Data Mining

Lecture 13
Hybrid Systems and Current Topics in Data Mining

Knowledge Technology

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

Overview

- An overview over data mining applications
- Hybrid architectures
- Data mining from sound and vision streaming
- Working with unreliable data

Data Mining Applications

- Data mining: A young discipline with broad and diverse applications
- There still exists a nontrivial gap between generic data mining methods and effective and scalable data mining tools for domain-specific applications

Data Mining Applications

- Some application domains (briefly discussed here)
  - Data Mining for Financial data analysis
  - Data Mining for Retail and Telecommunication Industries
  - Data Mining in Science and Engineering
  - Data Mining for Intrusion Detection and Prevention
  - Data Mining and Recommender Systems
  - Data Mining from Sound and Vision Streaming
  - Data Mining for Robotic Systems

Data Mining for Financial Data Analysis (I)

- Financial data collected in banks and financial institutions
  - often relatively complete
  - reliable
  - of high quality

- Design and construction of data warehouses for **multidimensional data analysis** and data mining
  - View the debt and revenue changes by month, by region, by sector, and by other factors
  - Data visualization of statistical information such as max, min, total, average, trend, etc.

Data Mining for Financial Data Analysis (II)

- Loan ***payment prediction***/consumer credit policy analysis
  - feature selection and attribute relevance ranking
  - Loan payment performance
  - Consumer credit rating

- Classification and clustering of customers for ***targeted marketing***
  - multidimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group

Data Mining for Financial Data Analysis (III)

- Detection of money laundering and other financial crimes
  - integration of from multiple DBs (e.g., bank transactions, federal/state crime history DBs)

  - Tools:
    data visualization,
    linkage analysis,
    classification,
    clustering tools,
    outlier analysis,
    sequential pattern analysis tools
    (find unusual access sequences)

Data Mining for Retail & Telcomm. Industries (I)

- Retail industry
  - huge amounts of data on sales
  - customer shopping history
  - e-commerce, etc.
  - Nowadays also online data acquisition and mining possibilities
- Applications of retail data mining
  - Identify customer ***buying behaviors***, shopping patterns and ***trends*** and improve the quality of customer service
    - Both offline (aisle layout, discount policies, etc)
    - and online (real-time data analysis, user)

Data Mining for Retail & Telcomm. Industries (II)

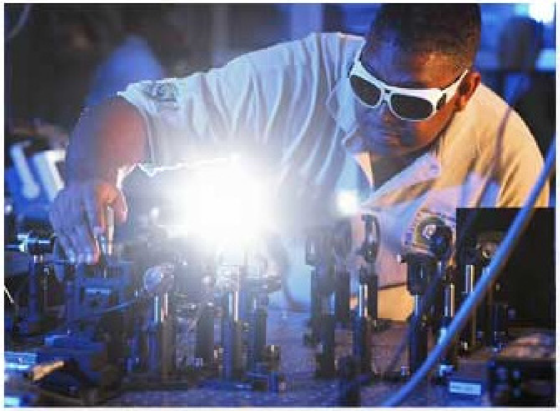
- Applications of retail data mining
  - Achieve better customer retention and satisfaction
  - Improve goods transportation and ***distribution policies***

- Telecommunication and many other industries:
  Share many goals and expectations of retail data mining

Data Mining Practice for Retail Industry

- Design and construction of data warehouses
- Multidimensional analysis of sales, customers, products, time, and region
- Use of visualization tools in data analysis
- Analysis of the ***effectiveness*** of sales campaigns
- Customer retention: Analysis of customer loyalty
  - Use customer loyalty card information to register ***sequences of purchases*** of particular customers
  - Use sequential pattern mining to investigate changes in customer consumption or loyalty
  - Suggest adjustments on the pricing and variety of goods
- Product ***recommendation*** and cross-reference of items
- Fraud analysis and the identification of usual patterns

Data Mining in Science and Engineering



- Data warehouses and data preprocessing
  - Resolving **inconsistencies** or incompatible data collected in **diverse environments** and different periods (e.g. eco-system studies)
- Mining complex data types
  - **Spatiotemporal**, biological, diverse semantics and relationships
- Graph-based and network-based mining

  Links, relationships, data flow, etc.
- Visualization tools and **domain-specific** knowledge
- Data mining in computer science: monitoring systems, software bugs, network intrusion

Data Mining for Intrusion Detection and Prevention (I)

- Majority of intrusion detection and prevention systems use
  - ***Signature-based detection***
    use signatures, attack patterns that are preconfigured and predetermined by domain experts
  - ***Anomaly-based detection***
    build profiles (models of normal behavior) and detect those that are substantially deviate from the profiles

Data Mining for Intrusion Detection and Prevention (II)

- What data mining can help
  - New data mining algorithms for intrusion detection
  - Association, correlation, and discriminative pattern analysis help select and build discriminative classifiers
  - Analysis of stream data: outlier detection, clustering, model shifting
  - Distributed data mining
  - Visualization and querying tools

Data Mining and Recommender Systems (I)

- Recommender systems:
  Personalization, making product recommendations that are likely to be of interest to a user

  - ***Content-based***
    Recommends items that are similar to items the user preferred or queried in the past

  - ***Collaborative filtering***
    Consider a user's social environment, opinions of other customers who have similar tastes or preferences

- …or a hybrid of both

Data Mining and Recommender Systems (II)

- Data mining and recommender systems
  - General: ***Customers C × items S***
    → extrapolate from known to unknown ratings to predict user-item combinations
  - ***Memory-based*** method often uses k-nearest neighbor approach
  - ***Model-based*** method uses a collection of ratings to learn a model (e.g., probabilistic models, clustering, Bayesian networks, etc.)
  - ***Hybrid*** approaches integrate both to improve performance (e.g., using ensemble)

Privacy, Security and Social Impacts of Data Mining

- Many data mining applications do not touch ***personal data***
  - E.g., meteorology, astronomy, geography, geology, biology, and other scientific and engineering data
- Many DM studies are on developing scalable algorithms to find ***general or statistically significant patterns***, not touching individuals

- The real privacy concern:
  - ***unconstrained*** access of individual records, especially privacy-sensitive information
  - ***Cross-database*** correlations involving personal data

Privacy, Security and Social Impacts of Data Mining

- How can we reduce the risk?
- Removing sensitive IDs associated with the data
- Data security-enhancing methods
  - Multi-level security model: permit to access to only authorized level
  - Encryption: e.g., *blind signatures*, *biometric encryption*, and ***anonymous databases*** (personal information is encrypted and stored at different locations)
- Privacy-preserving data mining methods

Data Mining Architectures are Hybrid Systems

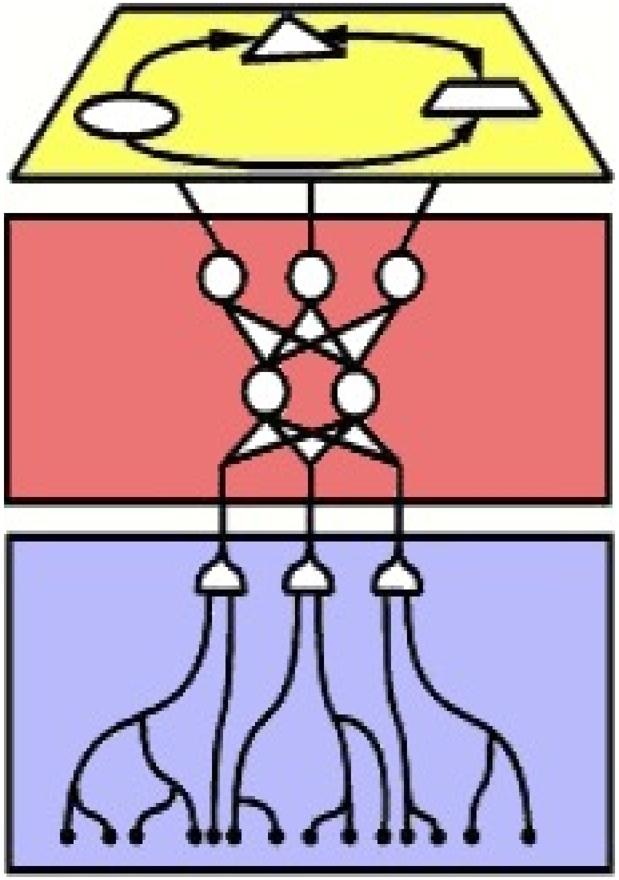
- Data Mining is always integrated into larger hybrid systems often combining several approaches
- ***Hybrid processing*** combines or integrates multiple modes of processing
- Hybrid systems in artificial intelligence and knowledge engineering for increasing performance

Benefit of Hybrid Representation Integration

	Neural/Statistical/ Sub-symbolic Data Mining	Symbolic/Structural/ Rule-based
Knowledge format	Numbers, Connections	Rules, Trees, Structure
Representation	Distributed	Local
Computational elements	Numerical associations Weights, Thresholds	Premises, Conclusions Rule strength, Predicates
Processing	Continuous numbers	Discrete symbols
Cognitive level	Low	High
Basic units	Neural Networks, Statistics	Rules
Manipulated by…	Continuous math	Logic
Representation	Compact but distributed	Verbose

NEST: NEural Symbolic Technology architecture for Hybrid Systems

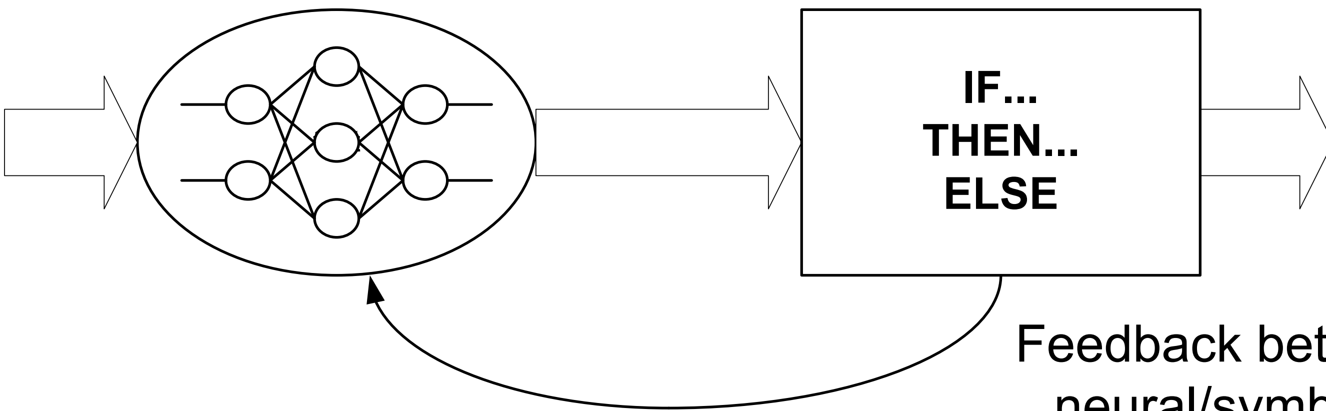
- ***Symbolic*** knowledge and understanding
- ***Neural/statistical*** knowledge representation
- ***Sensory input*** from several modalities (audio, vision, text, graphs…)



Modular Hybrid Systems

- Neural networks/statistical data mining and
- symbolic rule based systems cooperating

