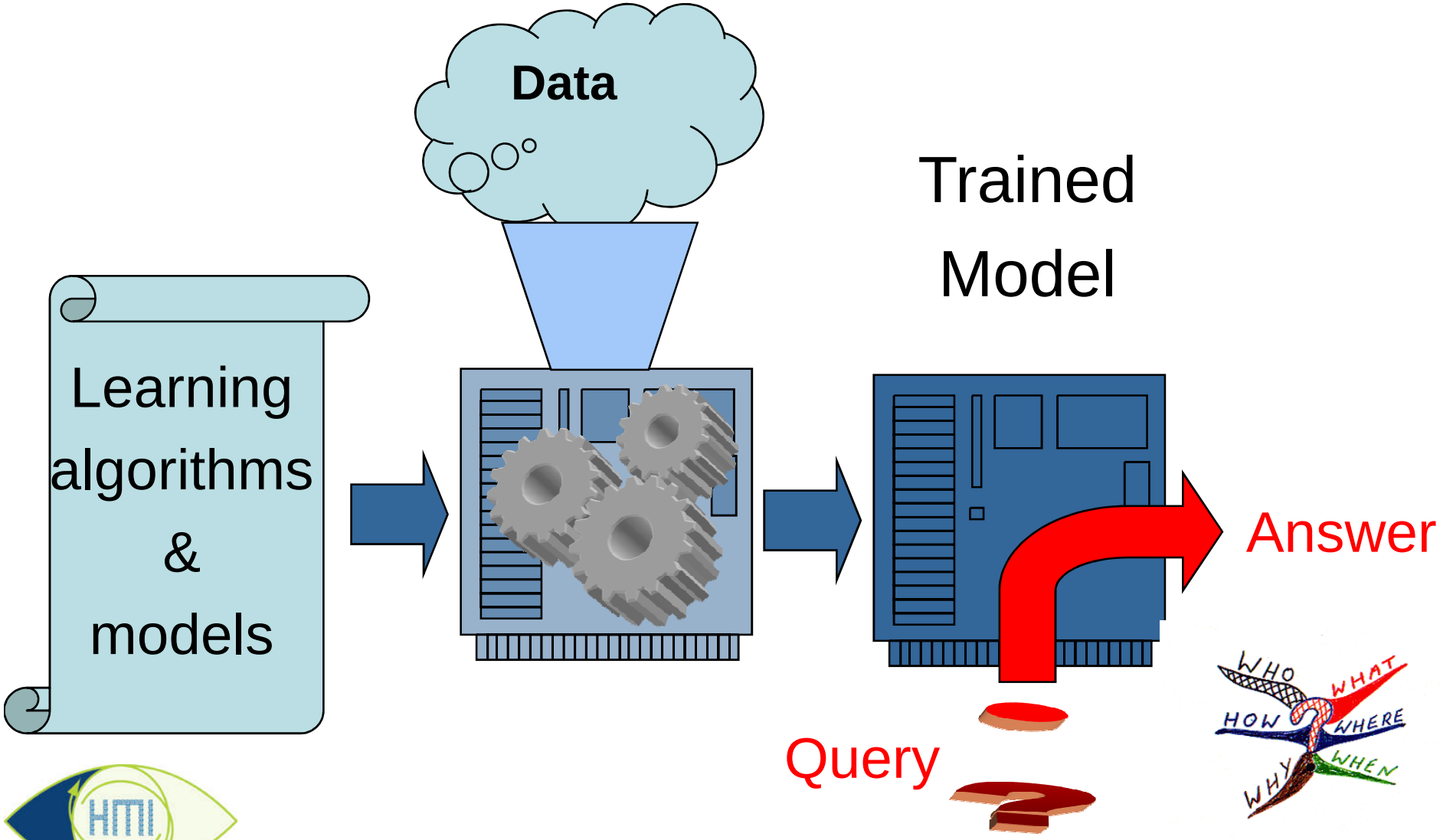
Introduction to Machine Learning

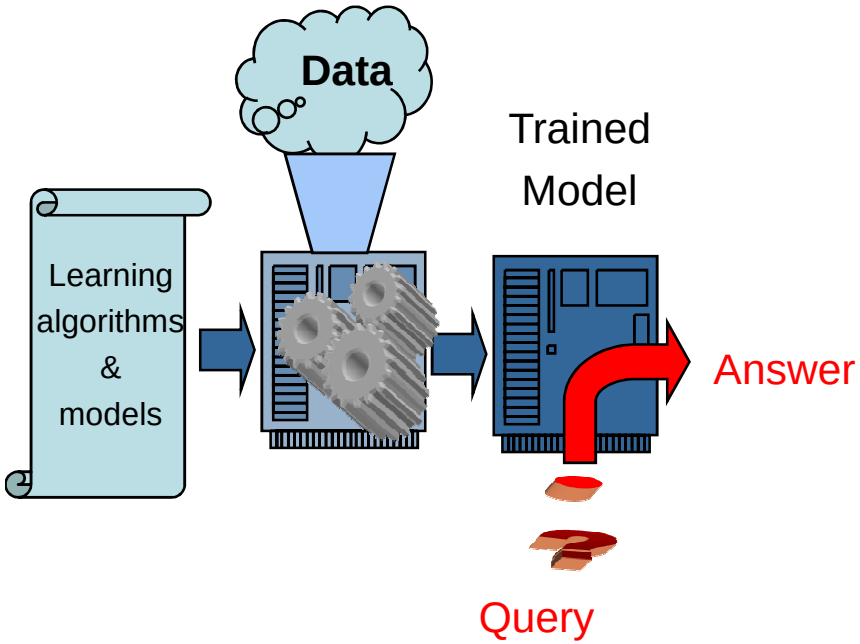
A case based approach

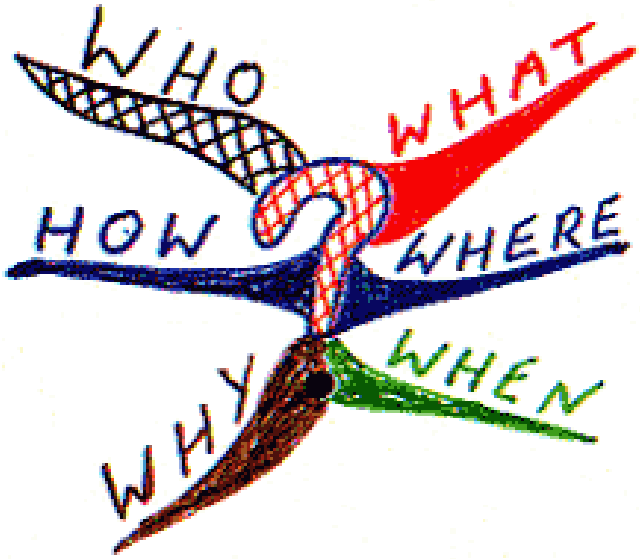
Mannes Poel
University of Twente
Dept. Computer Science
Human Media Interaction

What is Machine Learning?



What is Machine Learning?





What is Machine Learning

- Learning by Machines from Data.
- But there is No Free Lunch, meaning that there is no learning algorithm which performs best on all ML problems.
- A.k.a. Pattern Recognition, Data Mining, ….

Some applications in Human Media Interaction

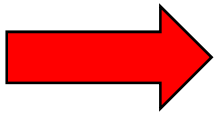
- Speech Recognition
- Body:
  - Body detection
  - Pose estimation
  - Gesture recognition
  - Action recognition
- Head:
  - Head detection
  - Head pose estimation
  - Visual Focus of Attention
  - Facial expressions

Main problem for a naturalistic setting is the data:

- How to get naturalistic data?
- Annotations
- Reliability of annotations

Human Media Interaction

Data in Machine Learning

- It is the interplay between the domain which generated the data and the learning algorithm which determines how good the model is!
- Assumptions on the data:
  - The data is good predictor of the future.
  - The data is or can be made available.
  - The data contains what we want to predict.
- Garbage in  Garbage out!!!!!

Case 1: Detecting heads

Challenge

Given a picture or a video stream, can we detect faces in the image (video stream)

IS1001c.L_24000.jpg

IS1001c.R_24000.jpg

Head detection: Possible approaches

- Shape based:
  - Head as a certain shape, detect such shapes in the picture.
- Skin color or appearance based:
  - Detect skin blobs in the image
  - Largest skin blobs are heads
- Combination of both
- Other

Focus: Skin color based approach

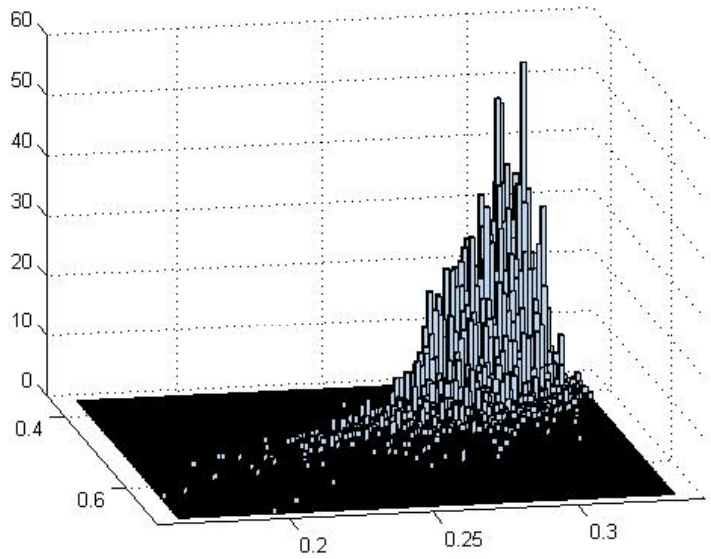
Model: Skin color detector

- Color pictures are coded in RGB values.
- Machine Learning problem is to construct a model *M* which classifies the color of a pixel *x.*
- The possible classes are *skin* or *not skin*.
- Need of data to train (construct) the model *M.*

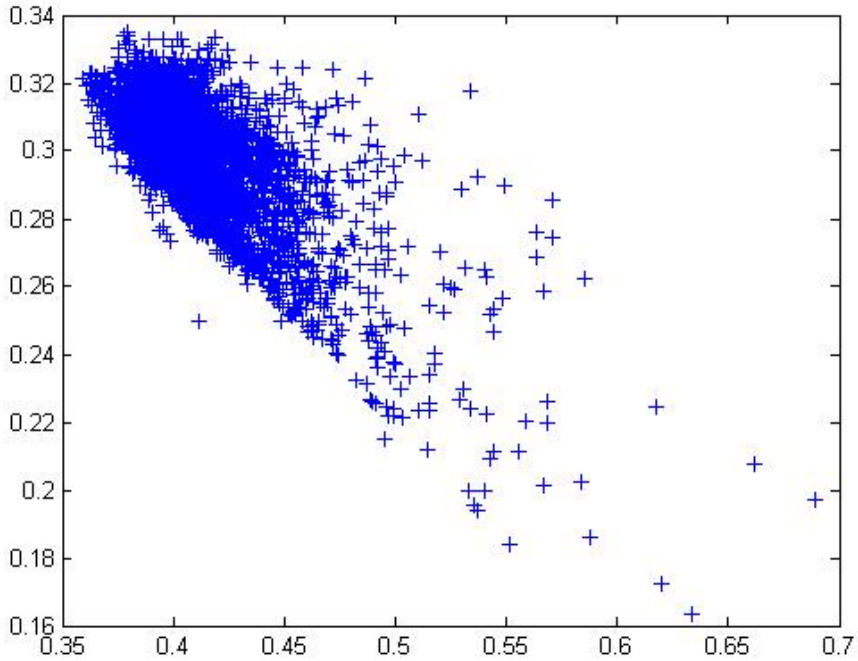
Skin color detector

- Use data:
  - Generate skin blob regions from your train data.
  - Determine normalized RGB value for each pixel in the skin blobs, this reduces illumination differences.

