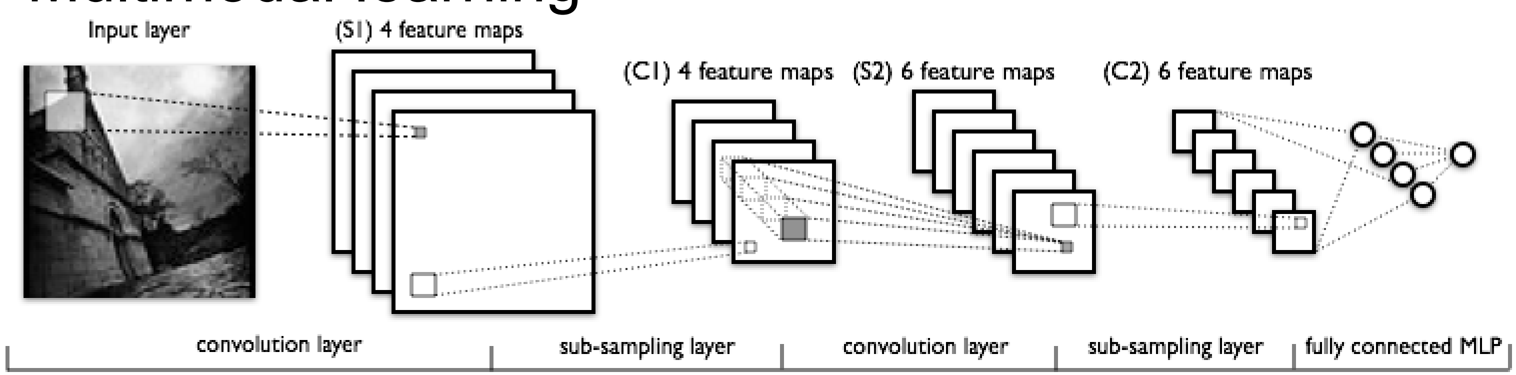
- - Derivation of bunch graphs
- - Computional performance:
  - It lasts several seconds to compare models and image
  - Online processing difficult
    - Impractical for recognition of sign language
  - Solution: hierarchical models of hand postures
    - Static postures can be combined to derive dynamic gestures

Posture Recognition Using CNN

- Problem statement:
- How to make a robot learn instructions with speech and gestures?
- Multimodal learning
- Real world application
  - Robot sensors
  - Noisy information
  - Performance issues

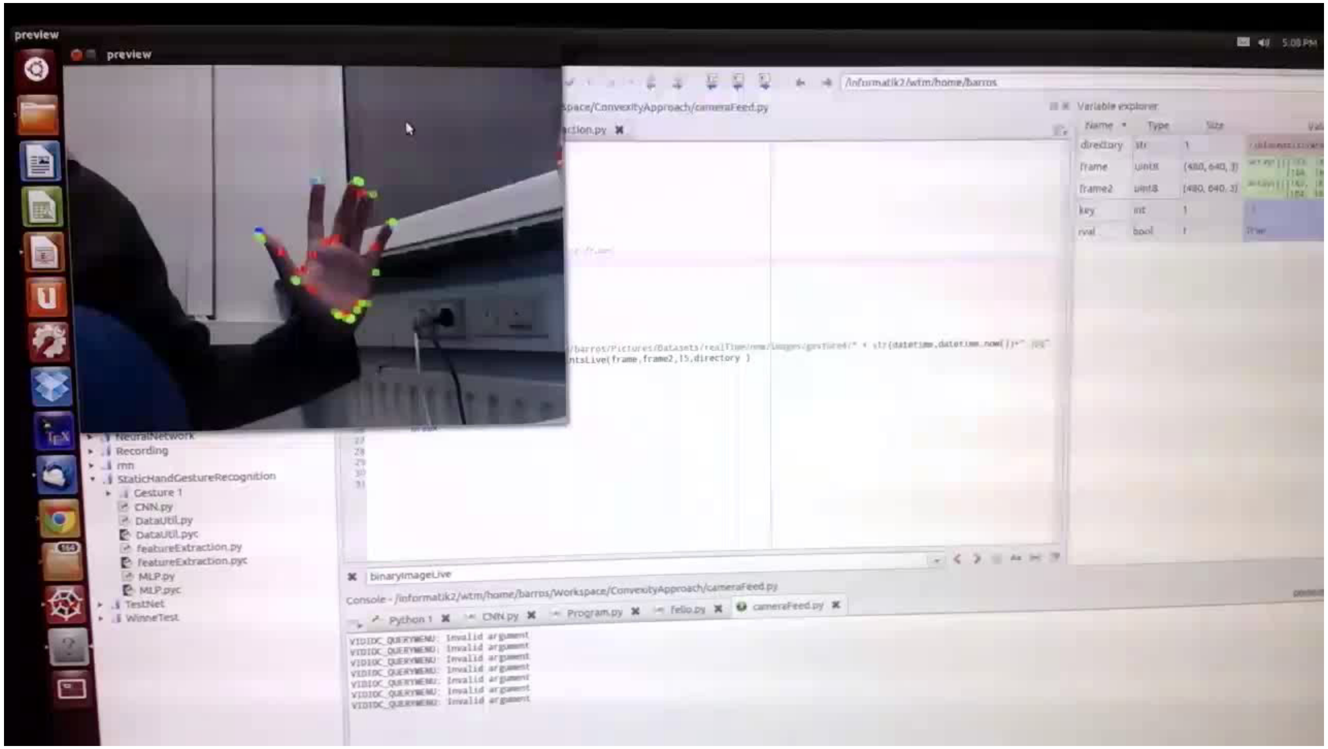
Deep Learning

- Bioinspired neural architecture
- Shared information representation between the layers
- Implicit feature extraction
  - Local features
  - Global features
- Robust against noise
- Multimodal learning



First Steps - Gestures

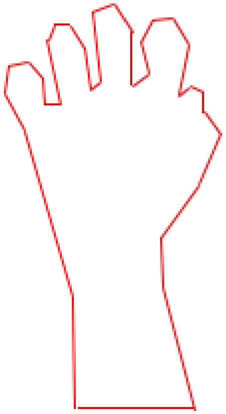
- Feature Extraction
  - Implicit
    - Mathematical/statistical approach



First Steps - Gestures

- Feature Extraction
  - Implicit
    - Convexity Approach [1]