Information flow
maybe sequential or
parallel

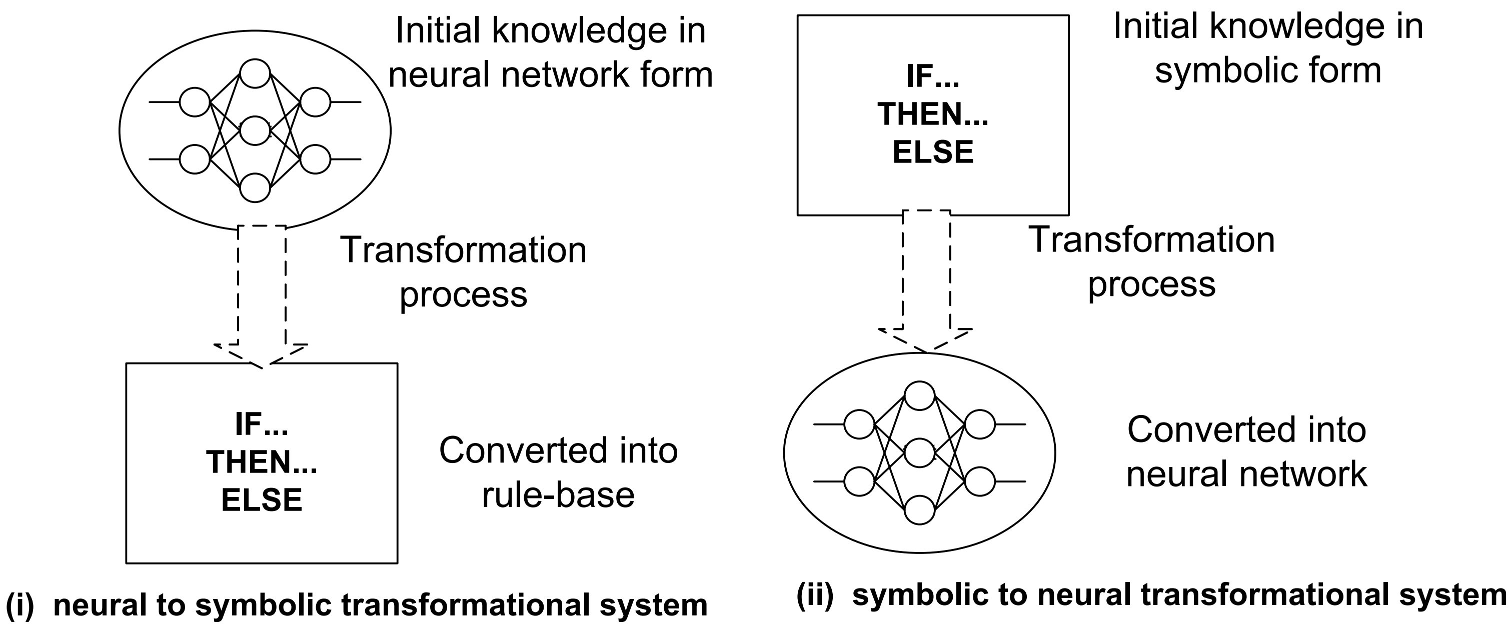


Feedback between
neural/symbolic
elements can occur

Modular Hybrid Systems

Transformational Hybrid Systems

- Converting neural / statistical representations into rule-based format and vice-versa

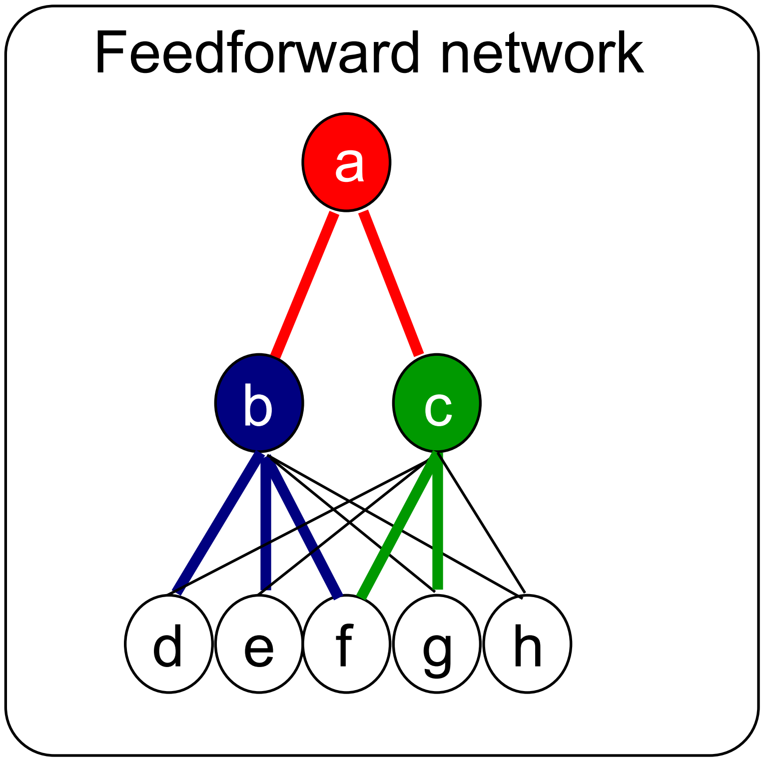


Transformational Hybrid Systems

Example: Transformational Rule Extraction (1)

- Weights contain the knowledge
- ***Problem:*** distributed representations are difficult to understand and modify
- Transfer of weights into symbolic rules
- Often used for feed-forward networks
- Grouping, elimination and clustering of weights
- N-of-M rules

Example: Transformational Rule Extraction (2)



Symbolic rules

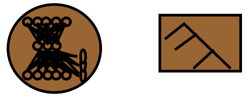
a :- b , c

b :- d , e , f

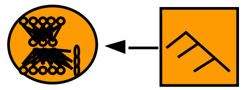
c :- f , g

Types of Coupling and Integration

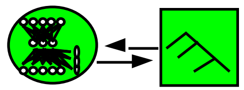
- Consist of symbolic/structural and neural/statistical representations at the same time



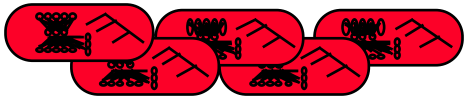
- Different forms of combination and integration
  - ***Loosely coupled***: symbolic/structural and neural/statistical modules separate and unidirectional communication



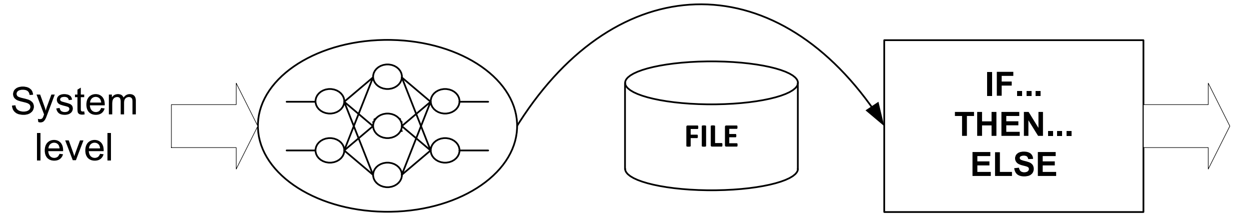
  - ***Tightly coupled***: symbolic/structural and neural/statistical modules separate and bidirectional communication



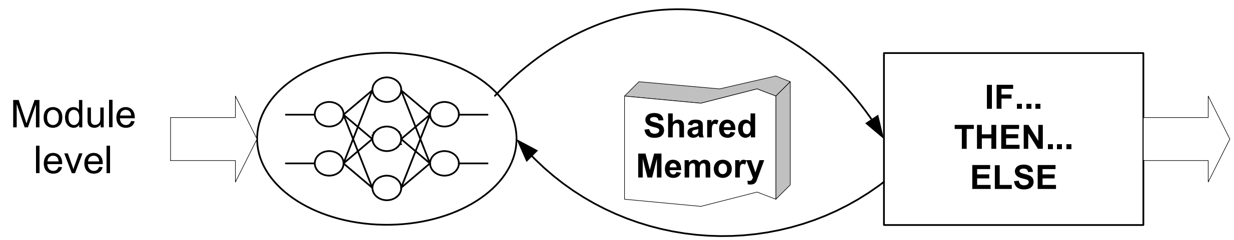
  - ***Integrated***: symbolic/structural and neural/statistical modules fully embedded and integrated