Visual inspection of skin color data

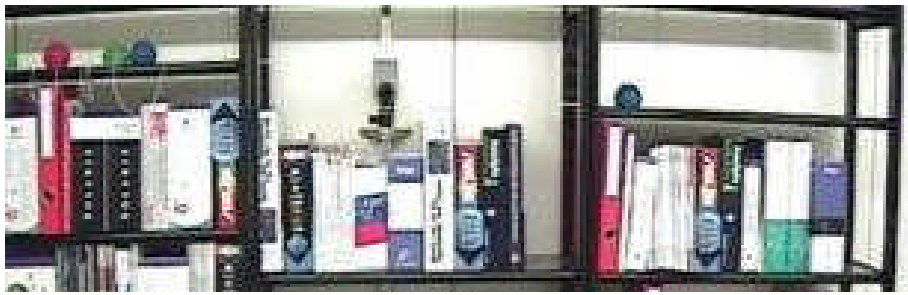


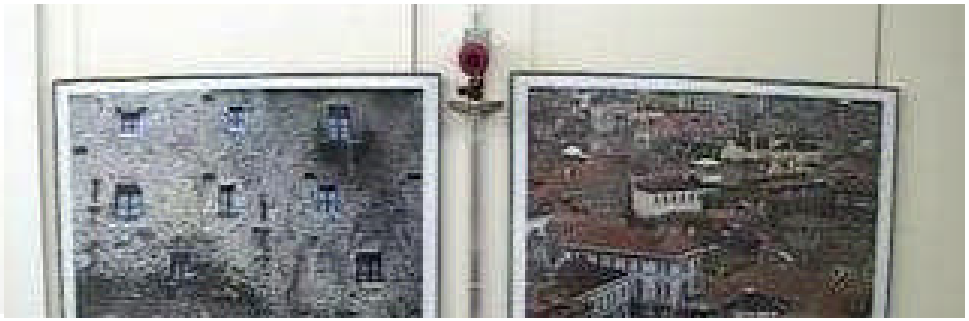
100 x 100 Histogram



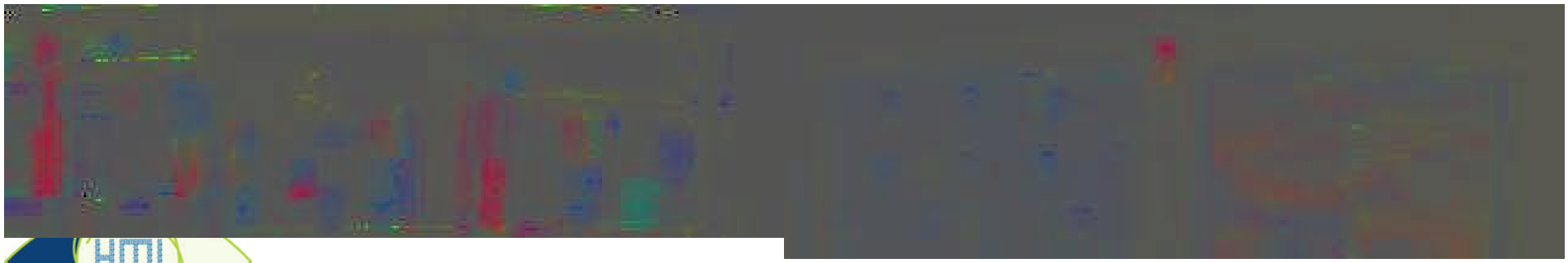
Data distribution

Negative examples

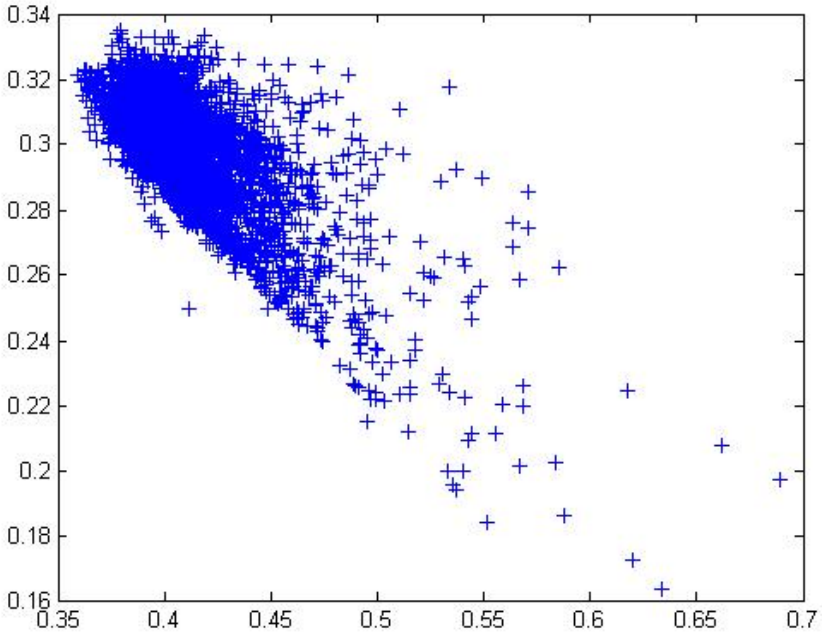




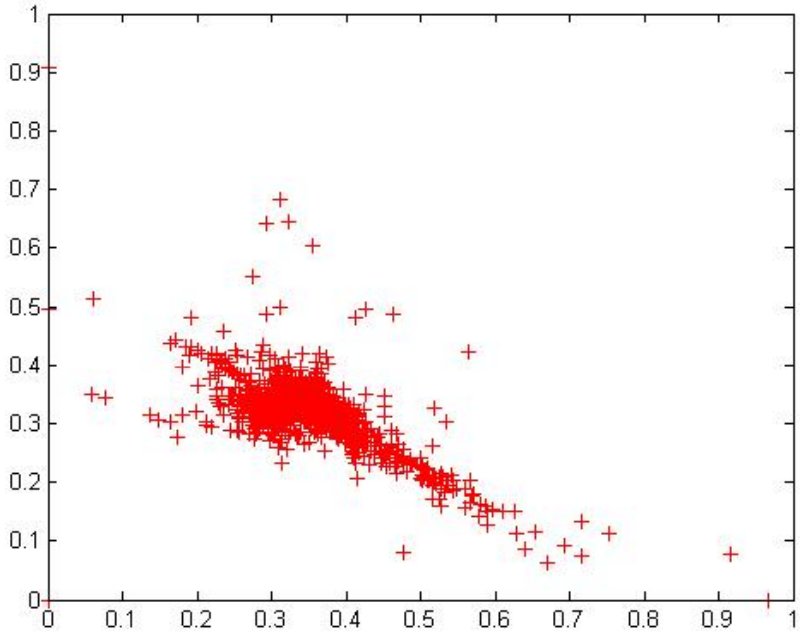
Normalized version



Data visualization



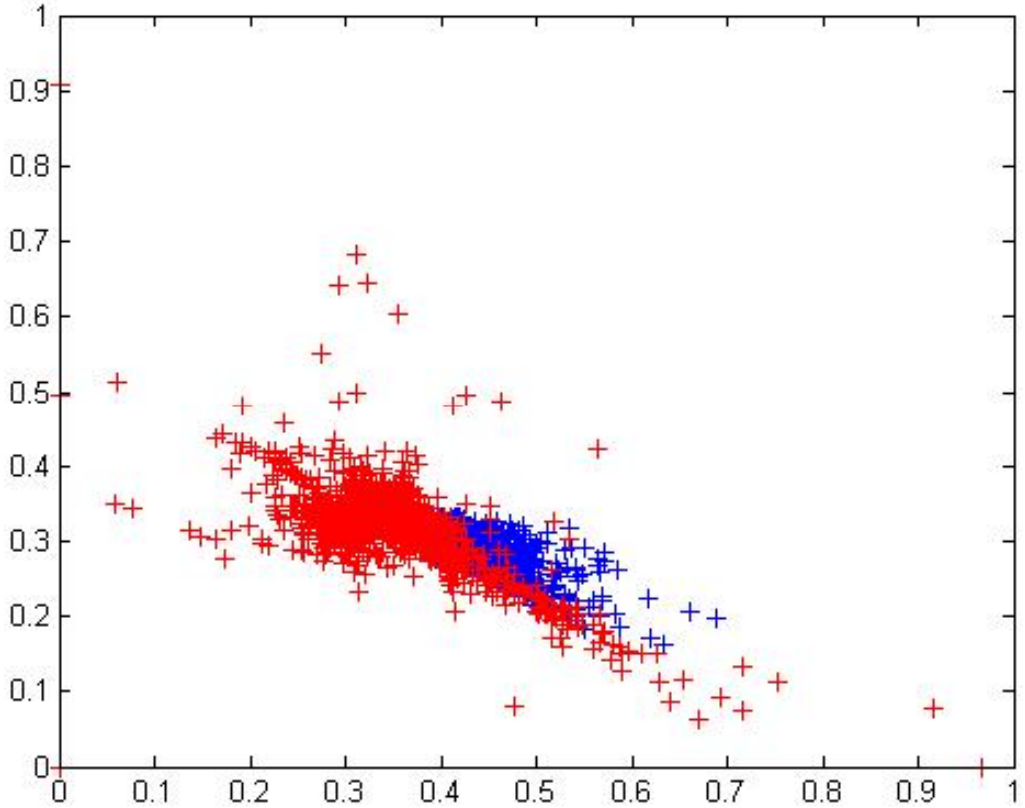
Skin



Non skin

Observe the different scales of the axis

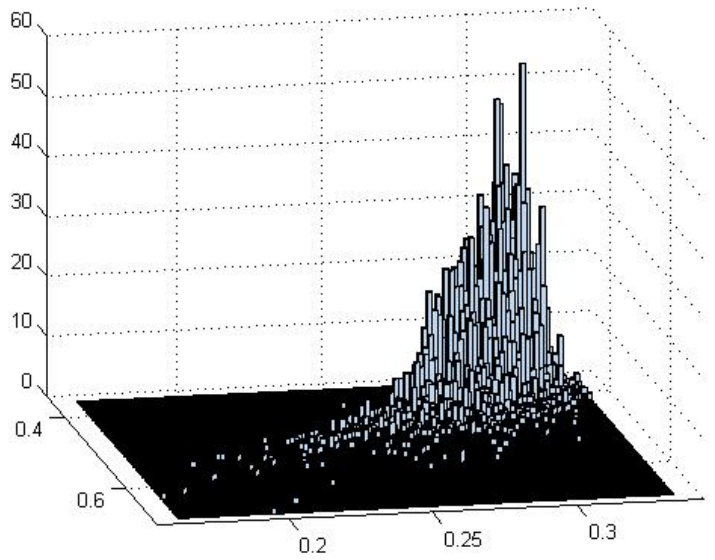
Data visualization: both classes



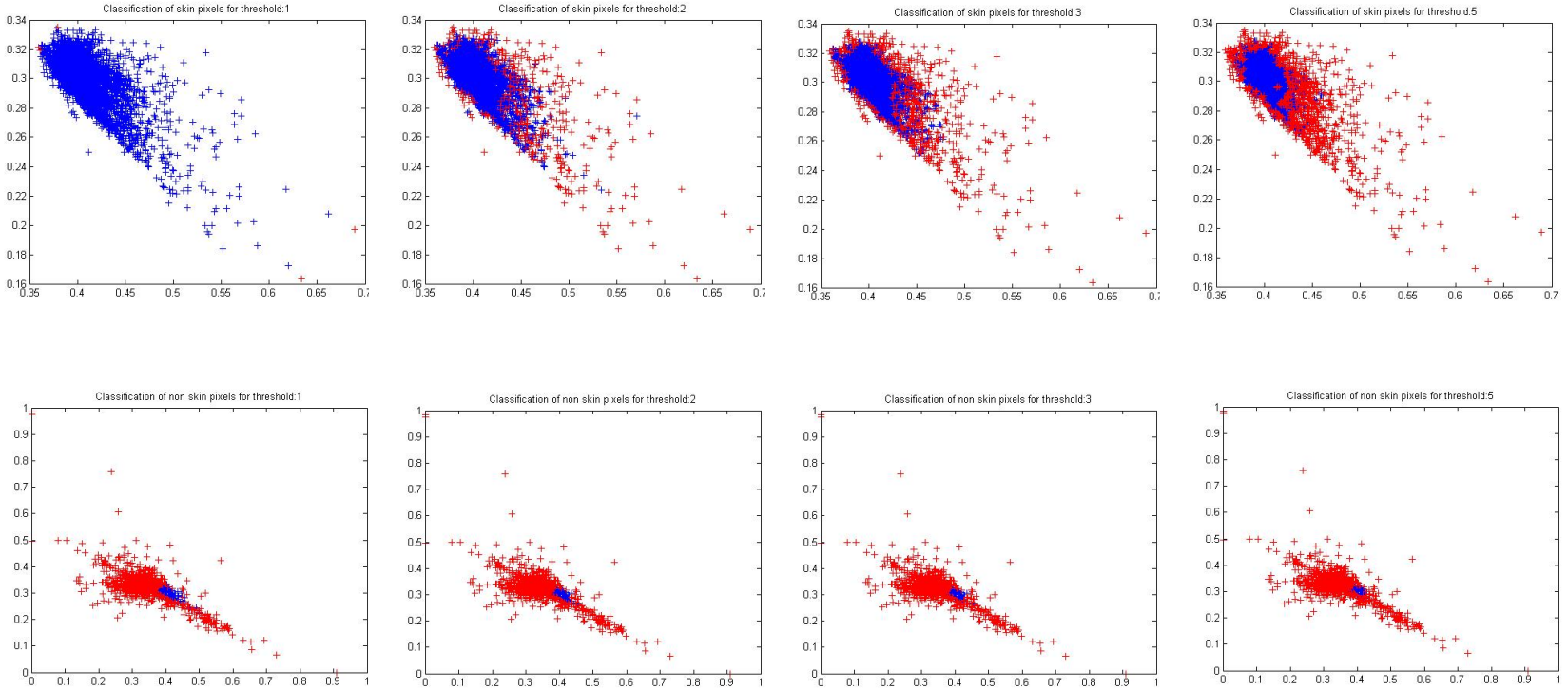
Conclusion: skin color classification is non trivial

Histogram based approach

- Given pixel *x* with value *(r,g,b)* determine the bin of *x*. If *bin value* > *t*, with *t* threshold then *x* is classified as skin.
- Threshold *t* is determined on train set:
  - select *t* with optimal performance on train set.
  - use visualizations
  - this can also be done with the help of Receiver Operator Curve (ROC). Plot Hit rate against False Alarm rate