Hand segmentation

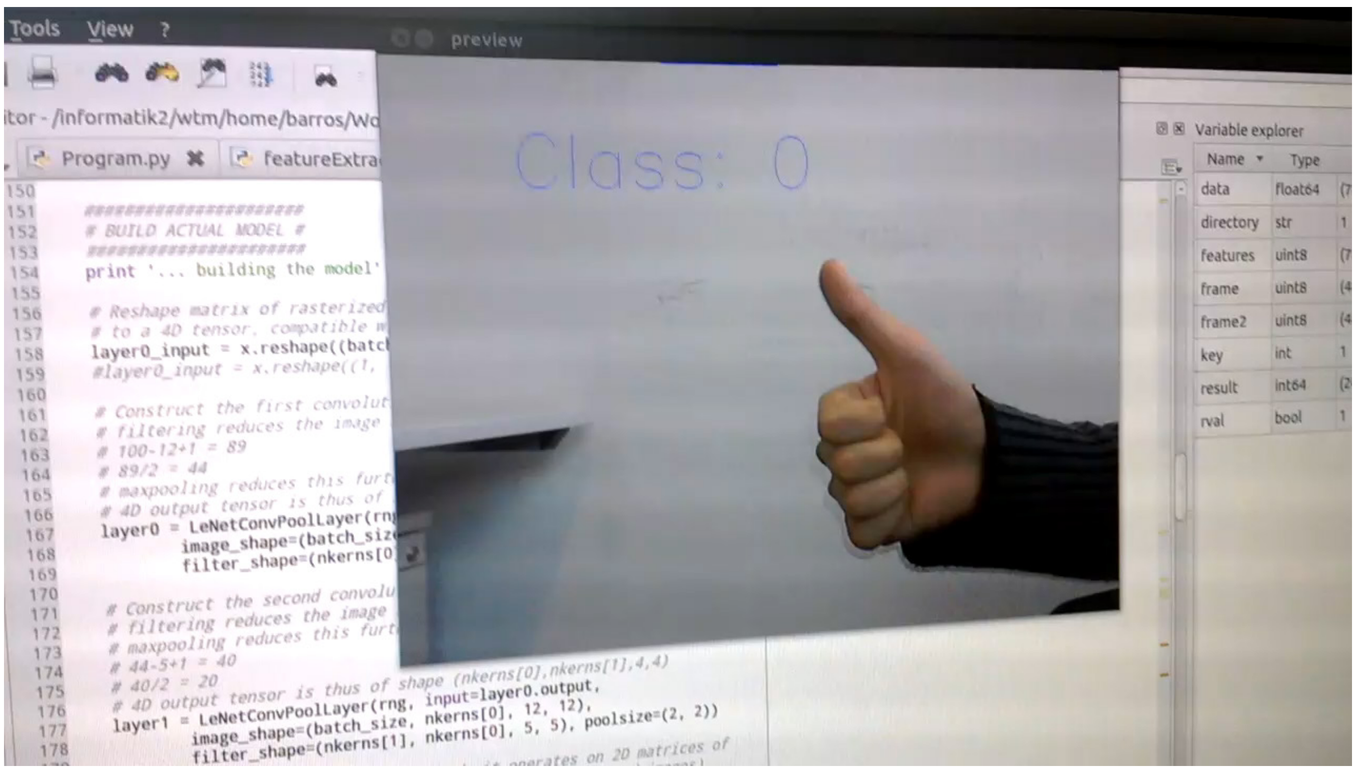


Model minimization

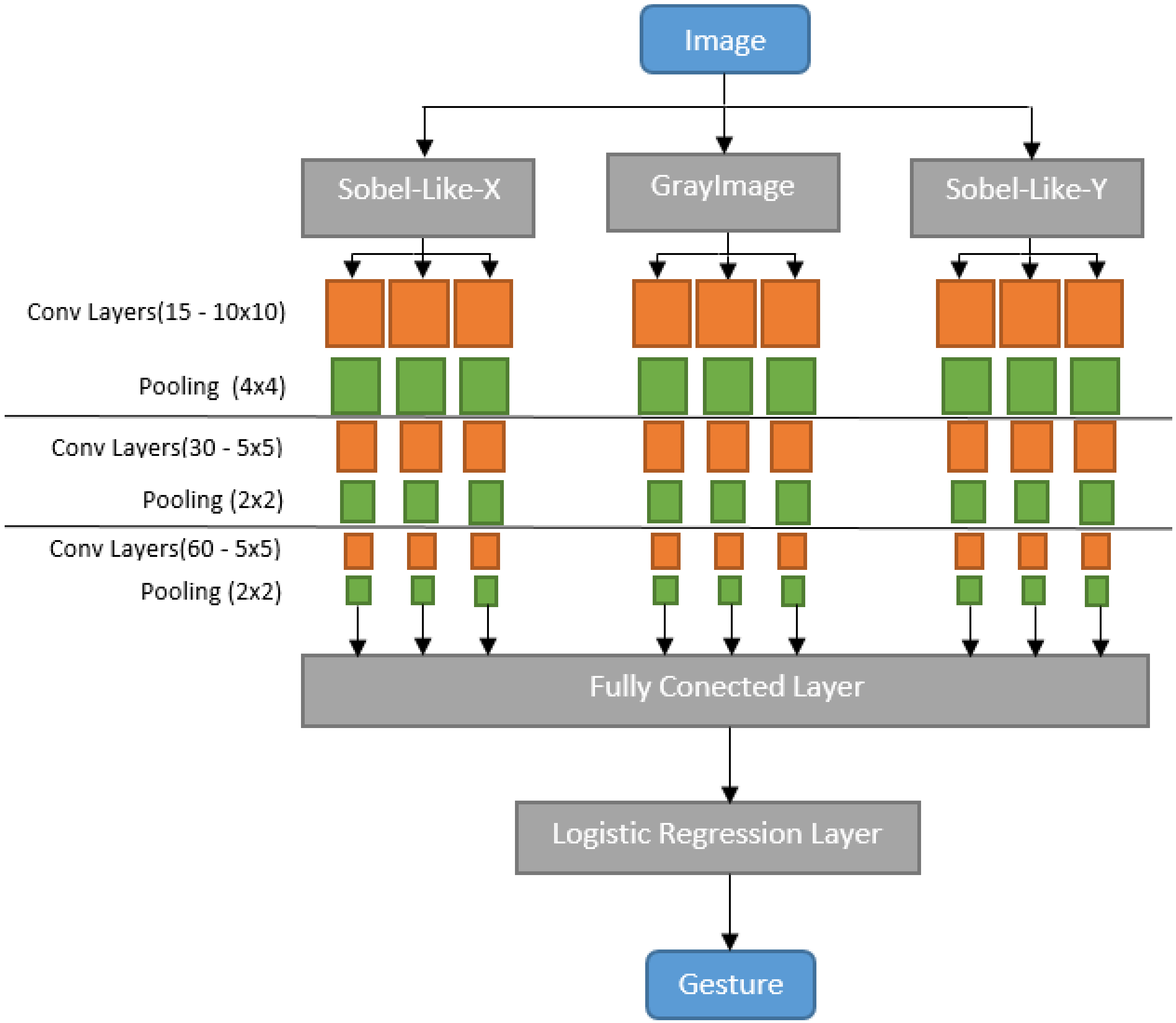
Points selection

First Steps - Gestures

- Feature Extraction
  - Explicit
    - Neural Approach: Convolutional Neural Networks