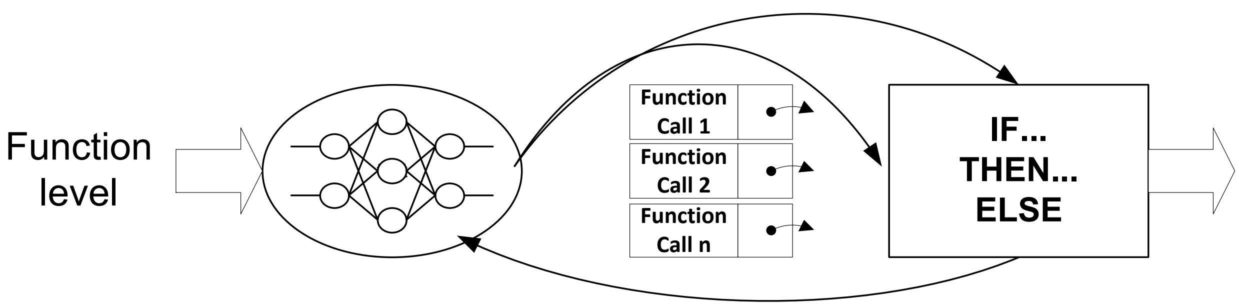
Implementation Strategies



Passively coupled by files

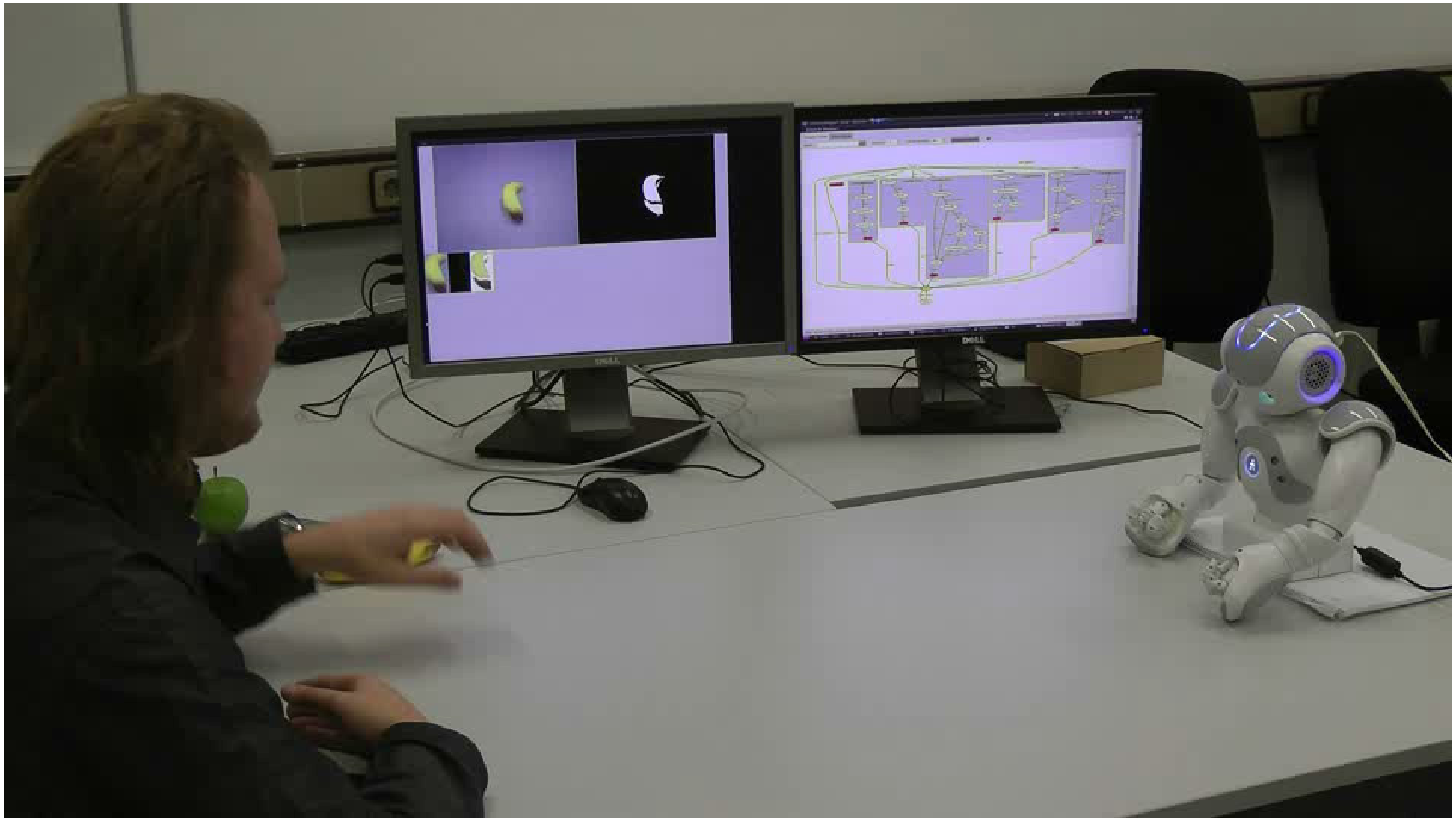


Actively coupled by shared memory



Interleaved by function calls

Recap: Example NEST Architecture

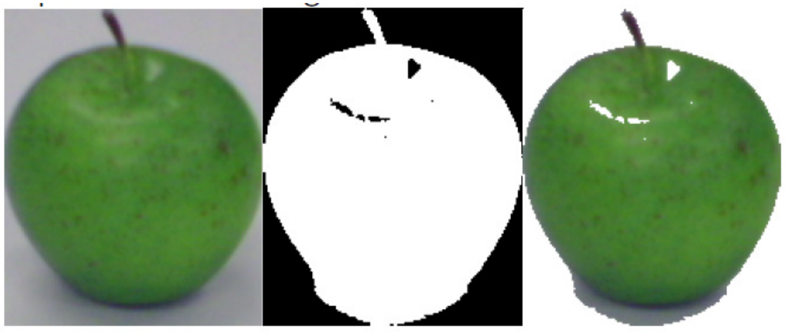


NAO data mining of objects based on an ***ensemble*** of ***neural networks in a symbolic architecture***

- Every network classifies based on ***different features***:
  pixel patterns, color & texture, or SURF features

http://www.informatik.uni-hamburg.de/WTM/teaching/WiSe11_HumanRobotInteraction_Pj.shtml

Example NEST Architecture



- Pre-processing of input data
  - Find region of interest
  - Extract features of object
  - Speech processing

- Learning to classify objects with 3 feed-forward neural networks

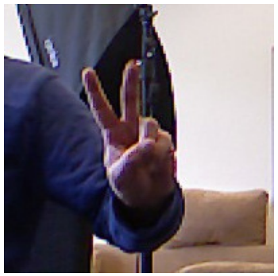
  →Ensemble Learning

- System controlled by symbolic state machine
  - Tightly coupled hybrid system

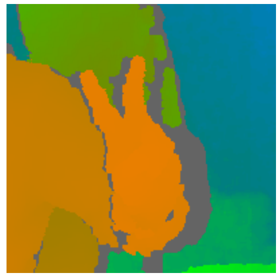
Data-Mining to Change Data Representation

- In a modular (hybrid) system, data-mining techniques incrementally can abstract from raw input data
  - Data representations change towards more abstract, symbolic data
  - Higher level data-mining can work on higher-level representations
  - Often several representations used in parallel to focus on different aspects

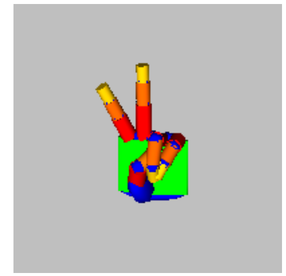
Example: Hand Model Fitting



Color



Depth



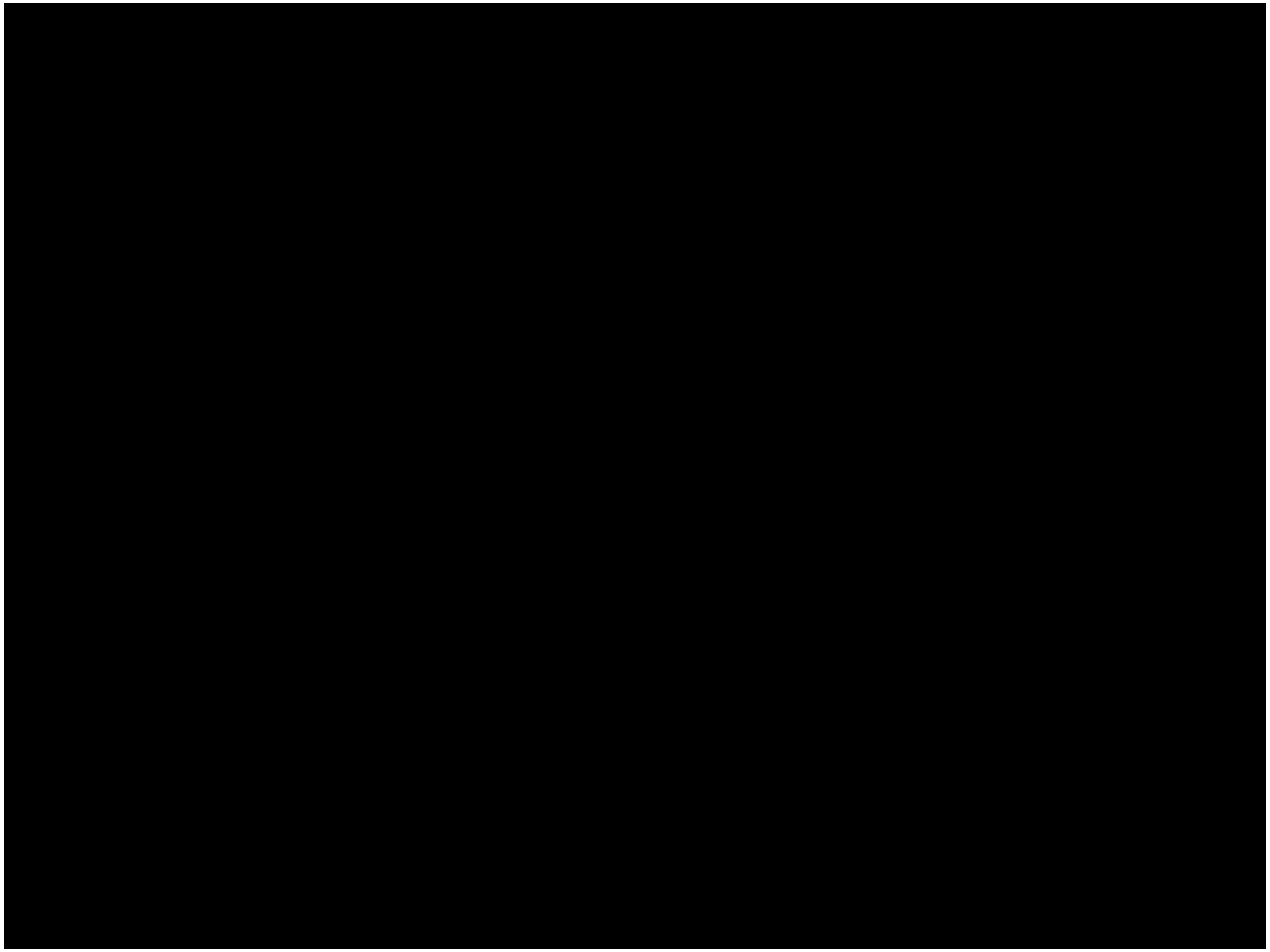
Hand Pose

- Raw RGB-D data
- Noisy
- Difficult to extract information on hand posture

- 26 DOF model of kinematic state
- Angles between joints plus orientation and position in space
- High-dimensional, high-level representation

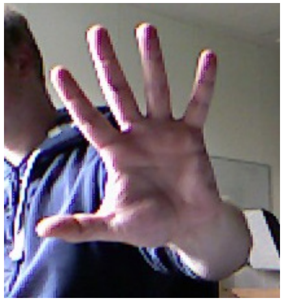
Hamester, D., Jirak, D., Wermter, S. Improved Estimation of Hand Postures Using Depth Images, 2013

Hand Model Fitting

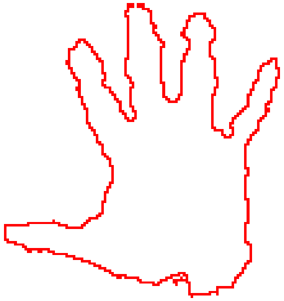


Detect the Hand

- To track the hand, it has to be detected
- Input: RGB-D data



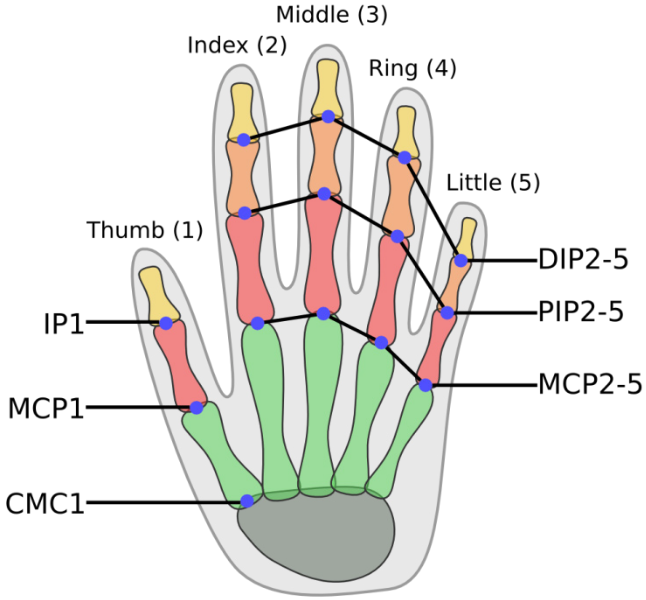
- Extract contour candidates
- Normalization
  - (affine transformations, Fourier descriptors)
- Classification with SVMs