Visualization: data level



threshold =1 threshold =2 threshold =3 threshold =5

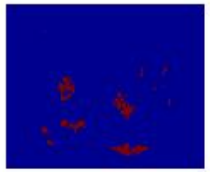
Human Media Interaction

Visualization

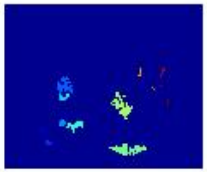
threshold = 5

raw image

skin pixels



skin blobs

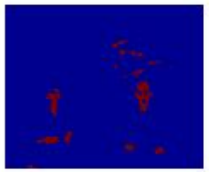


face blob

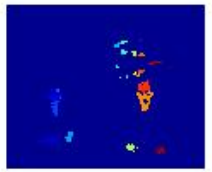
cropped face

raw image

skin pixels



skin blobs



face blob

cropped face

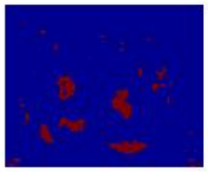
Human Media Interaction

Visualization

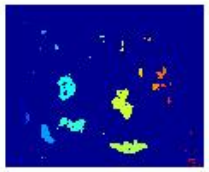
threshold = 1

raw image

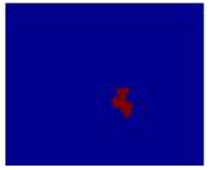
skin pixels



skin blobs



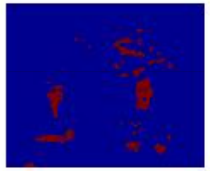
face blob



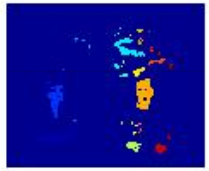
cropped face

raw image

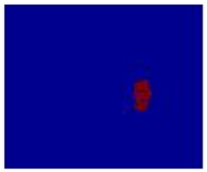
skin pixels



skin blobs



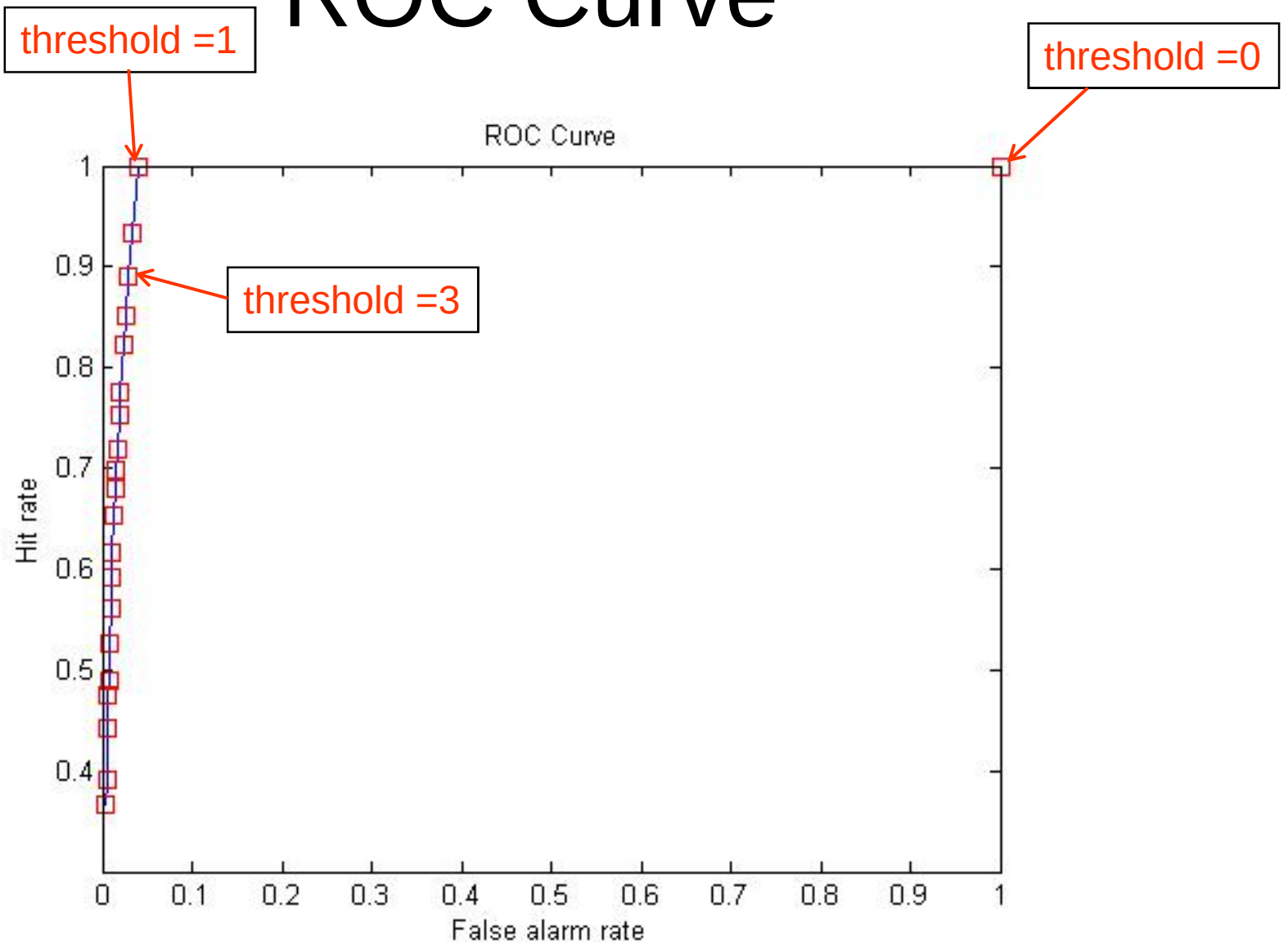
face blob



cropped face

Human Media Interaction

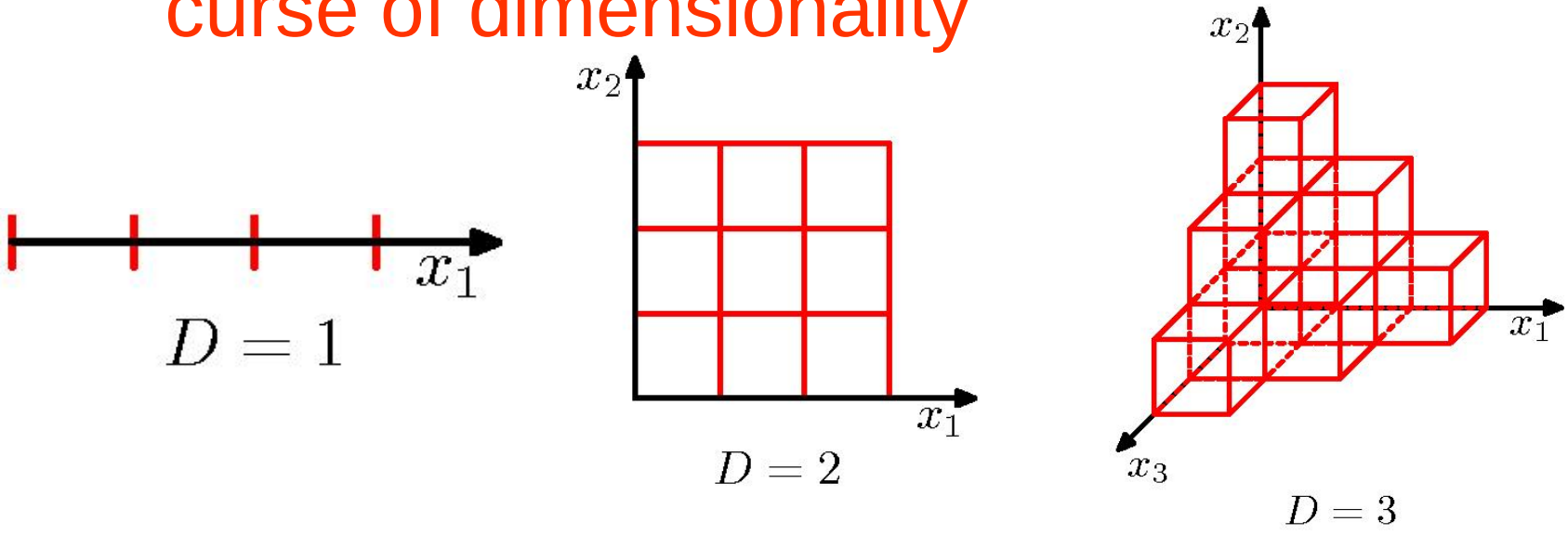
ROC Curve



Human Media Interaction

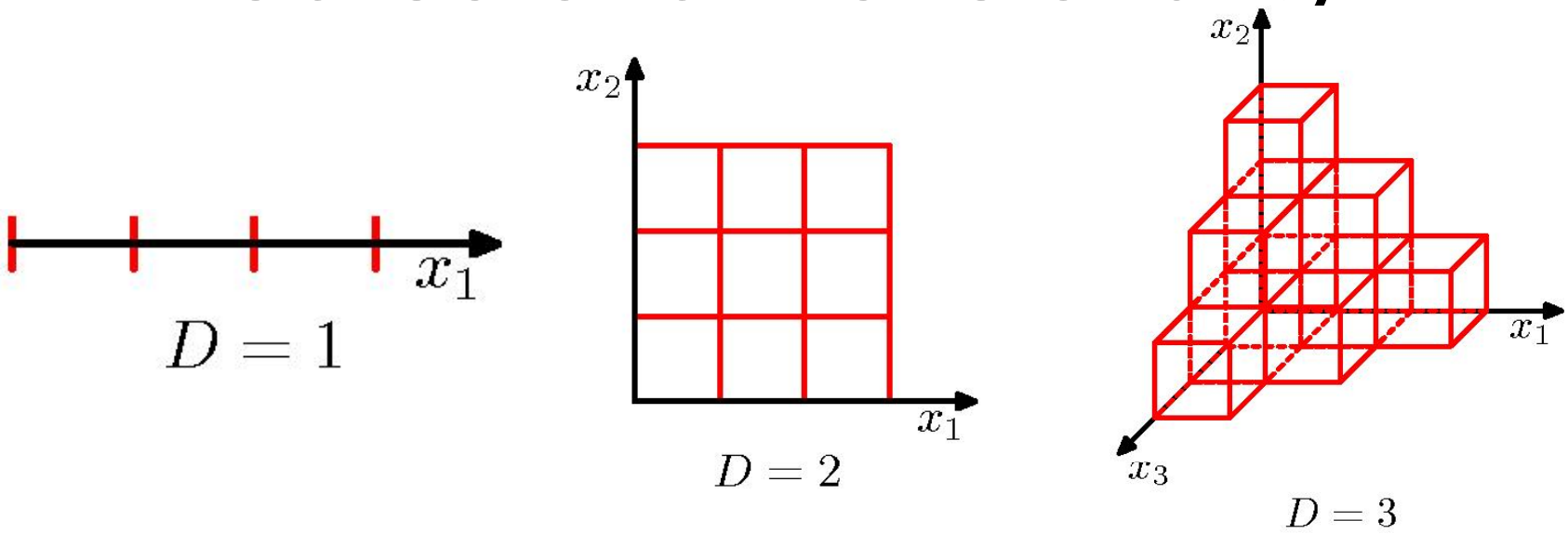
Curse of dimensionality

- In higher dimensions, histogram based approach leads to the so-called "curse of dimensionality"



number of bins grows exponentially with the dimension of the feature space

Curse of dimensionality



- Hence one needs exponentially large quantity of training data. 10 bins per dimension leads to 10^d bins in dimension *d*

Solution: Generate models from the data.

Generative models