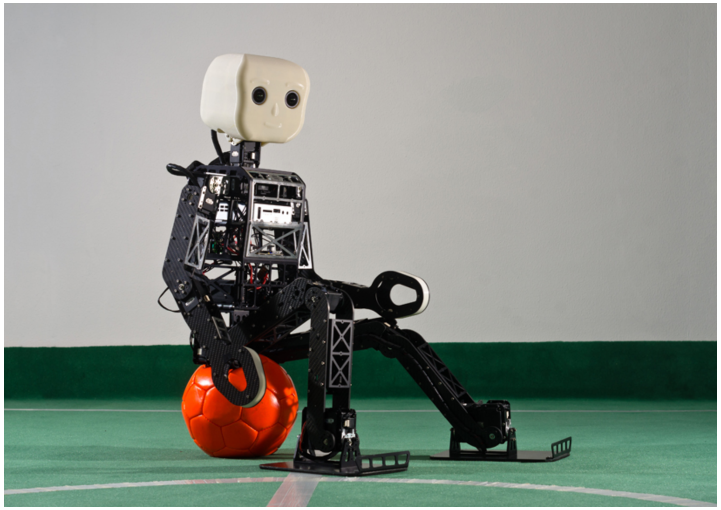
First Steps – Deep Architecture



Further Planned Experiments

- Simulation of a real world scenario
- Set of instructions
  - Only Gestures
  - Only Speech
  - Gesture and Speech

- Use of a humanoid robot
  - Nimbro



Dynamic Gestures: Preprocessing

Image Sequence

↓

RGB to YCbCr

↓

Skin Regions/Masks

↓

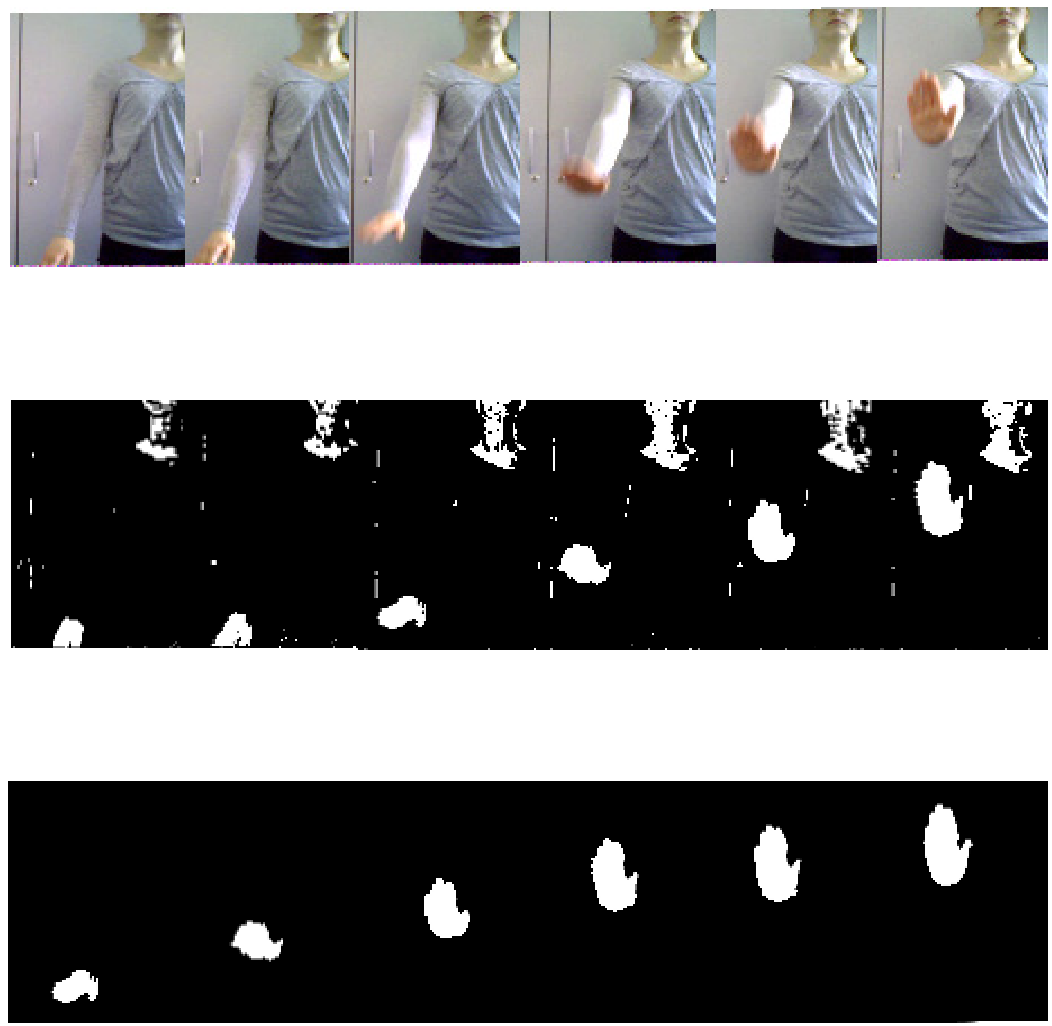
Connected Component Analysis/Labelling

↓

Morphological operations

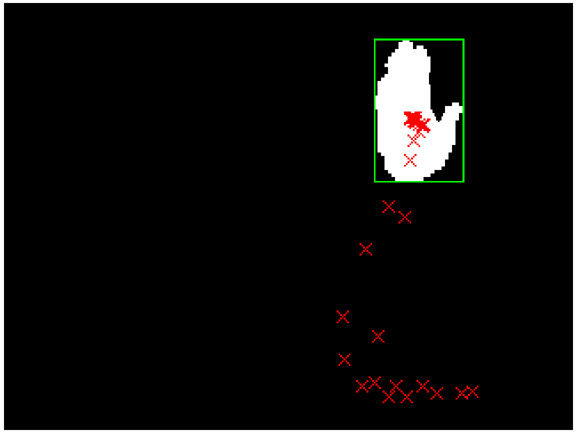
↓

Threshold: Hands



Dynamic Gestures

- Gestures as varying time-series
  - Trajectories derived from different reference frames
    - Fingertips, marker-based
    - Hand centre, 2D vision
    - Wrist joint, Kinect
- Gesture vs. Gesticulation
- Gesture spotting: start, end
  - Important in sign-languag
    - ‘Visual syntax‘