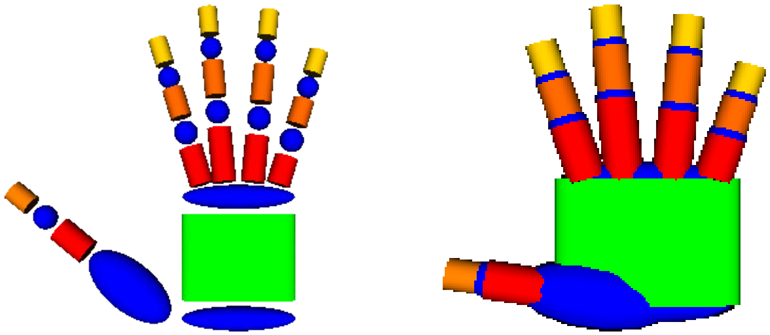


- Output is the position and classification as hand/no-hand

Hand Pose Model

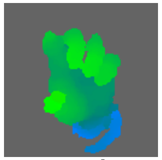
- A hand model can now be fitted to detected candidates



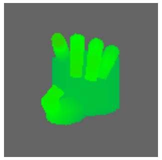


- Hand pose can be modelled by 20 DOF for joint angles plus 6 for orientation and position in 3D space
- Rendered by 12 ellipsoids and 15 elliptical cylinders

Fitting the Model to the Data

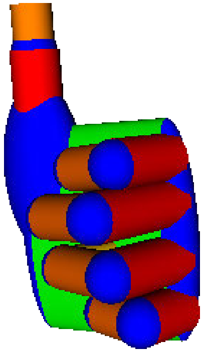


(Observation O)



(Rendering R)

- Perform PCA to reduce dimensionality
  - first 12 dimensions explian 99% of variance
  - Axes now are eigenvectors
- Render a hand model of current hypothesis in 3D and compare with observed data
  - Sum of values in difference image $|O - R|$
    → Estimate of quality
  - Problem now an optimisation problem
    → solved using particle swarm optimisation (PSO)



Complex Hybrid Architectures in HRI

- Case Studies of a complex HRI scenarios
- Such systems usually incorporate
  - Sensory processing of noisy auditory, proprioceptive, and visual input
  - Data generally high-dimensional and with temporal component
  - Classification and clustering of such data for object/person and speech recognition
- What to data-mine from? Raw input?
  - Decide on level of abstraction to reduce complexity for single task

Representing and Mining Motion

- Movements are represented as sequences of body postures in time
  - Body posture is a set of joint locations
  - Often used: 3D Skeleton model
  - Several algorithms available

- One action defined as sequence of postures over time
  - How to segment continuous motion into actions?

Florian von Stosch, MSc Thesis, in preparation

Learning Motion Sequences

- Sequences can have different lengths
  - One approach: Only take sum of history into account

- Recursive SOM to incorporate predecessors
  - Difference vector at step n now has two weighted parts:

$$y_i(n) = (1-\alpha)y_i(n-1) + \alpha(x(n) - w_i(n))$$

  - Behaves like a normal SOM for $\alpha = 1$
  - Towards $\alpha = 0$, all units tend towards mean of input

Learning Motion Sequences

- Graphical representation of 2D SOM
- Weight vectors represent body postures
- Clustering of similar postures evident

Learning, Recognizing and Naming Actions

- Human 3D motion tracking
  - Extraction of ***spatio-temporal properties*** from moving targets
  - Use of depth and color information

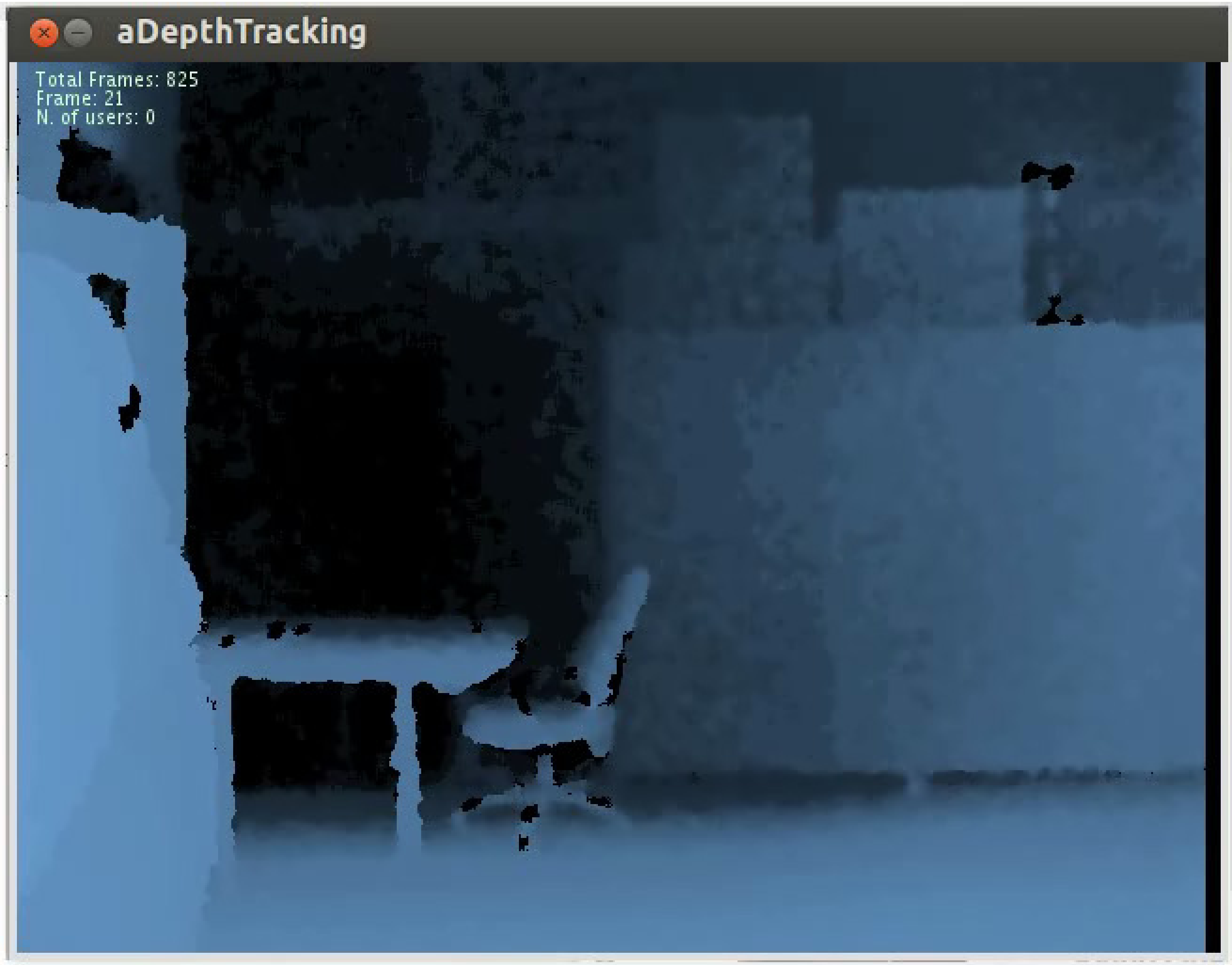
- Unsupervised novelty detection
  - Neural-statistical architecture
    based on self-organizing maps (SOM)

- Challenges:
  - Robust to changes in light conditions
  - Highly occluded targets

- How to represent human motion?

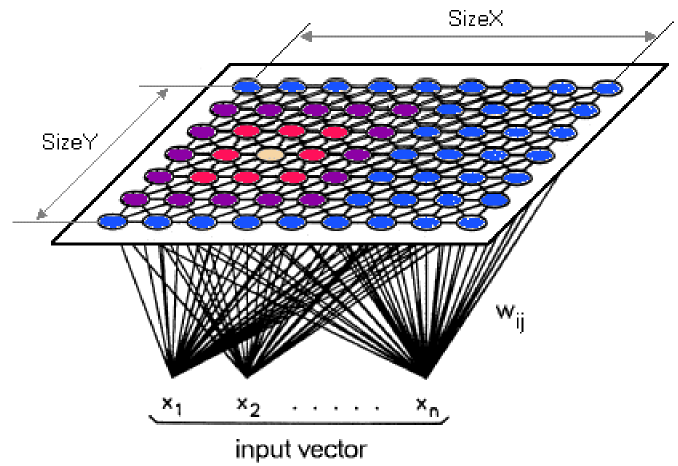
Motion Representation

Detecting Normal Actions – like Standing and Walking



Detecting Abnormal Actions – like Falling

Modular Neural Architecture



- SOM-based neural architecture

- What is a normal action?
- Degree of novelty is larger than the given threshold, the observation is reported as abnormal

Learning stage:

Motion Extraction

Outlier Preprocessing

Trajectories clustering | Body features clustering | Directions clustering

Detection time:

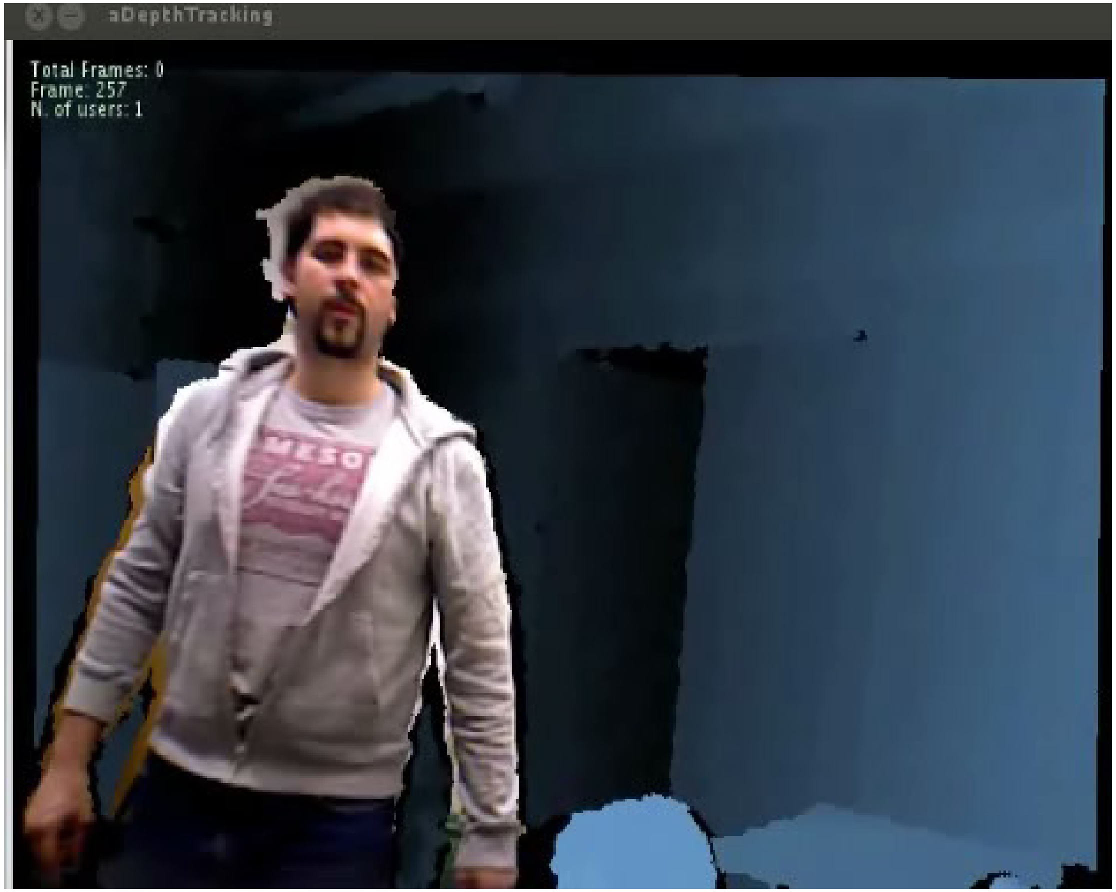
Motion Extraction

Trajectories Novelty detection | Body features Novelty detection | Directions Novelty detection

Weighted voting

Active Following:
Using Action for Perceiving

- Moving tracked target followed around the environment



ASUS Xtion Pro Live sensor
on humanoid NAO
1 user actively tracked

Quick Summary

- Data-Mining modules often part of a hybrid structure
  - Different types of coupling between neural/symbolic modules
- Questions that have to be answered for each module
  - What data representation do I need as input/output?
  - Which type of system is „best“ for each task?
- Problem with real-world data: Noise?
  - How do I incorporate unreliable observations in my system?
  - What effect do such observations have over time?