- Generate models *M_skin* and *M_non_skin* s.t. *M_skin(x)* (*M_non_skin*) is the probabilility that *x* is *skin* (*non_skin*).
- *x* is classified as skin if *M_skin(x)>M_non_skin(x)*
- Examples:
  - Bayesian approach; Bayesian networks
  - Hidden Markov Models

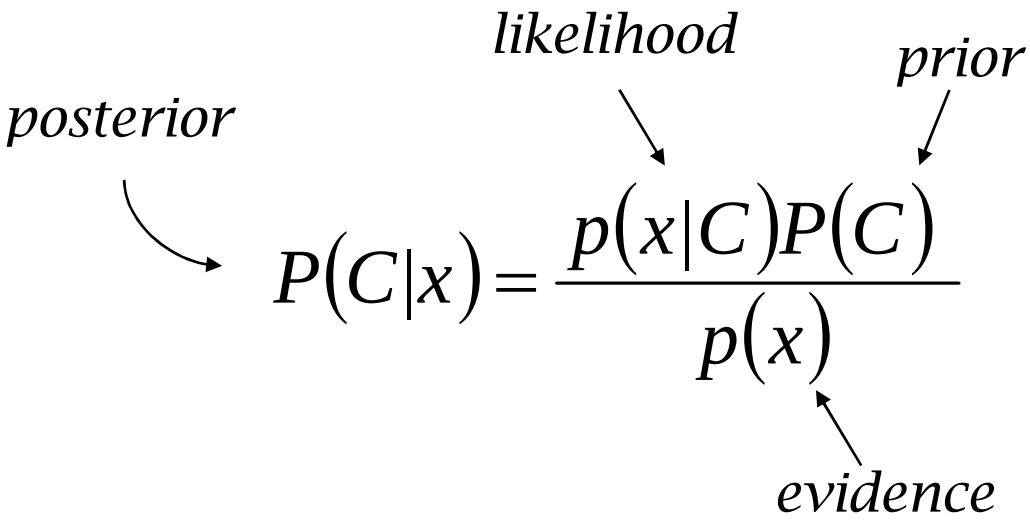
Discriminative models

- Generate real valued function *f_skin* such that *x* is classified as skin iff *f_skin(x)>0.* Also *f_skin(x)* can be seen as a confidence measure.
- Examples:
  - linear discriminant analysis
  - neural networks
  - support vector machines

Discrete Discriminative models

- Generate discrete function *f_skin* with values in *{0,1}* such that *x* is classified as skin iff *f_skin(x)=1*
- Examples:
  - Decision Trees

Bayesian Approach



Classification of *x:*

$max_k P(C_k|x) = max_k\, P(x|C_k)P(C_k)$

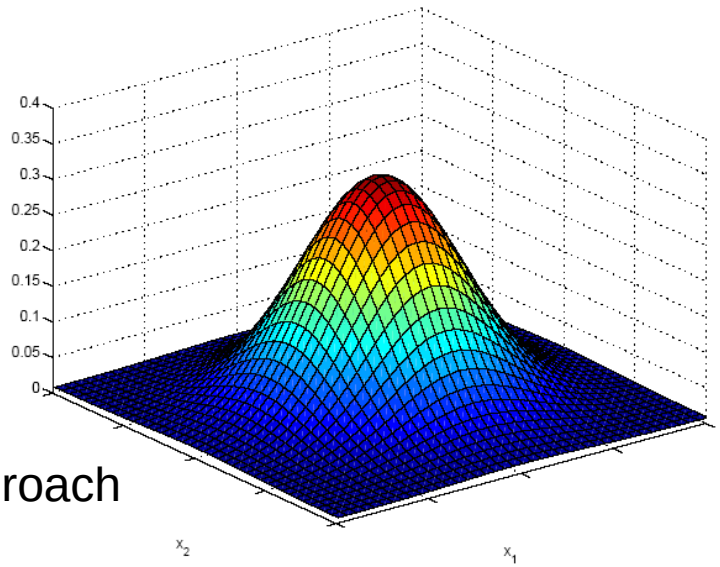
$P(C_k)$ can be estimated from trainings set, hence needs model for $P(x|C_k)$

Bayesian Approach

- Need model for *P(x|C);* the probability that *x* is generated by class *C.*
  - for instance a normal distribution.
  - Estimate mean and covariance

$$\text{Sample mean } m: m = \frac{\sum_{p=1}^{N} x^p}{N}, i = 1,\ldots,d$$

$$\text{Covariance matrix } S: S = \frac{\sum_{p=1}^{N} \left(x^p - m\right)\left(x^p - m\right)^T}{N}$$

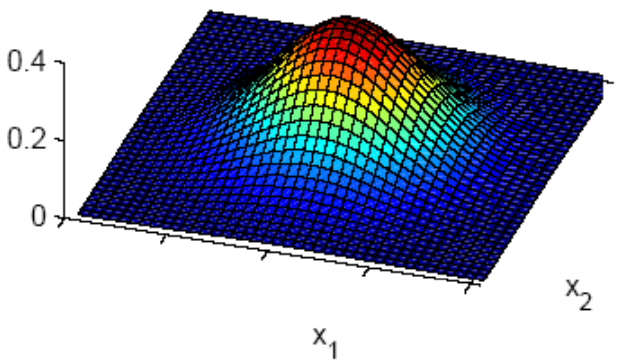


for 2 dim. only 5 parameters instead of 100^2 parameters in the histogram based approach