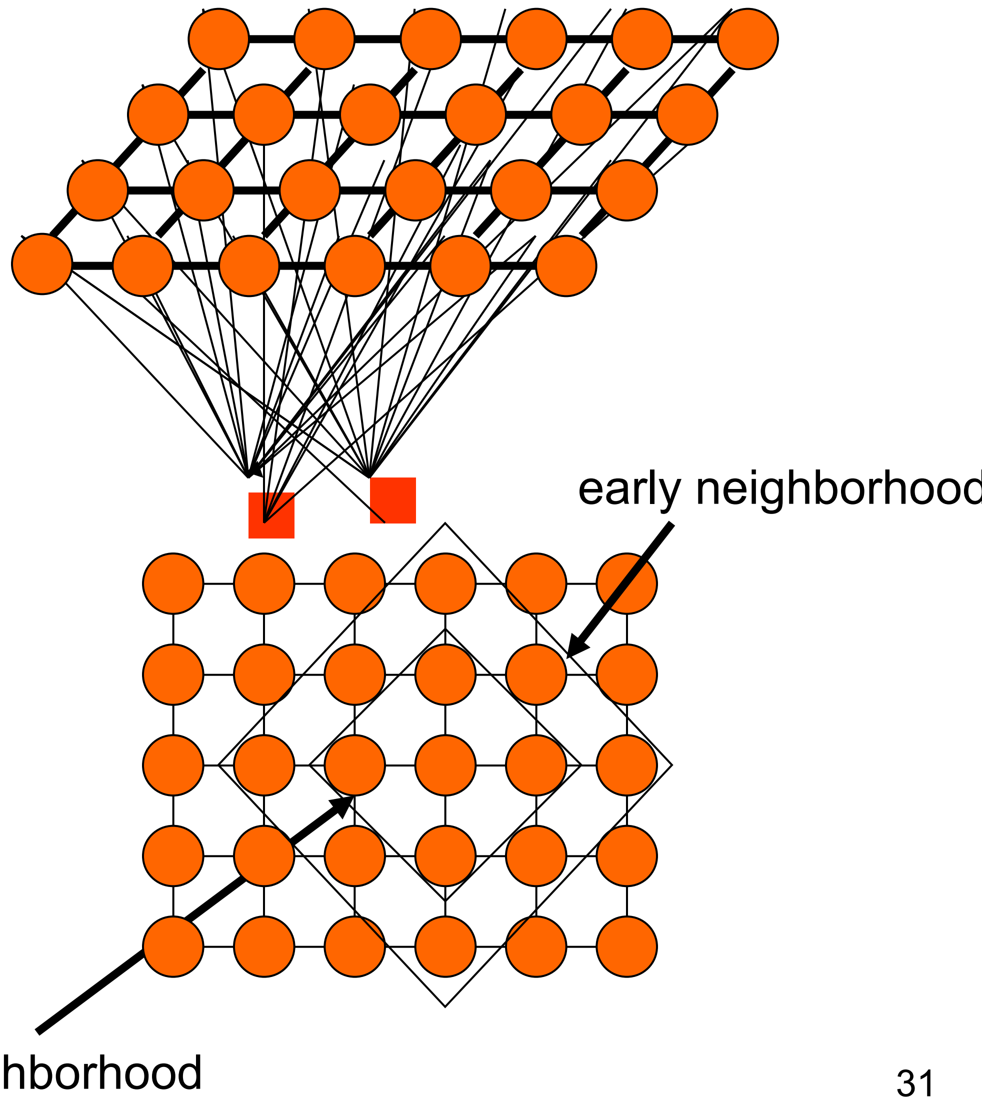


Approach for Dynamic Gesture Recognition

- Uses self-organizing maps (SOM)
- Gesture DB consists of
  - **Commands**: abort, zoom, turn-around
  - **Italian gestures**: needs inclusion of head information
- Latter also known as co-speech
- Cultural dependence
- Co-speech underlines what a person is saying
  - Iconic classification
- Saliency-based approach to attract attention to the hands
  - Encodes stimulus perception on retina
  - i.e. moving objects will provoke eye gaze to its direction

Self organizing maps

- The activation of the neuron is spread in its direct neighborhood
  - neighbors become sensitive to the same input patterns
- The size of the neighborhood is initially large but reduced over time during training as the network neurons become more specialized

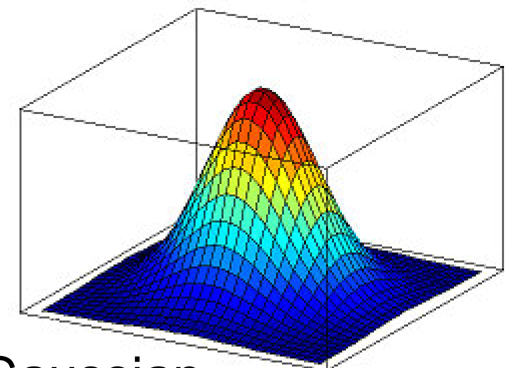


Neighborhood Function

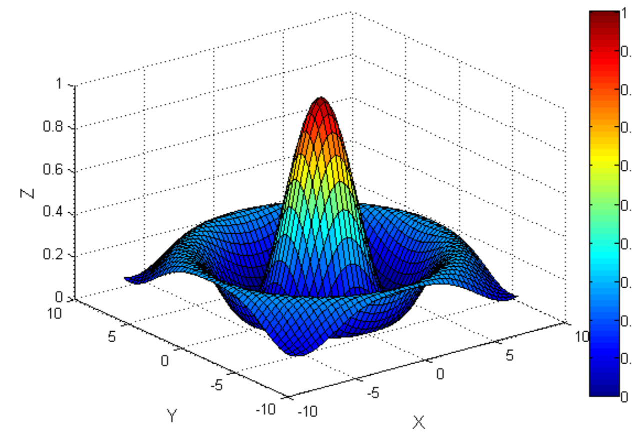
- The neighborhood function $h(n_b,t)$ determines the degree of weight vector change of the neighbors

$$w_j^T \leftarrow w_j^T + \eta(t) \cdot h(n_b,t) \cdot \left(x - w_j^T\right)$$

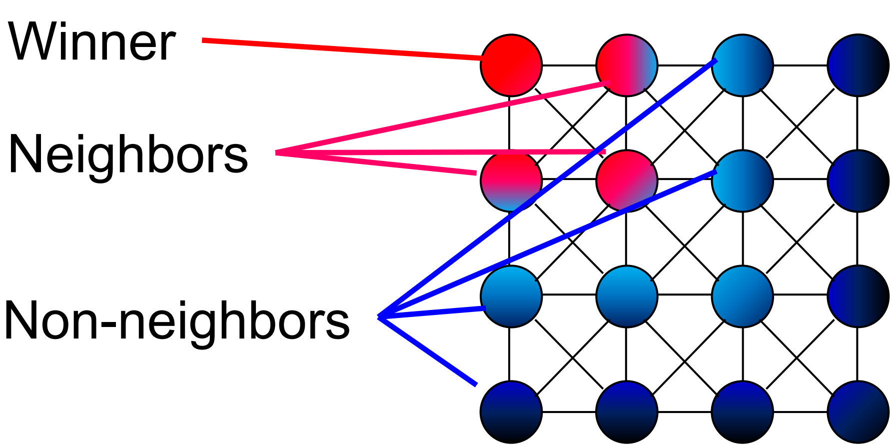
  - Mostly used: Gaussian or Mexican Hat function
  - Goal: ***Preserve*** the ***topology***