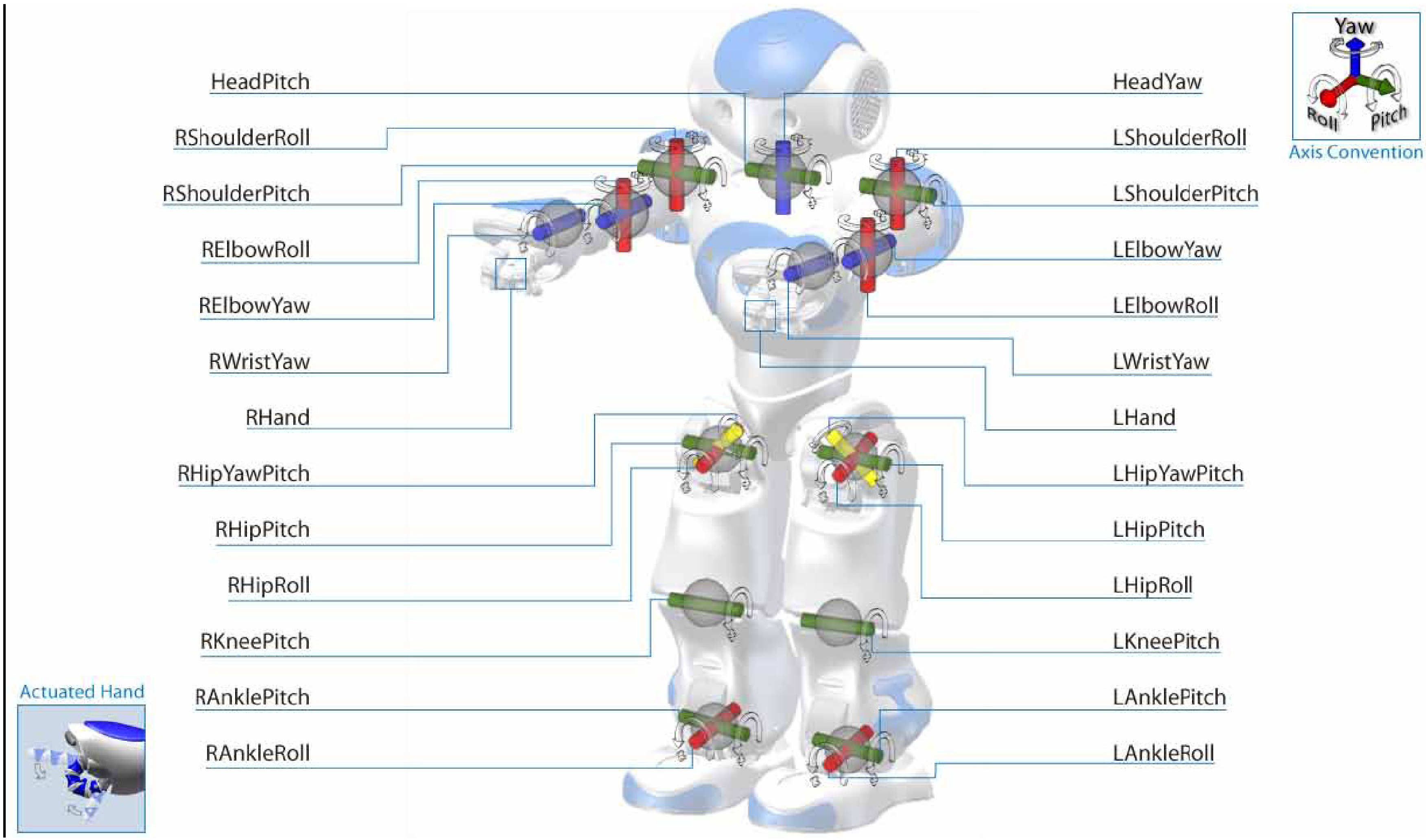
Case Study: Localisation and Navigation

- Typical task for a robot in a complex environment
  - Find the exact location
  - Navigate safely to a different location

- Hybrid systems that use different modalities
  - Detect objects or obstacles to aid localisation or navigation
  - Use of visual or proprioceptive information

- Environment is complex and dynamic
  - What features to extract and use?

Robot Mobile Behaviour

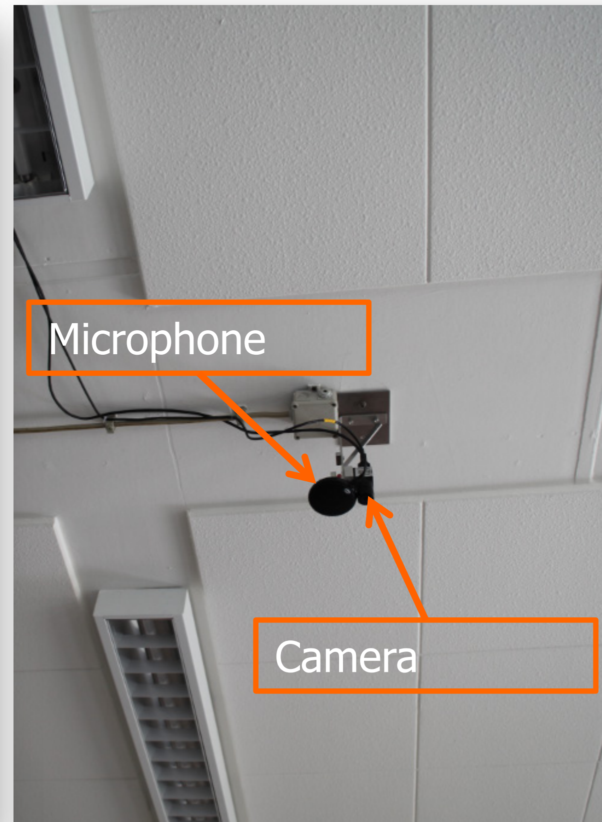


Experimental Environment

- Domestic Environment with different objects

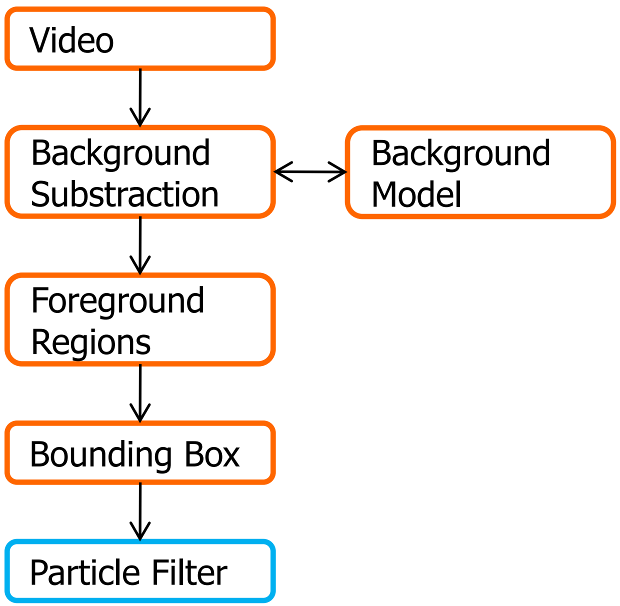
Experimental Environment (Cont.)