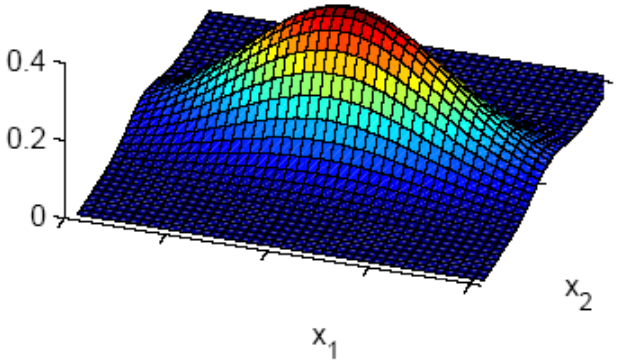
Human Media Interaction

Bayesian Approach

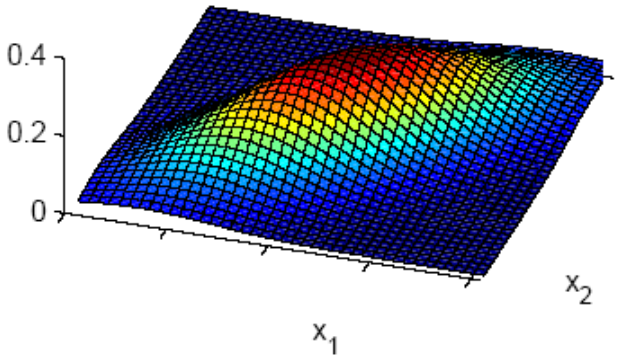
$Cov(x_1,x_2)=0$, $Var(x_1)=Var(x_2)$



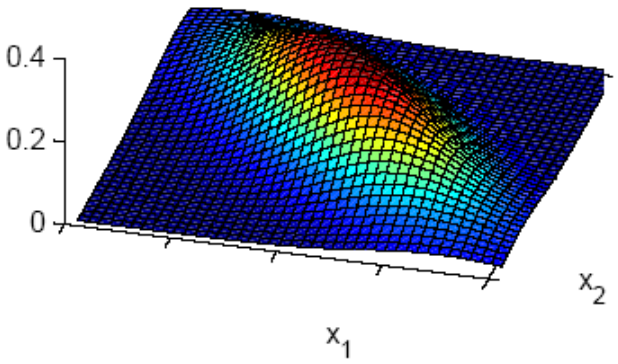
$Cov(x_1,x_2)=0$, $Var(x_1)>Var(x_2)$



$Cov(x_1,x_2)>0$



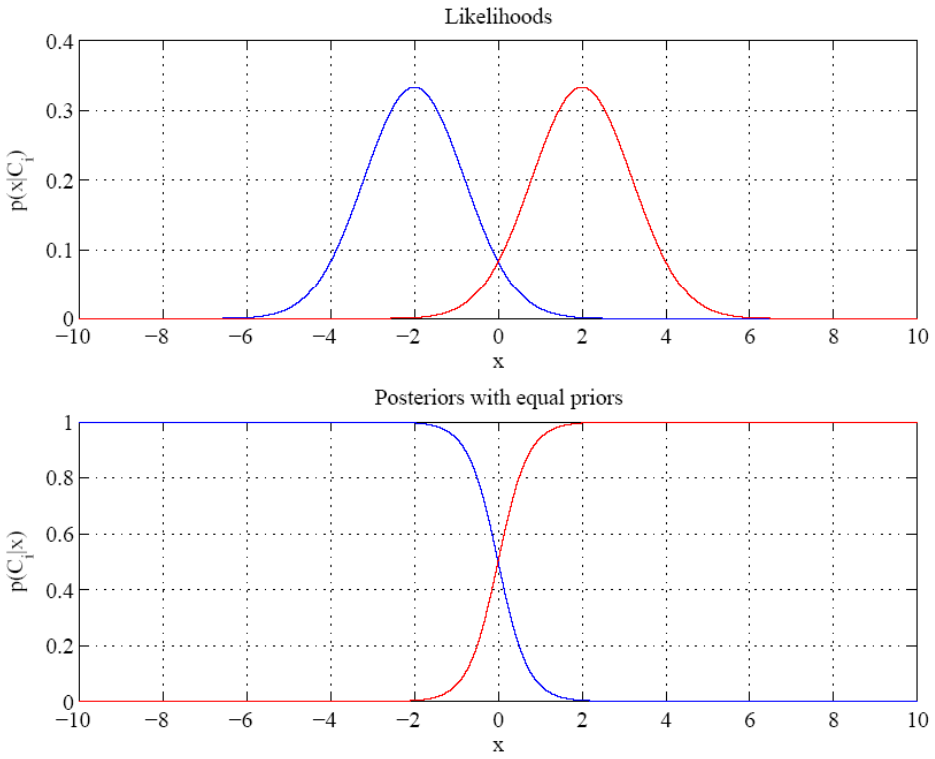
$Cov(x_1,x_2)<0$



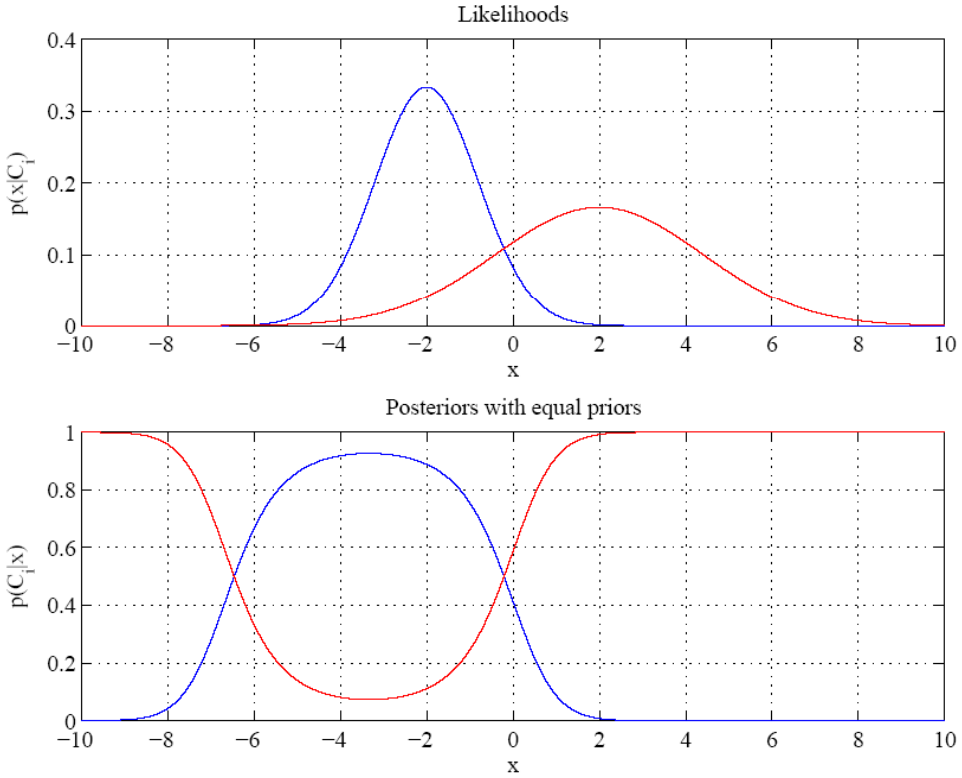
Human Media Interaction

Bayesian Approach

Equal variances



Variances are different

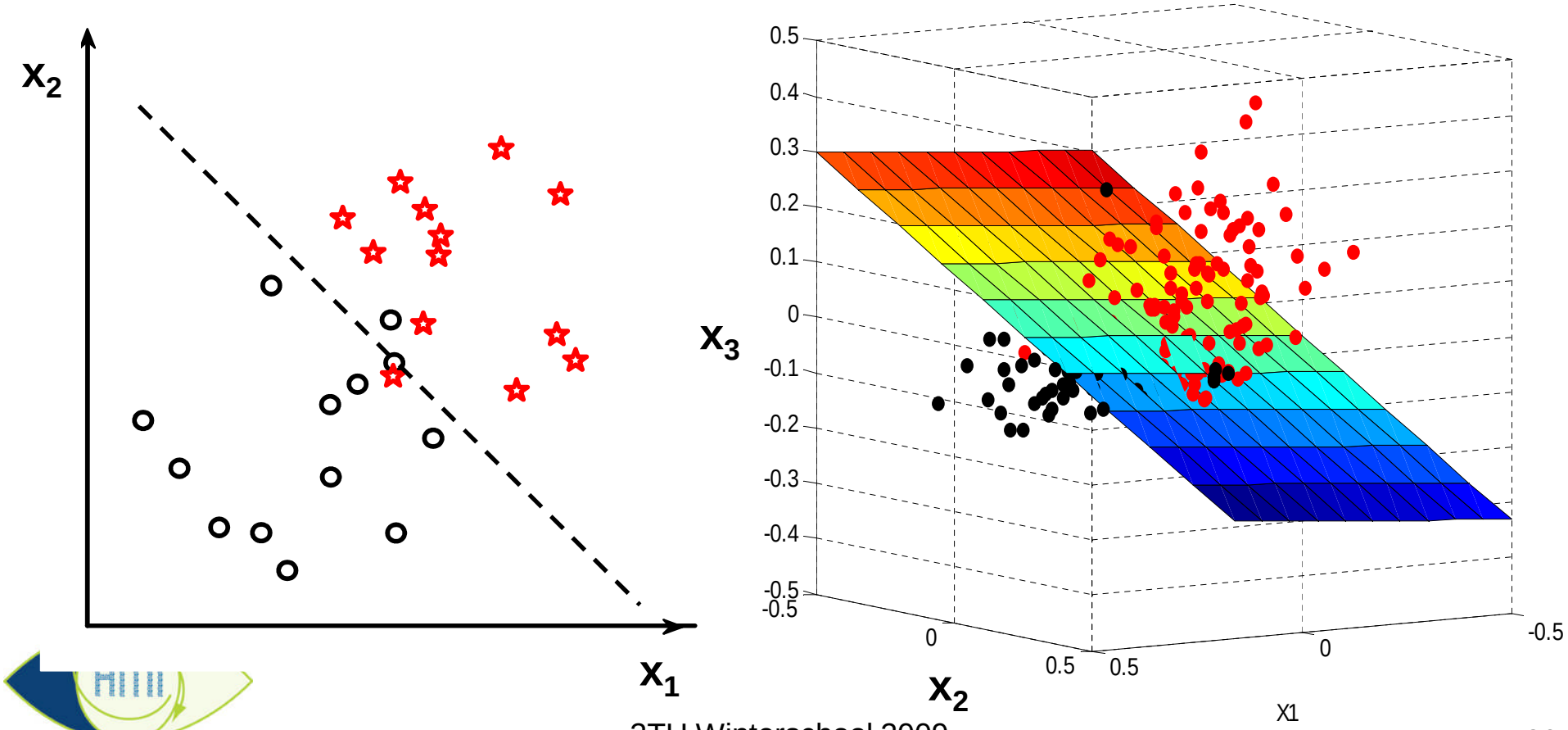


Discriminative: Linear classifiers

- A linear classifier partitions in a *clever* way the feature space into two parts using a hyperplane.
- Classification is base on which side of the hyperplane the feature vector lies.
- In 2 dimensions a linear classifier is determined by a line: y=w.x+b. Hence 3 parameters instead of 5 for the Bayesian approach