Gaussian

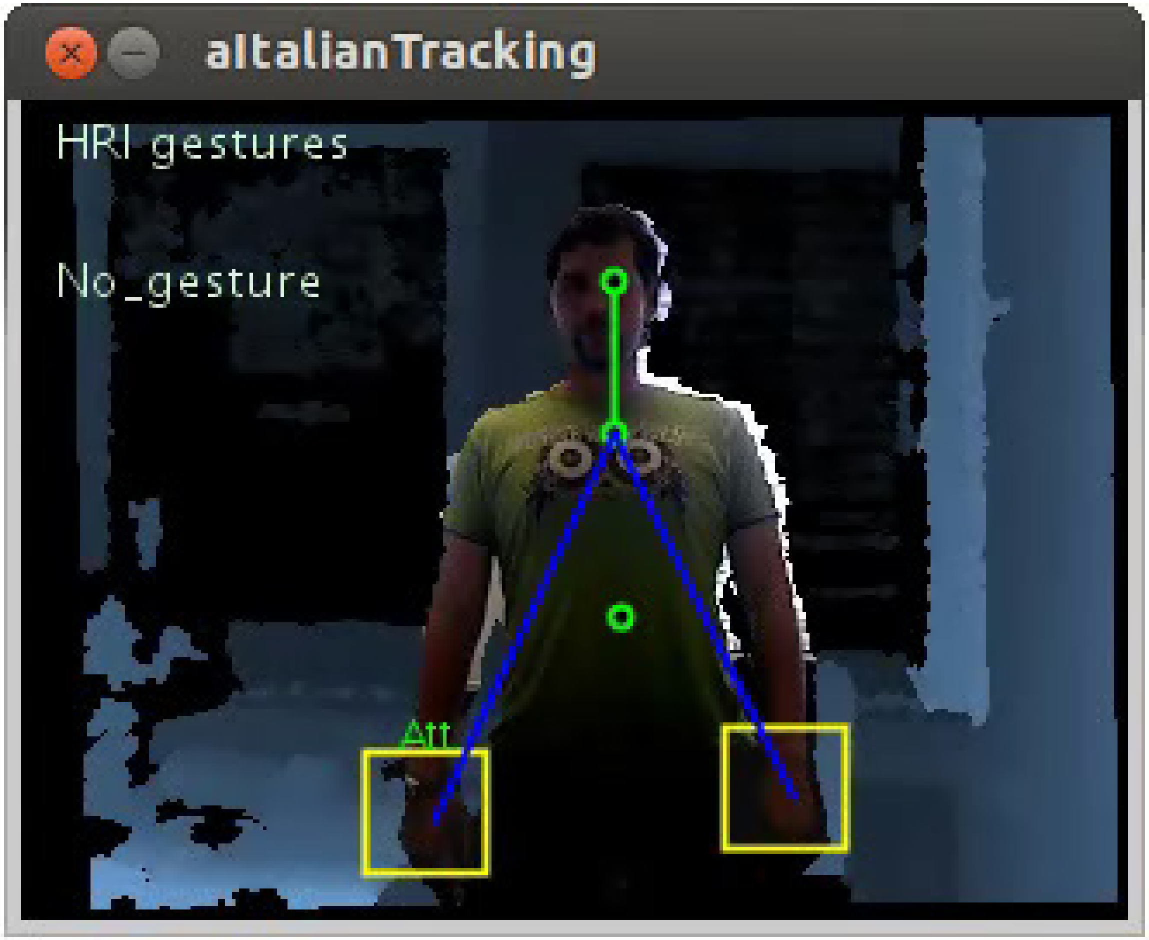


Mexican Hat



Approach for Dynamic Gesture Recognition

- SOM-based motion clustering
- Saliency-based encoding
- Hand independent
- Labeled training gestures
- Real time

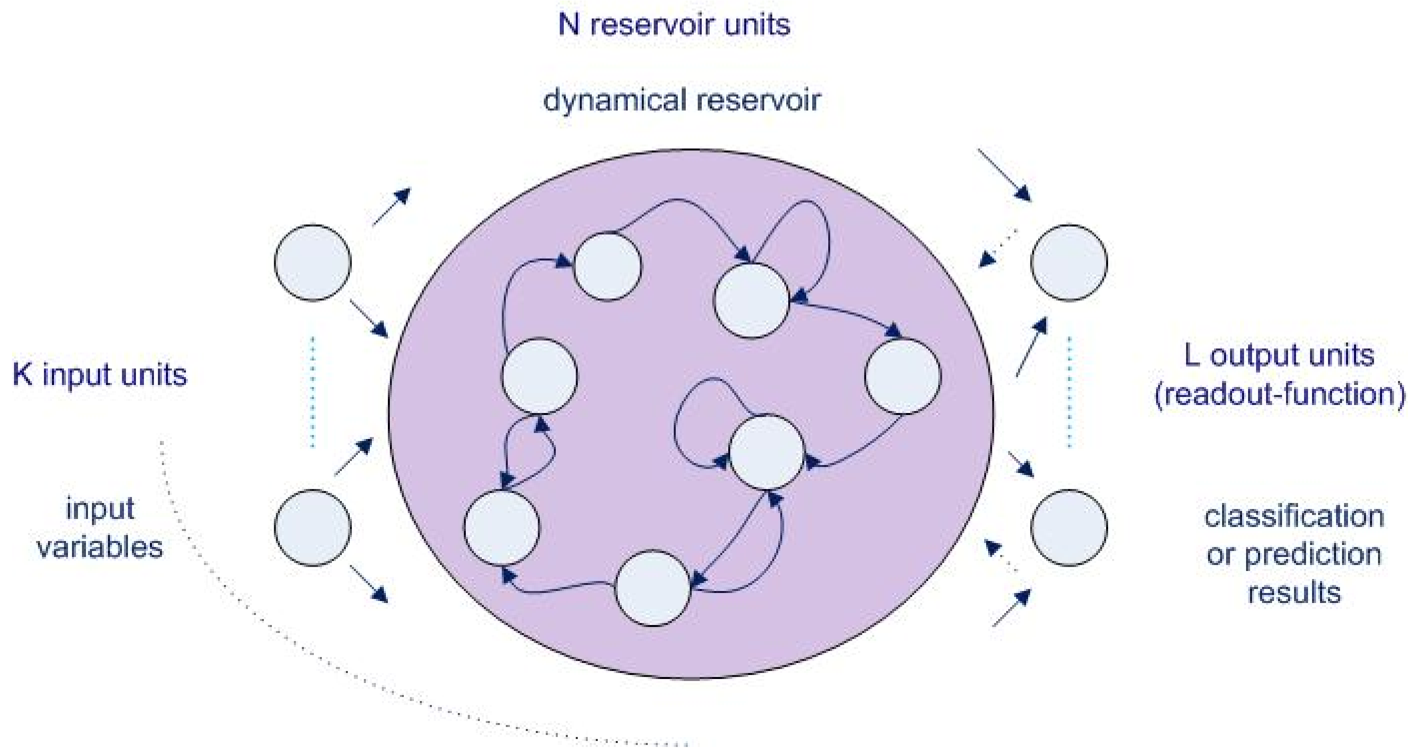


Neural Network Approaches

- For static images: MLP classifier
- For sequences: Recurrent Neural Networks (RNN)
  - capture image dependences
  - Provide context
  - Parameter: e.g. number of hidden layer, number of neurons
  - Training with Backprogagation Through Time
- Bio-inspired approach (Dominey, 1995):
  - Neurons in prefrontal cortex (PFC) assumed as ‘reservoir‘
  - Nonlinear dynamics ‘echo-ed‘ by neurons
  - Reservoir computing: Training only on linear read-out
  - Echo State Networks, Liquid State Machines, Temporal RNN

Neural Network Approaches

- ESN provide short-term memory
- Important when dealing with dynamic gestures, which follow a specific ‘syntax‘
- Approach successfully applied to language comprehension