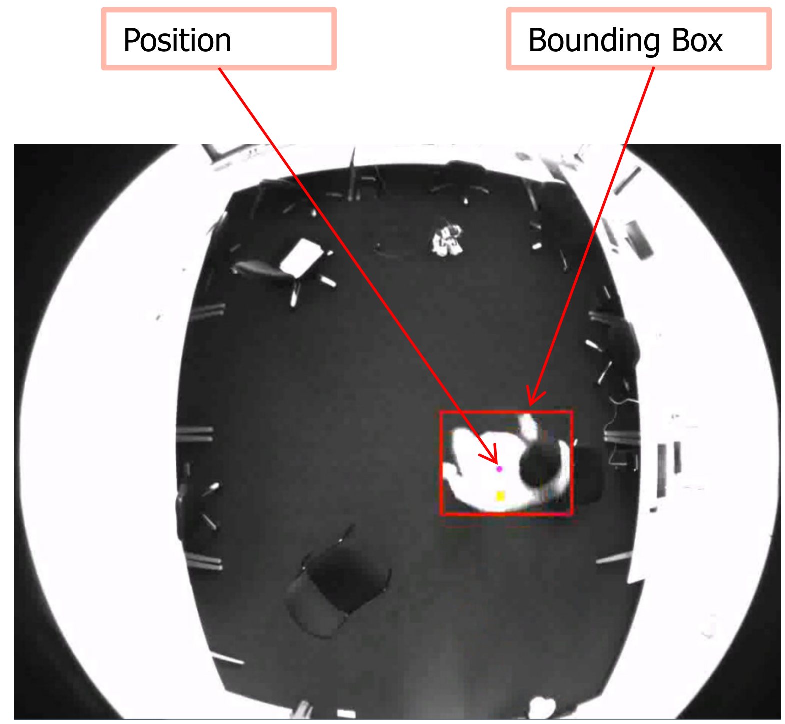
- Ceiling-mounted Camera & Microphone



Person Recognition for Approaching: Decide what to data-mine from

- Find the position of user or Nao using ceiling camera with fish-eye lens





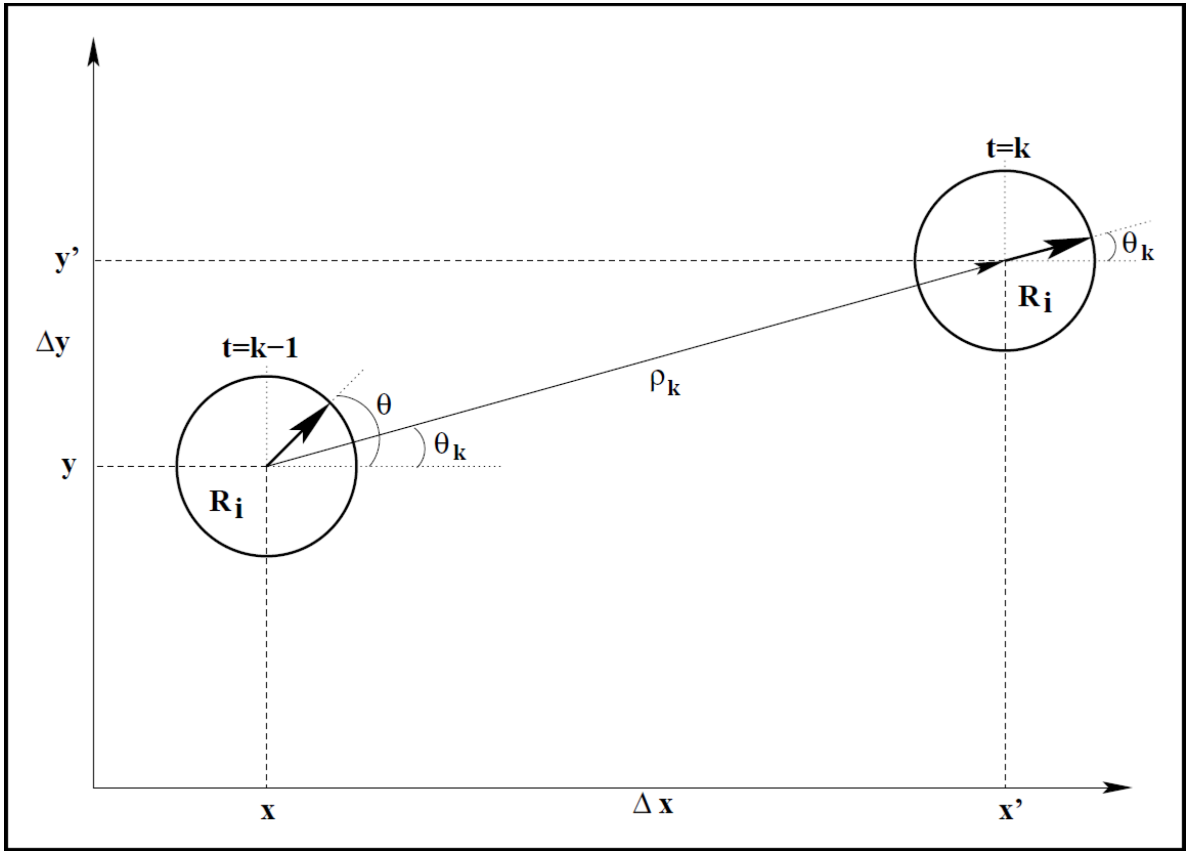
Particle Filter

- Main Aim:
  - Track a variable of interest as it changes over time

- Idea:
  - Represent the variable by a sample of possible states (particles)
  - Each particle represents a „belief“ taking into account previous observations
  - Once I have an observation, select particles that are most likely to be correct
  - and continue.......

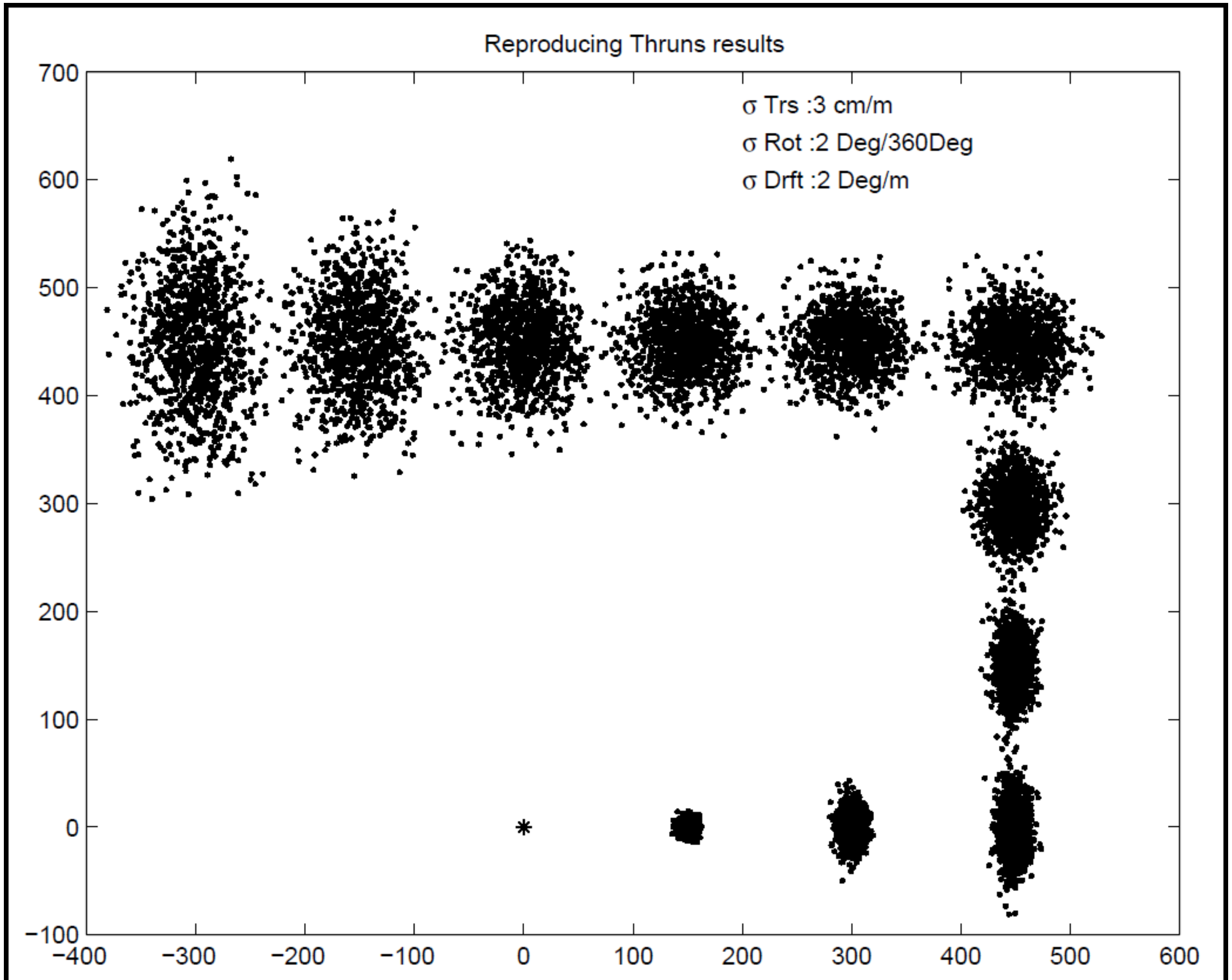
Particle Filter for Moving Robots

- Divide movement of robot into rotation and translation



Rekleitis, Ioannis M. "A particle filter tutorial for mobile robot localization.", *McGill University, Tech. Rep.* 2004.

Particle Filter for Moving Robots



- Each action introduces uncertainty, leading to accummulating error in position estimate

Rekleitis, Ioannis M. "A particle filter tutorial for mobile robot localization.", *McGill University, Tech. Rep.* 2004.

Modelling Uncertainty in Data

- Difficult to know noise
  - Particle *P* usually modelled with ***Gaussian noise*** with mean μ and variance σ:

Position $\mathbf{x}$ is a vector over coordinates *x* & *y*, and the angle θ

$$P(\mathbf{x}_{m_i}^{t+1} \mid \mathbf{x}_s, \mathbf{z}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mathbf{z}-\mu)^2}{2\sigma^2}}$$

Position of particle i at next time step

Position of robot

Estimated tracker measurement

Gaussian white noise

  - Quality of estimate depending on used variances
    - Could be fixed…
    - …or dynamic over the position:

$$\sigma^{t+1} = h(\mathbf{z},\sigma) = \mathrm{asin}\left(\sigma \Big/ \sqrt{dx^2 + dy^2}\right)$$

Particle Filter Algorithm

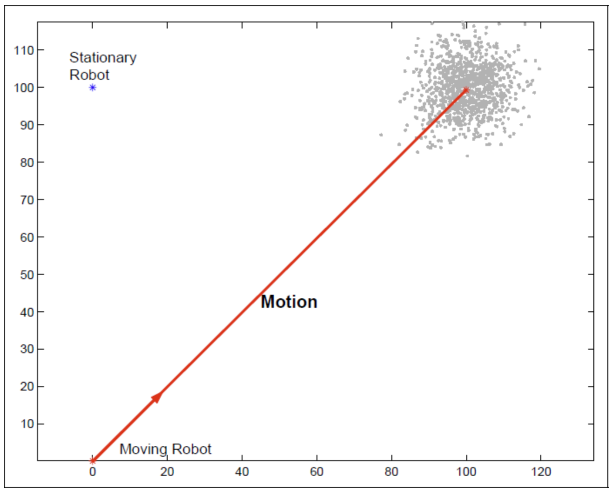
1. Initialise particles randomly
2. For N steps do
    1. For all particles p do
        1. If number of particles < threshold: Resample
        2. Update particles
        3. Change weights depending on observation
        4. Normalise weights

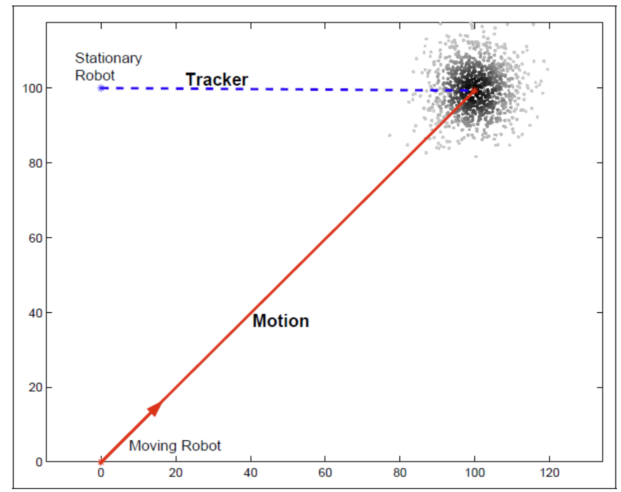
- Weight of particle = Level of certainty