Linear classifier

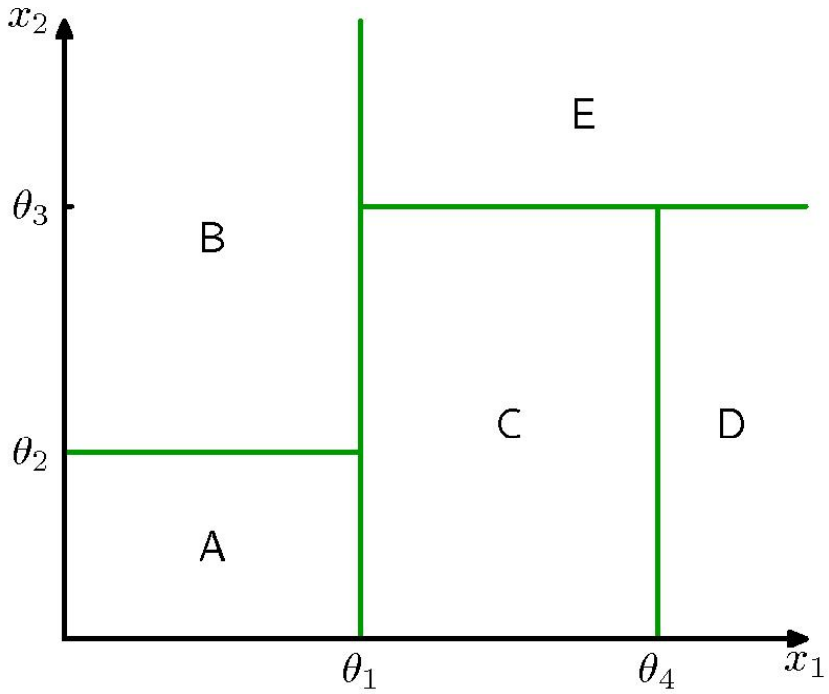
- Simple decision boundary: hyperplane

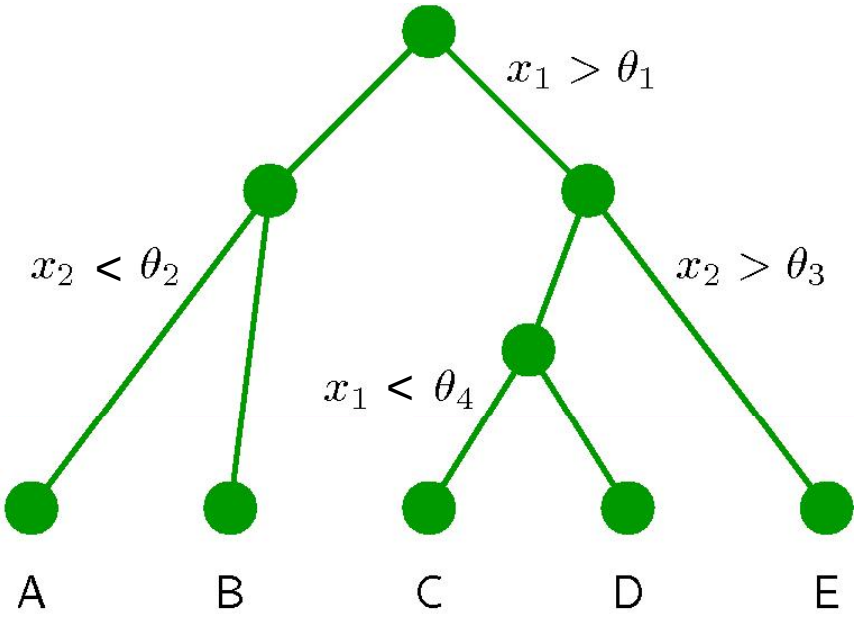


Decision Trees

- A decision trees partitions in a *clever* way the feature space into rectangles and assigns a class to ever rectangle.
  Similar to a non-uniform histogram, but the bins are automatically determined by the learning algorithm.

Decision Trees





The variables to split upon and the thresholds are determined by the learning algorithm.

Neural Networks

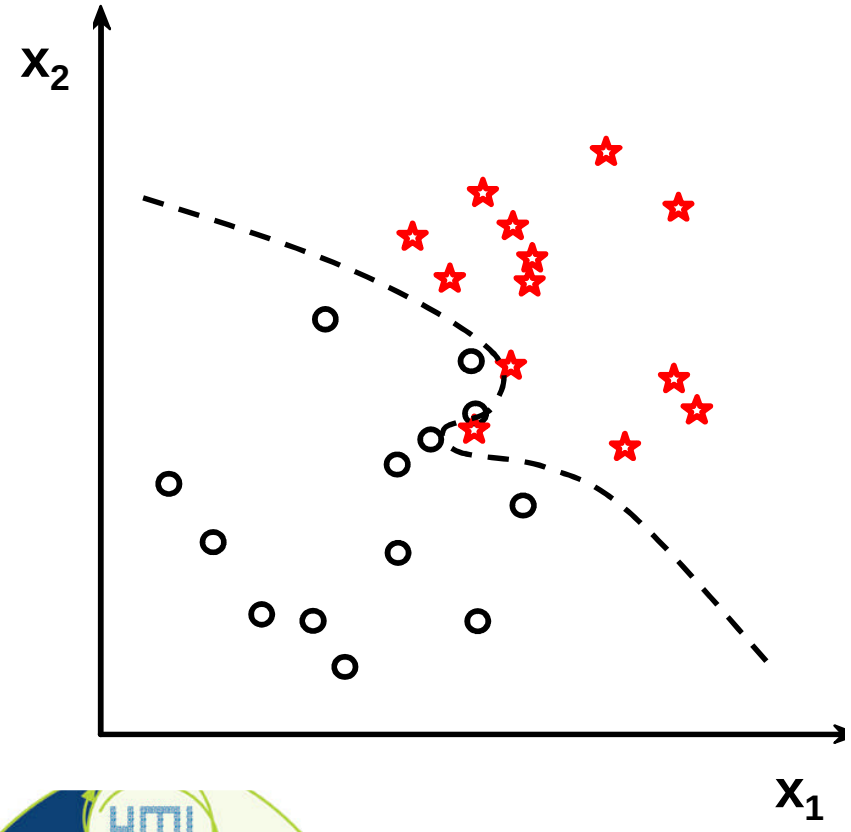
- Structure of simple neurons

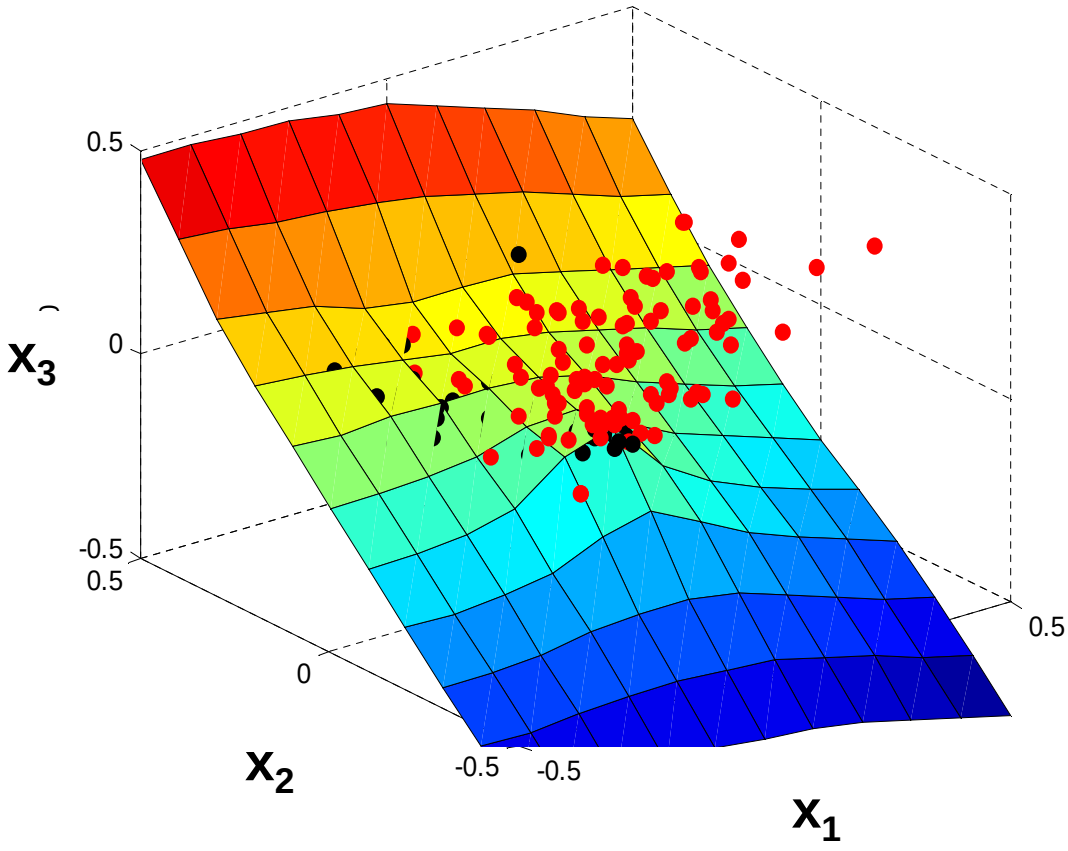
McCulloch and Pitts, 1943

$$f(\mathbf{x}) = \varphi(\mathbf{w} \bullet \mathbf{x} + b)$$

Neural Networks

- More complex decision boundaries

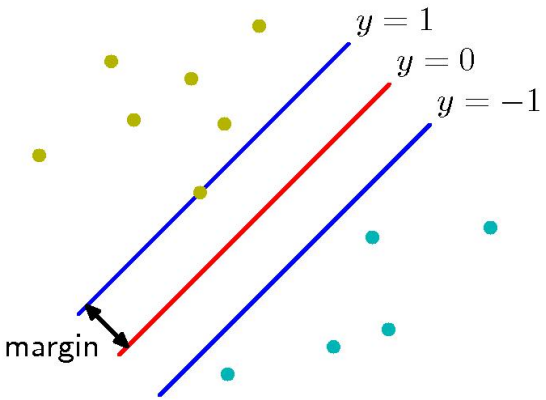


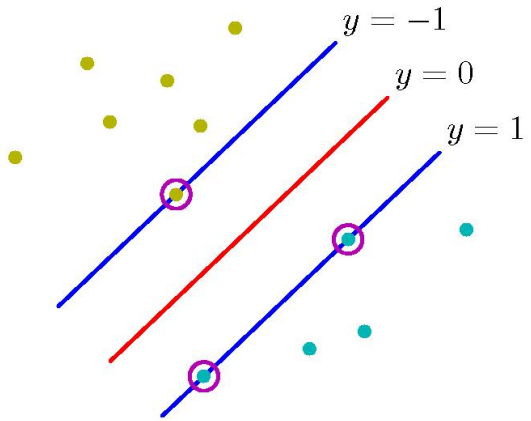


Human Media Interaction

Support Vector Machines a.k.a. Kernel Machines

1. Map feature vectors to a different space, this determines the kernel.
2. Calculate *optimal* linear decision boundary in this transformed space.