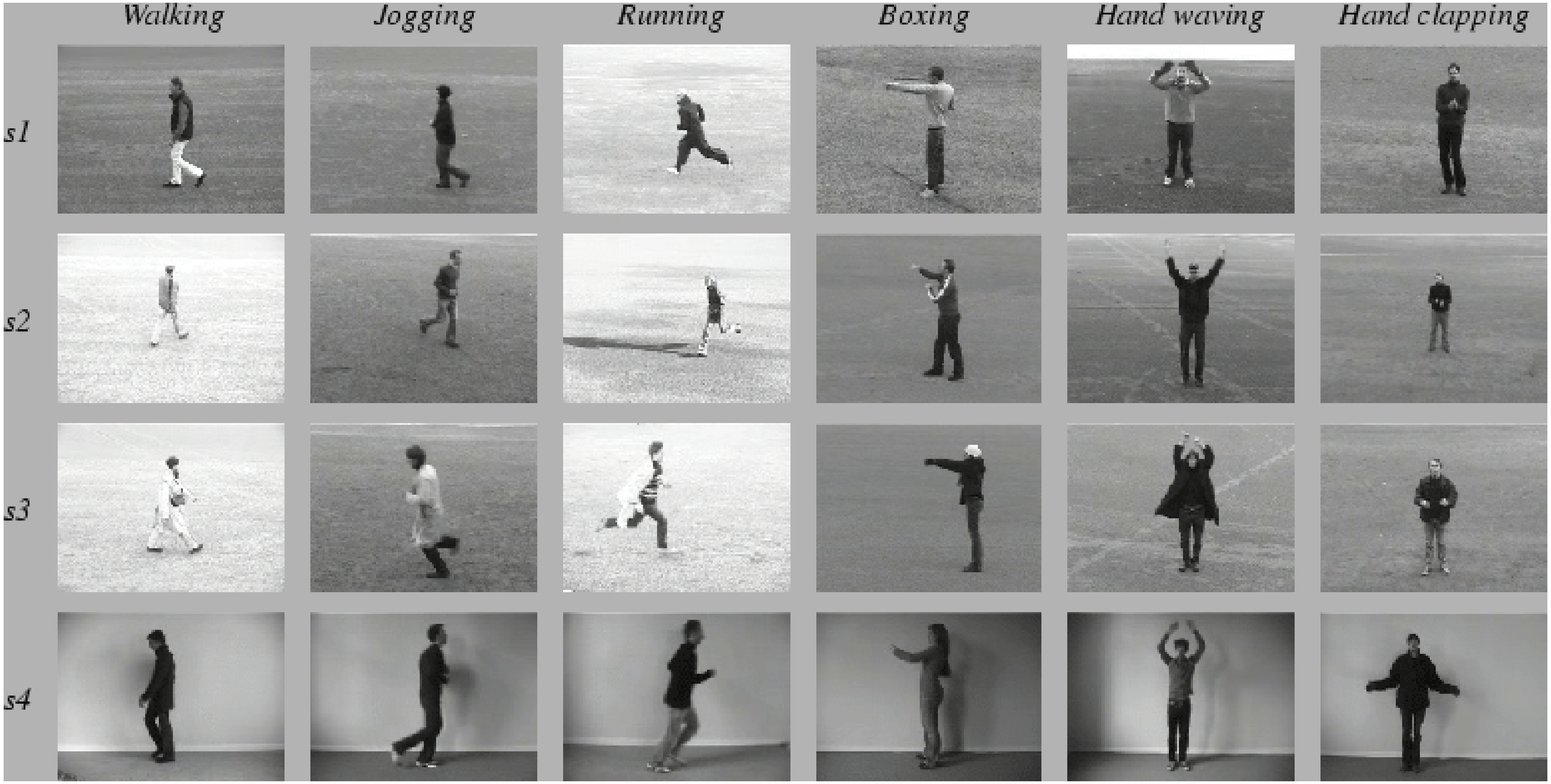


Interim Summary

- Gesture Recognition is essential for:
  - Assistive system component in e.g. health care
  - Human-Machine-Interaction
  - Establishment of multimodal systems, e.g. combination of speech and gestures for communication
- Static gestures vs. dynamic gestures
- Different devices available, but vision-based gesture recognition provides most intuitive interface
- Still real-time challenges, but new devices help to overcome
- Interesting topic to derive new neural network models
  - Add also: Attention, context, imitation, ….?

Action Recognition

- Different subjects, perspectives, and lightning conditions makes action recognition a challenging task

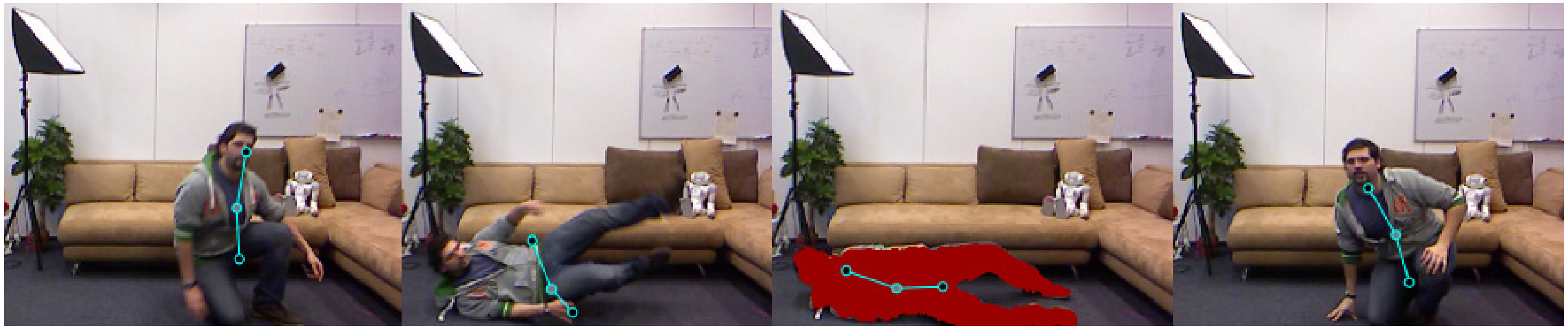


Motivation

Human Action Recognition

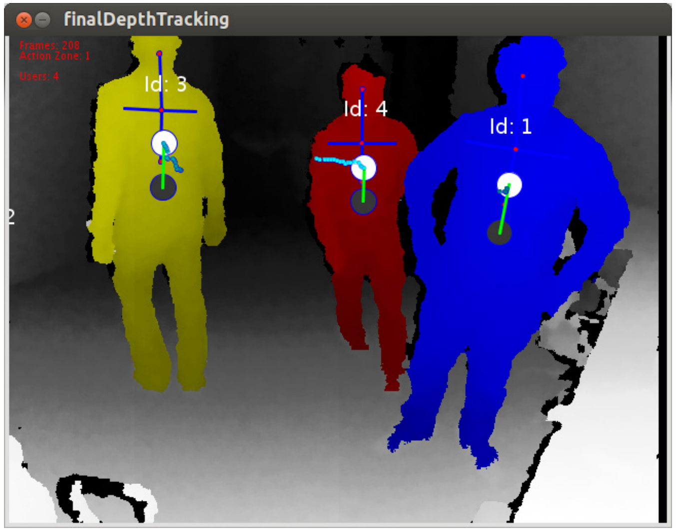
Visual-based applications for real world scenarios:

- Robust (light condition, occlusions, …)
- Adaptive (application domain)
- Fast (real time recognition)



Depth Information

- Images that contains information relating to the distance of the surfaces of the objects in the scene
- Estimation of depth under varying light conditions
- Computationally efficient for segmentation tasks
- Depth sensors
  - Time-of-Flight cameras
  - Stereovision
  - Structured Light



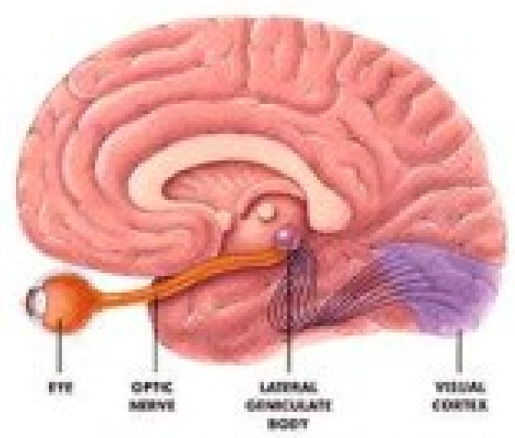
Bio-Inspired Approach

Biological principles from our visual system

- Dynamics of cognitive/perceptive processes
- Efficient computational models
- Learning systems for adaptive and robust HAR

Issues:

- Huge amount of visual information
- Sensor noise
- Representation of human actions



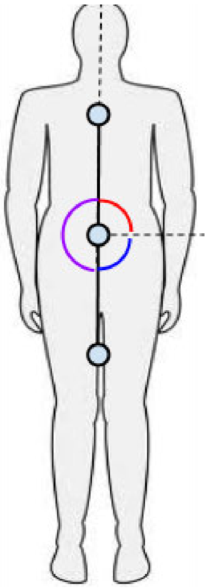
Motion Extraction

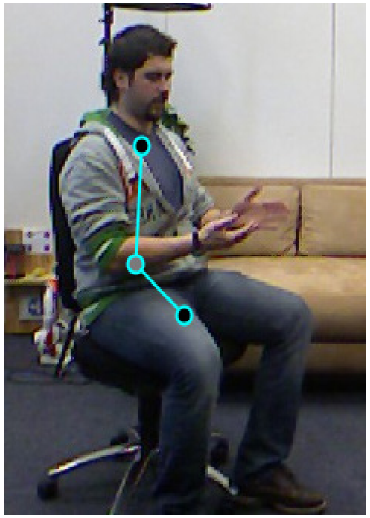
Visual attention

- Estimation of motion heuristics
- Saliency-based encoding
- Reduced amount of processed information

Noise Reduction

- Unsupervised outlier detection
- Perceptual-motivated interpolation
  - No loss of information

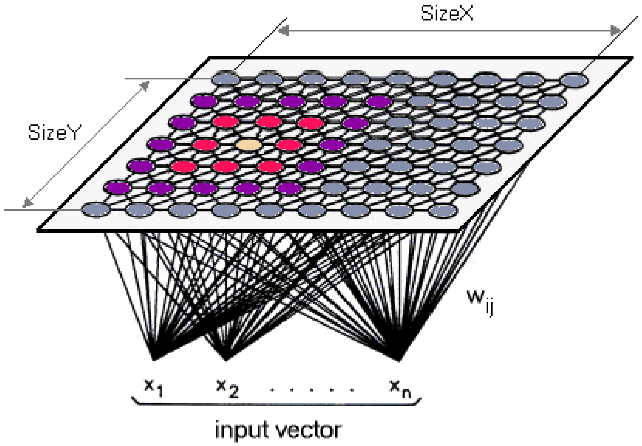




Actions as Motion Sequences

- Representation of actions
  - Sequence of body postures and temporal dependencies
  - Concatenation of flow motion vectors (time windows)
- Semantically different actions can be computationally fuzzy
  - Modeling of relevant spatiotemporal motion properties

Learning Framework



Dynamics of the visual system

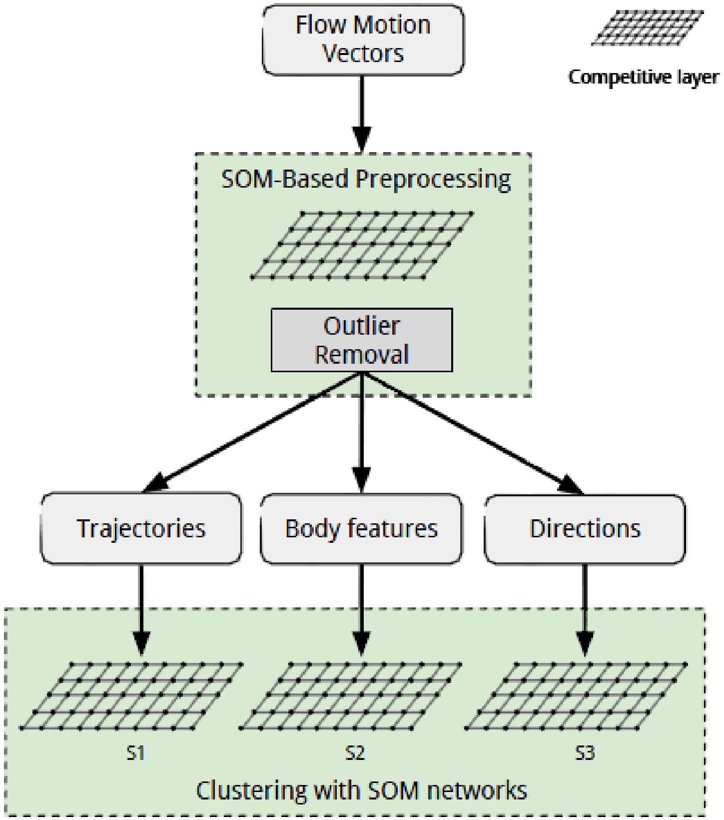
- Distributed, hierarchical architecture
- Cortical input-driven self-organization
  - Self-tuning to the distribution of inputs

Application-oriented viewpoint

- Plausible computational model of the visual cortex
- SOM and similar extensions
  - Behavioral patterns in terms of multi-dimensional vectors
  - Unsupervised + supervised learning for labelled actions