Particle Filter in Action

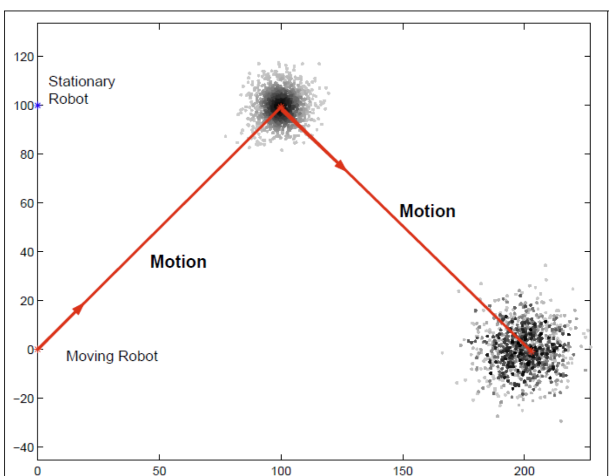
Prediction phase

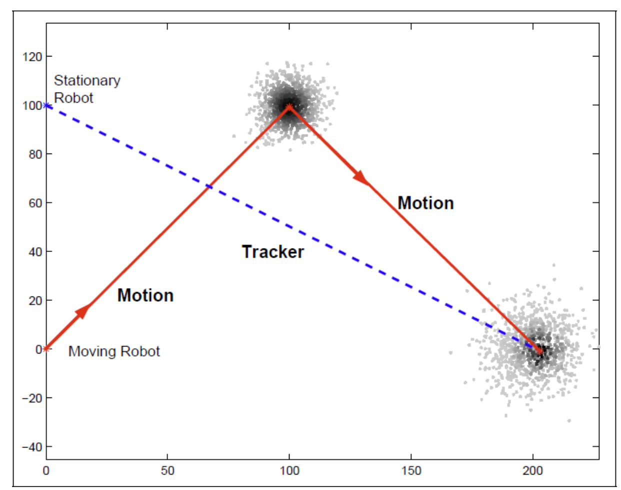
Update phase





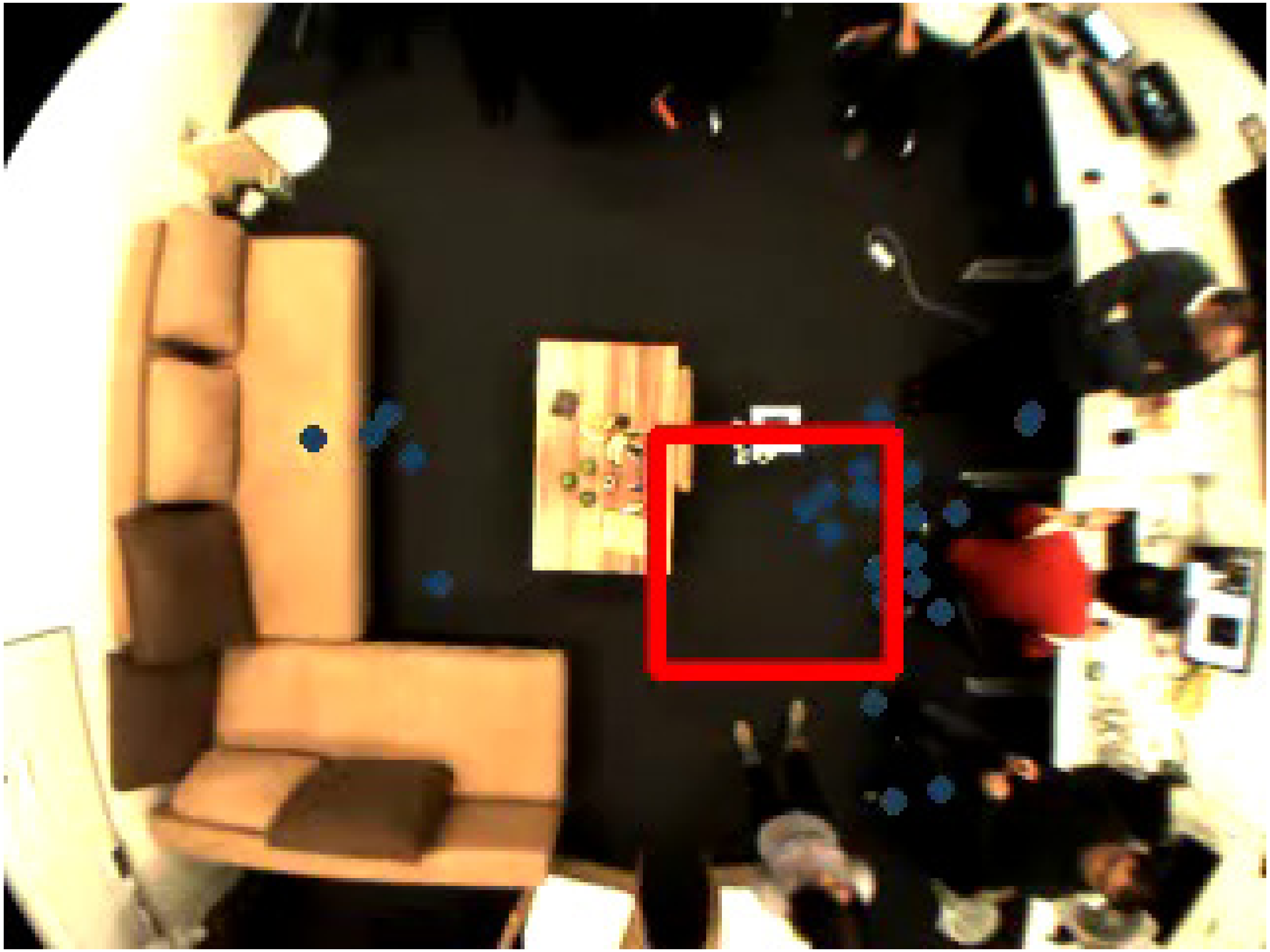
First
step





Second
step

Particle Filter for Person Tracking



Yan, W., Weber, C., Wermter, S. A hybrid probabilistic neural model for person tracking based on a ceiling-mounted camera. Ambient Intelligence and Smart Environments, Vol. 3(3), pp. 237-252, 2011.

Particle Filter for Person Tracking

- How to estimate uncertainty in person detection?
  - Bounding box changes with body movements

- Predict positions only by Gaussian distribution:

Beware! Different notation!

$$P(\mathbf{s}_t \mid \mathbf{s}_{t-1}^{(i)}, a_{t-1}) = \frac{1}{\sqrt{2\pi\sigma(a)^2}} e^{-\frac{(\mathbf{s}_{t-1}^{(i)} - \mathbf{s}_t^{(i)})^2}{2\sigma(a)^2}}$$

Assuming Gaussian noise again!

State (position) of person

State of particle i at previous time step

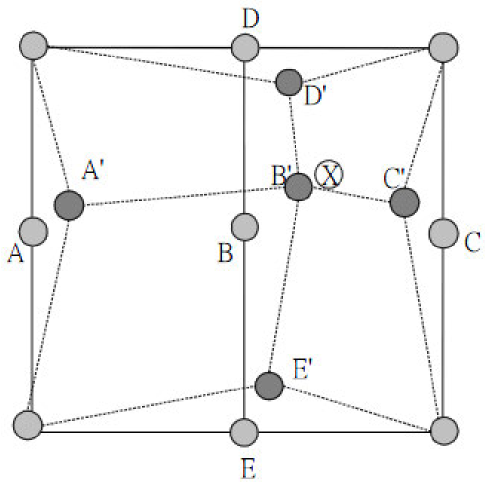
Previous selected action

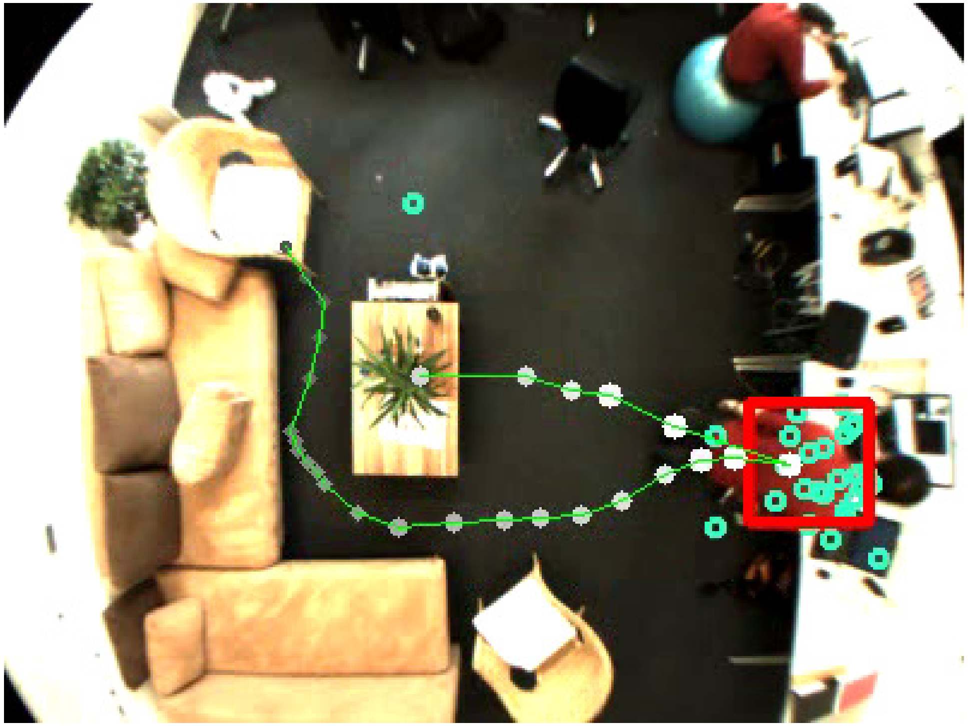
Variance is function over the action

- where $\sigma(a)$ can take one of two values
  - high if motion is detected
  - low if no motion is detected

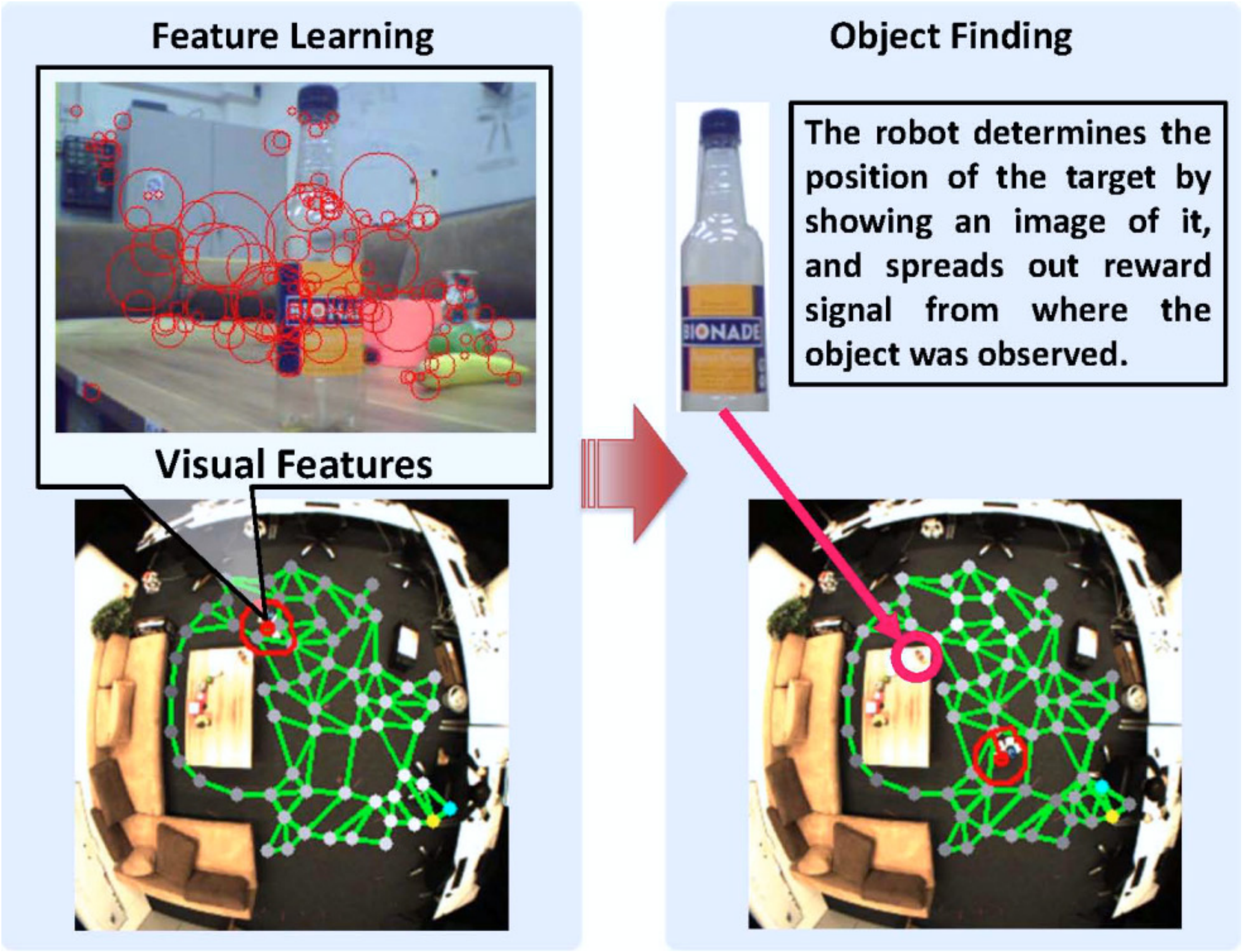
Integration of Color, Shape, & Movement Cues