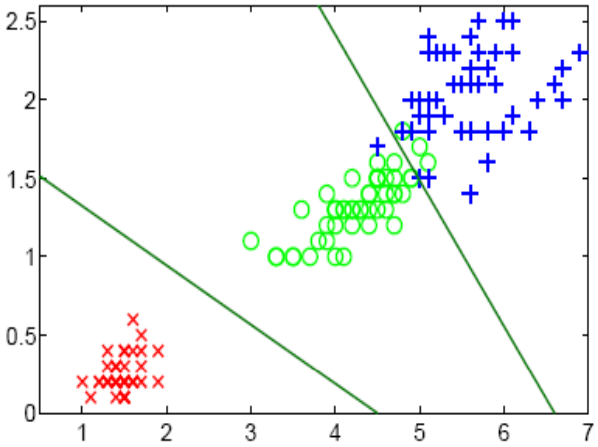


Human Media Interaction

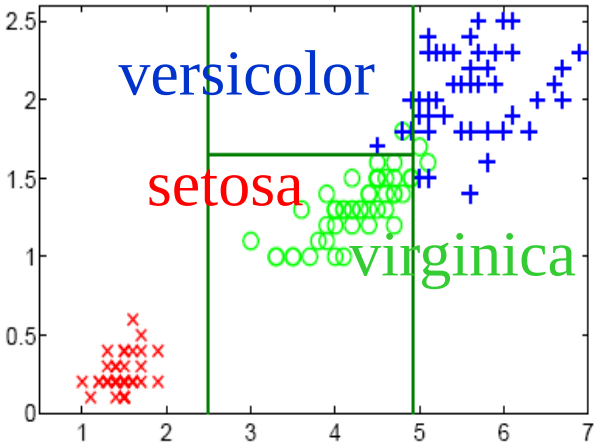
Overview ML Methods

Figure from Norbert Jankowski and Krzysztof Grabczewski

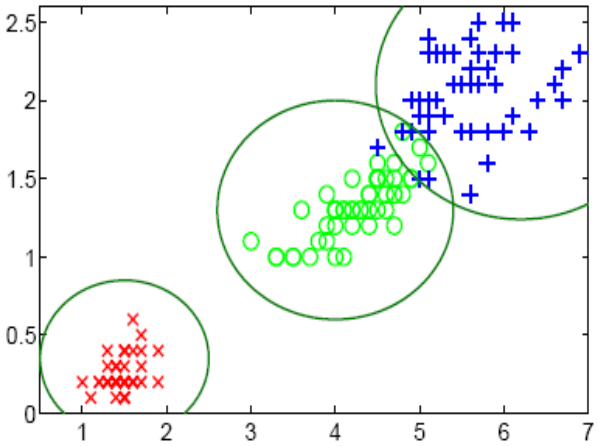
Linear discriminant



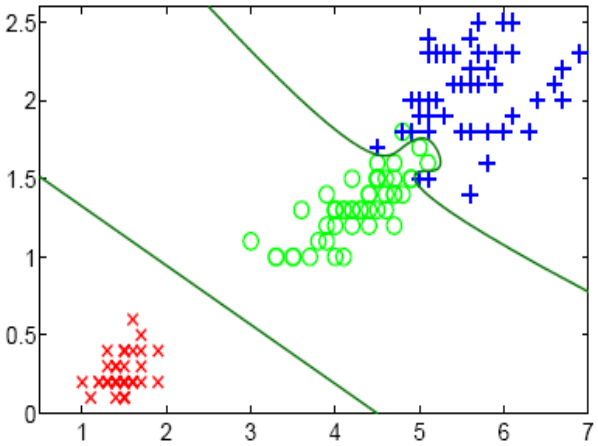
Tree classifier



Gaussian mixture



Kernel method (SVM) and NN



Case 2: Detecting (upper) bodies

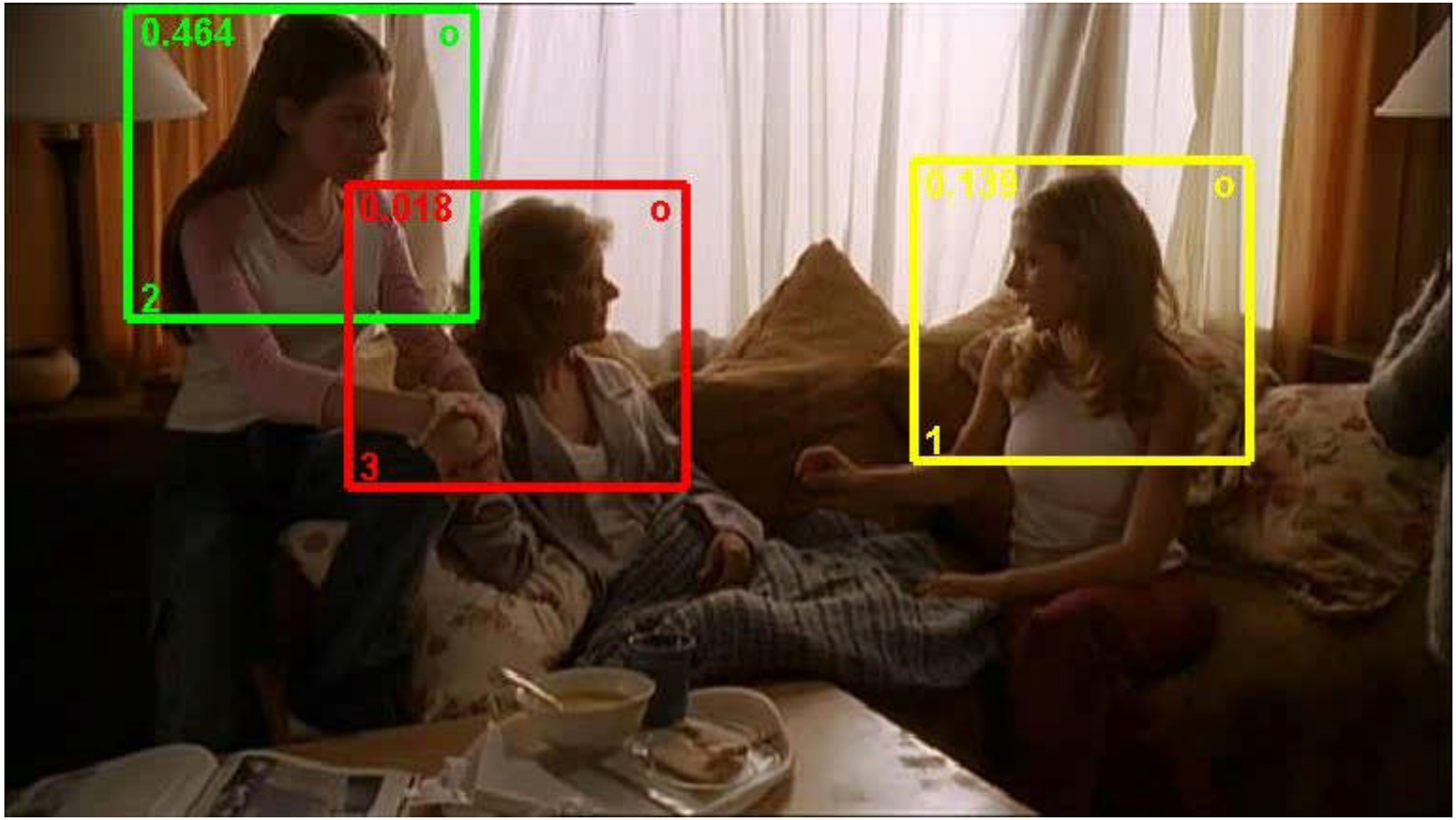
Picasso:
Portrait of Wilhelm Uhde

Edges play an important role

Detecting (upper) bodies

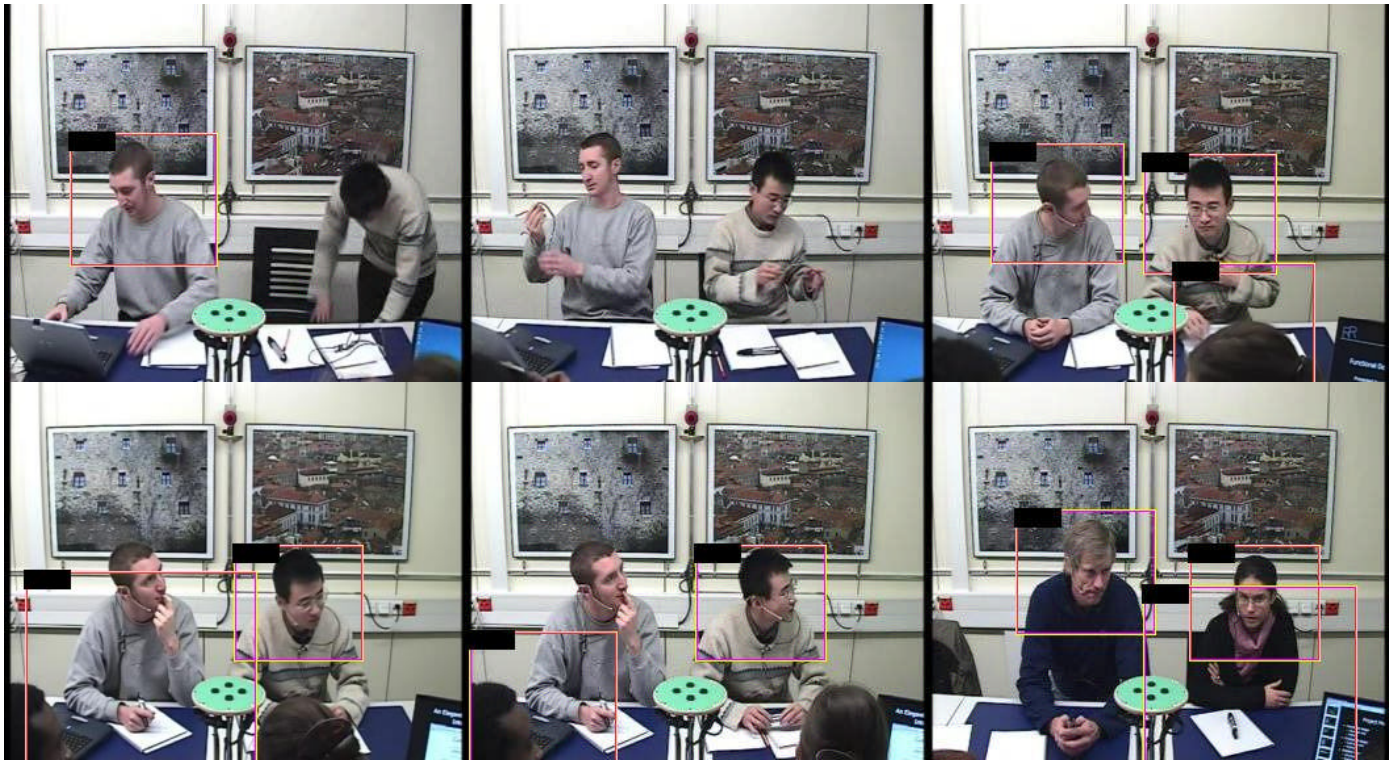
- Approach (based on work of Dalal):
  - sliding window
  - window consists of cells
  - for each cell histogram of oriented gradients (edges); edge information
  - feature vector is the concatenation of the cell histograms
  - confidence based classification using Support Vector Machine