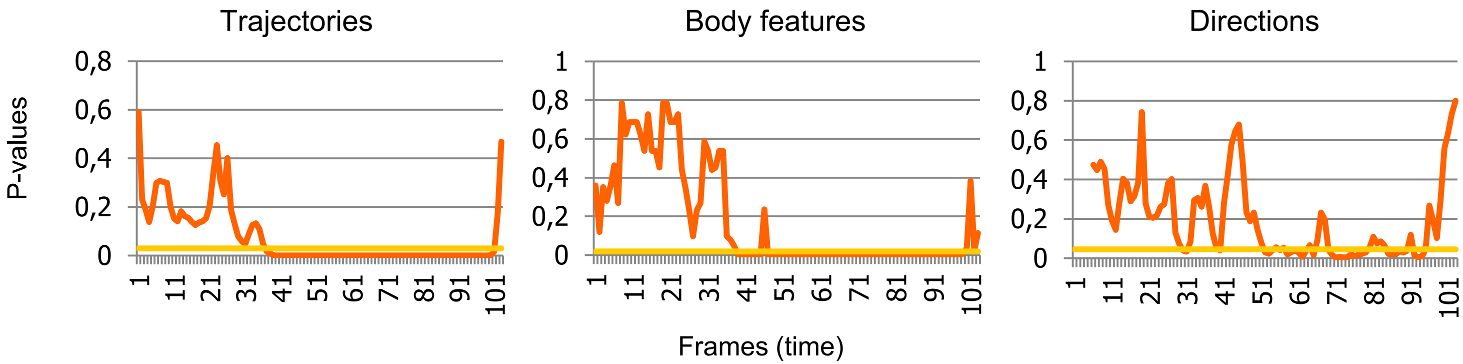
SOM-based Novelty Detection

- Detection of novel behavioral patterns
- System trained on domestic actions in terms of motion descriptors:
  - Trajectories
  - Body features
  - Directions
- Neural-statistical architecture for detecting novel observations



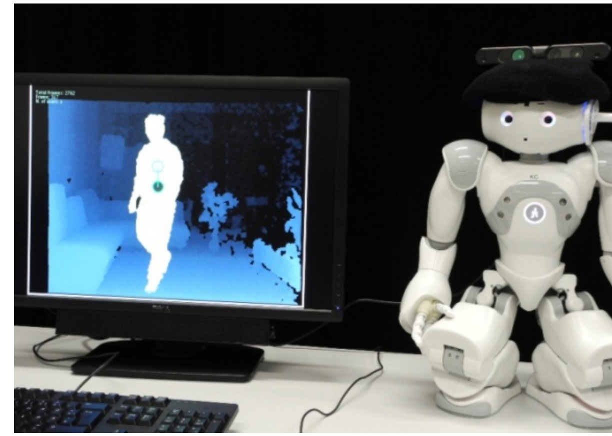
Results on fixed sensor

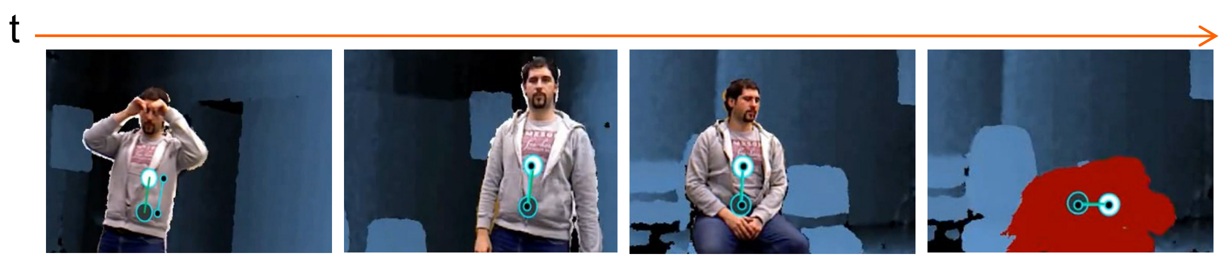
- The system reported novel actions not presented during the training
  - Falling down, fainting, crawling, jumping, visiting novel areas
- P-values for novel behavior lie under the novelty threshold
  - Abnormal behavior is reported if five consecutive frames are novel

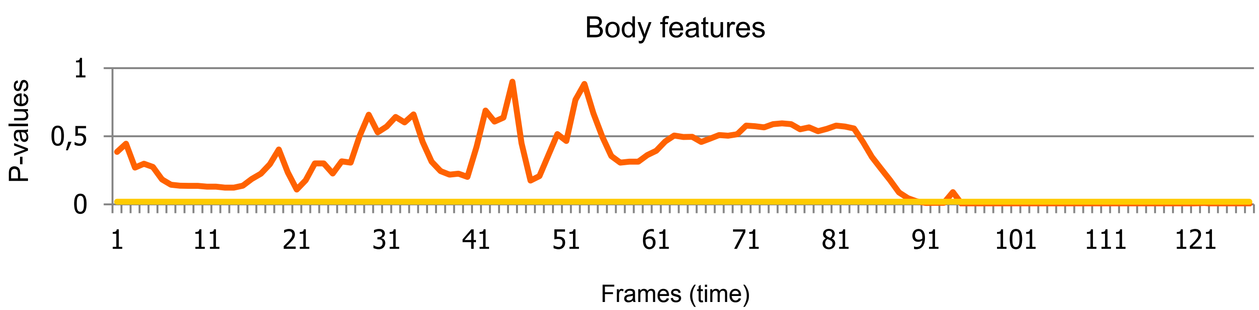


Active tracking with moving target

- Target moves around the environment
- Only body features are considered
  - Detected actions: falling down, crawling, jumping
  - P-values for novel activity are below the threshold







Summary

Learning systems for HAR

- Representation of human actions
- Introduction of biological principles
  - Attention, noise reduction
- Encoding of action dynamics
- Hierarchical learning architectures

Need of an extensive data set of human actions!

References

- Villareal M, Fridman EA, Amenqua A, Falasco G, Geschcovich ER, Ulloa ER, Leiguarda RC, Neural Substrate of Gesture Recognition, *Neuropsychologia*, 46(9):2371-82, 2008]
- Barros, P. V. A.; Junior, N. T. M.; Bisneto, J. M. M.; Fernandes, B. J. T.; Bezerra, B. L. D. & Fernandes, S. M. M., An Effective Dynamic Gesture Recognition System Based on the Feature Vector Reduction for SURF and LCS., *in* Valeri Mladenov; Petia D. Koprinkova-Hristova; Günther Palm; Alessandro E. P. Villa; Bruno Appollini & Nikola Kasabov, ed., 'ICANN' , Springer, , pp. 412-419, 2013 .
- G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, 29(6):82-97, 2012.
- Shuiwang Ji; Wei Xu; Ming Yang; Kai Yu, "3D Convolutional Neural Networks for Human Action Recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol.35, no.1, pp.221,231, Jan. 2013
- Jochen Triesch and Christoph von der Malsburg Robust Classification of Hand Postures against Complex Backgrounds, *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, *pp 170-175*, *IEEE Computer Society Press*, Killington, Vermont, USA, October 14-16, 1996.
- G. I. Parisi, S. Wermter, S. ”Hierarchical SOM-Based Detection of Novel Behavior for 3D Human Tracking” *Proceedings of International Joint Conference on Neural Networks*, pp. 1380–1387, Dallas, US, 2013
- Xavier Hinaut, Peter Ford Dominey (2013) Real-Time Parallel Processing of Grammatical Structure in the Fronto-Striatal System: A Recurrent Network Simulation Study Using Reservoir Computing, e52946. In *PLoS ONE 8 (2),* 2013.