- Tracking movement of user and robot to plan navigation
- Growing neural gas algorithm for cognitive map learning

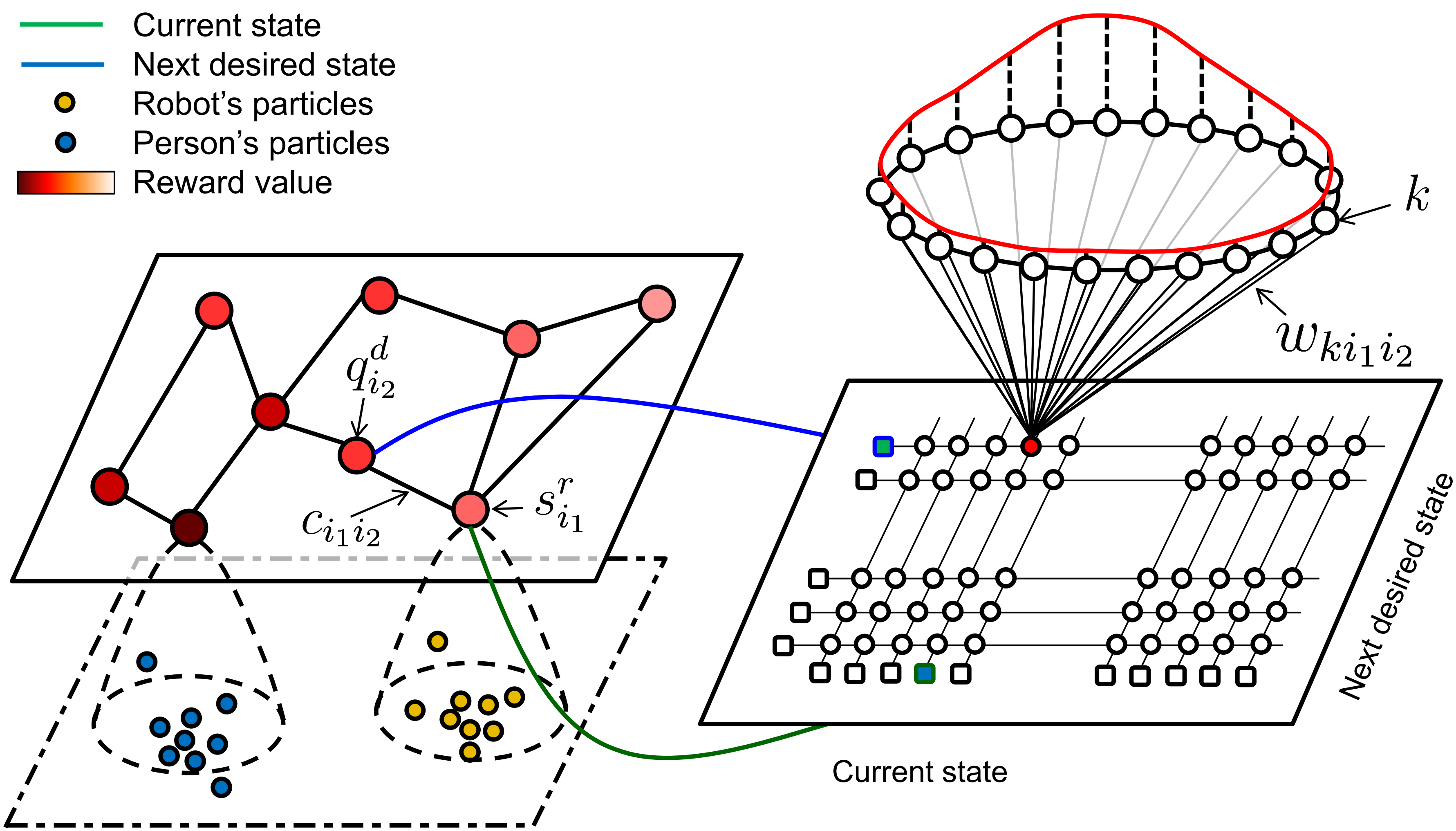




Anchoring Appearance Features at Map Nodes

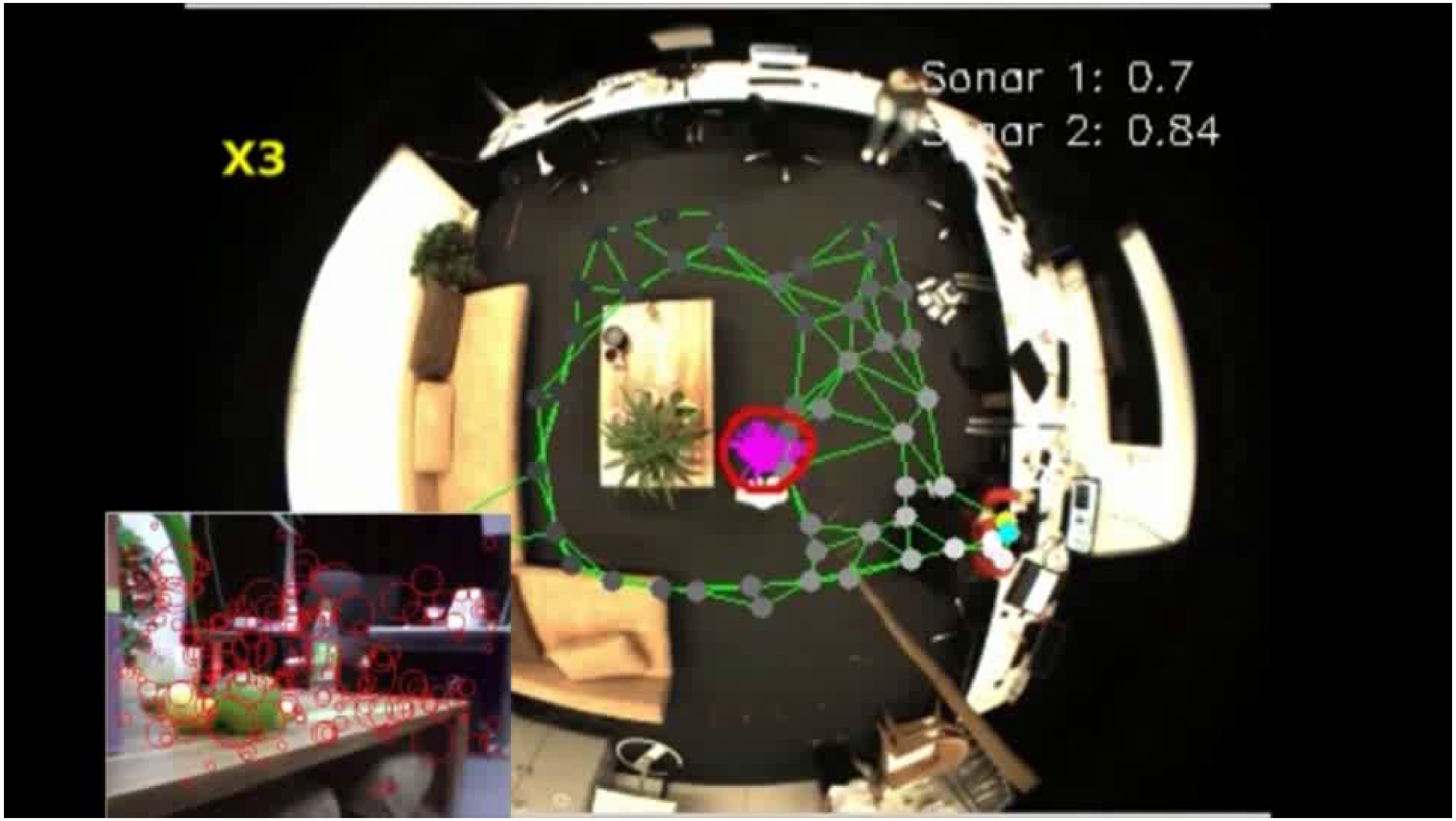


Architecture: Neural Gas and Neural Fields



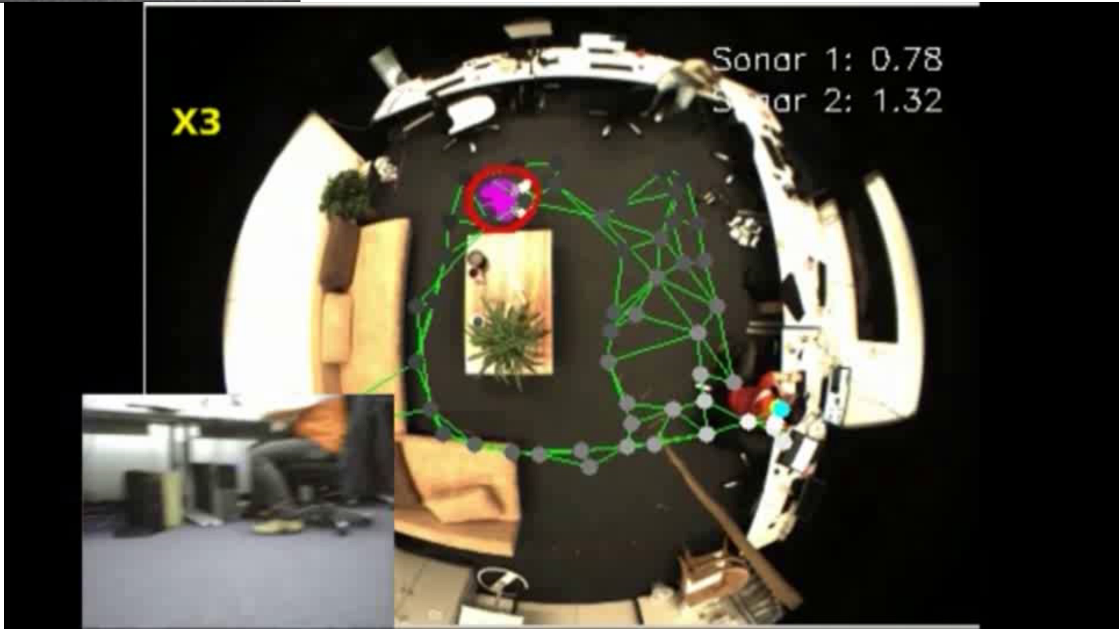
Yan, Weber, Wermter 2013

Building the Map and Storing the Features at Neurons of the Map



Person Tracking and Robot Navigation based on Ceiling Camera

Grasping Bottle and Bringing to Person



Summary

- Data-mining has many fields of application
  - Often domain-knowledge and specific representations used
- Mostly used as part of larger architectures
  - Hybrid approaches combine advantages of neural/statistical and symbolic techniques
  - Data mining to change low-level representations incrementally towards high-level, symbolic representations
- More examples can be found in the publication list on the Knowledge Technology website