Detecting (upper) bodies: Example



System developed by Ferrari based on the work of Dalal

Detecting (upper) bodies: Applied to meeting setting

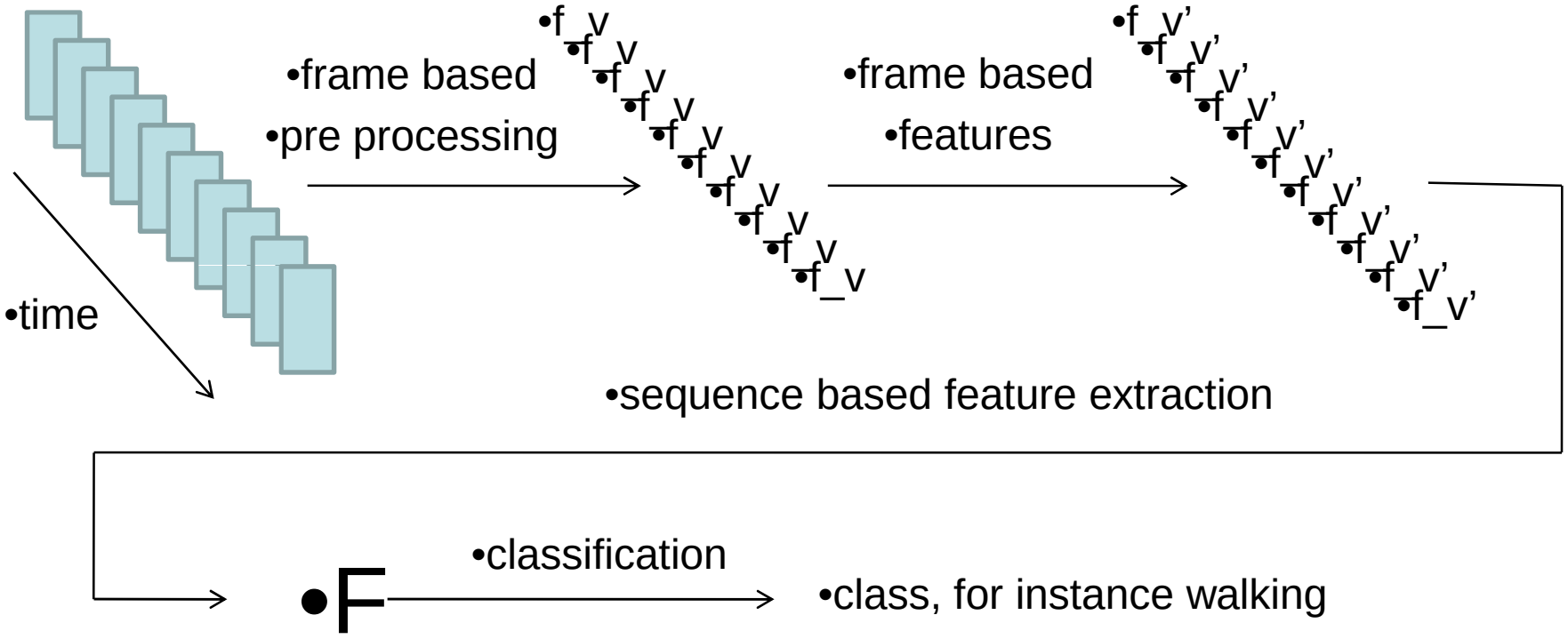


Case 3: Action Recognition

joined work with Ronald Poppe

- Discriminative approach to action recognition
- Main ingredients
  - Histogram of Silhouette Edges
  - Common Spatial Patterns
  - Simple 1 versus 1 mean CSP based classification technique
- Data set: Blank (Weizmann) data set (http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html)

Action Recognition



Weizmann data set

- Video's of 9 persons performing 10 actions: bend, jack, jump, p-jump, run, side, skip, walk, wave1 and wave2.

Some actions are performed from

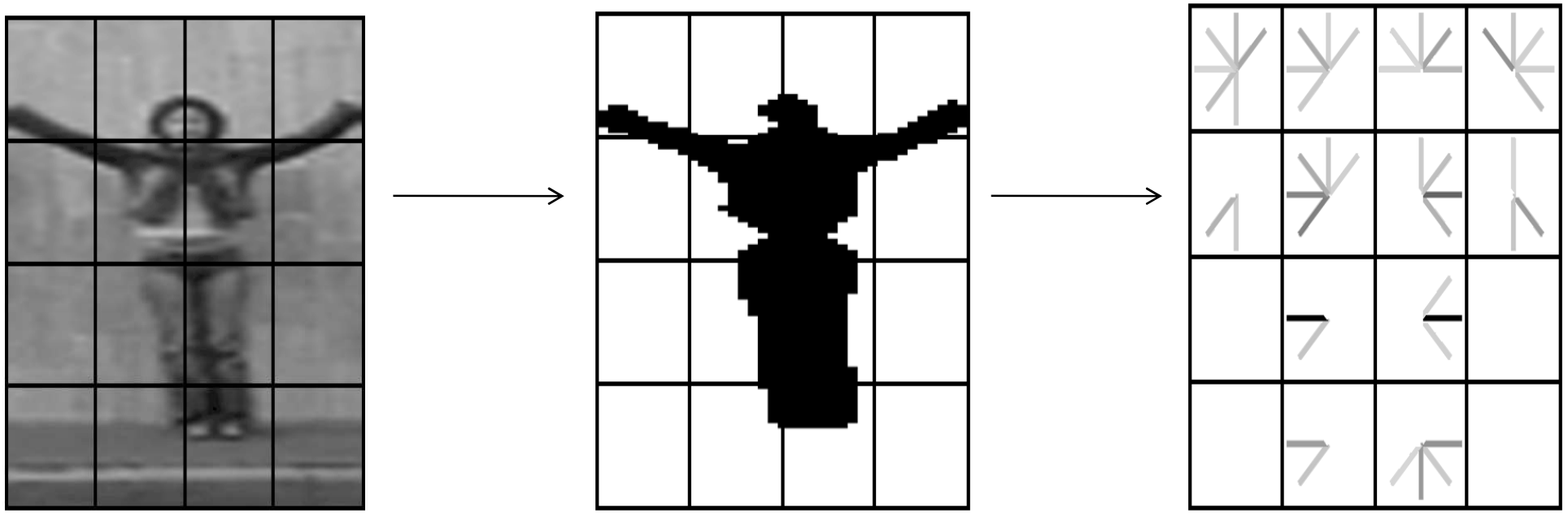
left to right by one person and from right to left by another person

Feature selection process

ROI

Silhouette mask

HoSG



Region of interest is divided in 4x4 cells. Histogram of silhouette gradients divided into 8 bins, each bin covering 45˚.
Result: 128-dimensional descriptor per frame.

Human Media Interaction

Common Spatial Patterns (CSP)

- Frame based feature extraction → sequence of 128-dimensional descriptors.
- How to take into account temporal information?
- CSP is a technique which can be used to discriminate between two classes on bases of the difference in temporal variance.

Idea behind CSP

Training sequences for two actions *a* and *b.* Each sequence can be seen as nxm_p with *n* the number of features and m_p the number of frames. C_a (C_b) is the concatenation of a the training samples of action *a* (*b*).

$$C = C_a C_a^T + C_b C_b^T$$

C the temporal variance over time.

Assumption: Temporal mean is 0!

CSP(2)

C is symmetric : $D = UCU^T$

with D a diagonal matrix. Define $V = \sqrt{D}^{-1}$

then

$$I = (VU)C(VU)^T$$

$$C = C_a C_a^T + C_b C_b^T$$

$$I = \underbrace{(VU)C_a C_a^T (VU)^T} + (VU)C_b C_b^T (VU)^T$$

•But this one is also symmetric, hence can diagonalized → W

CSP(3)

Result:

$$I = (WVU)C_a C_a^T (WVU)^T + (WVU)C_b C_b^T (WVU)^T$$

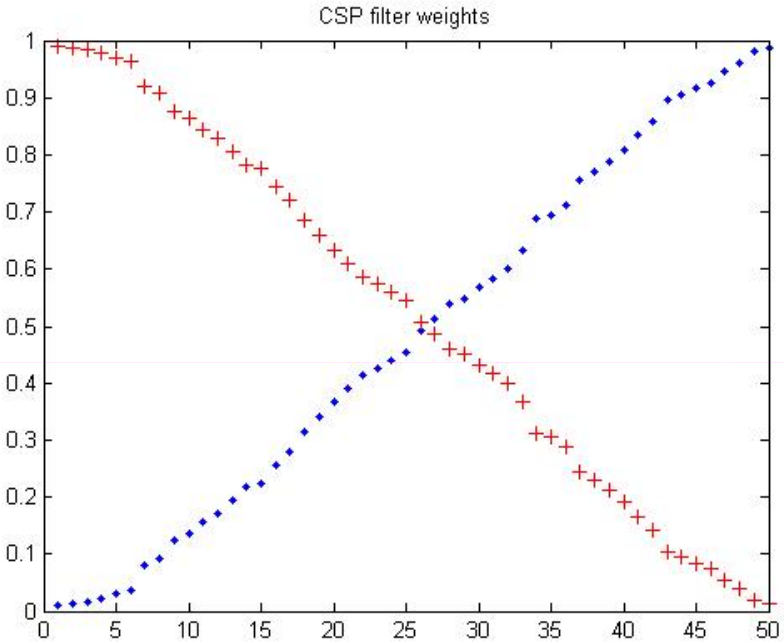
$$\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} v_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & v_n \end{bmatrix} + \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}$$

$v_1,\ldots,v_n$ decreasing, hence $w_1,\ldots,w_n$ increasing.

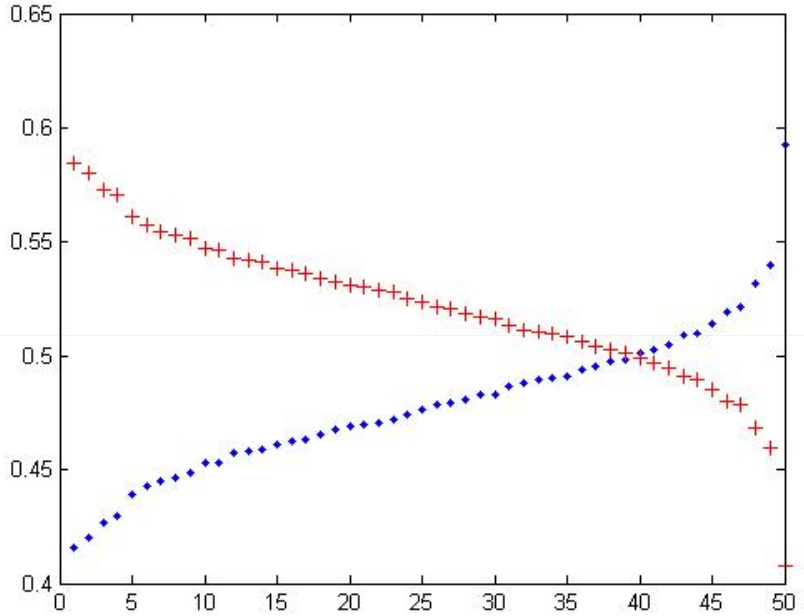
CSP(4)

- *L=WVU* is a linear transformation such that:
  - the sequences of class *a* have high temporal variance in the first components
  - the sequences of class *b* have low temporal variance in the first components
  - the sequences of class *a* have low temporal variance in the last components
  - the sequences of class *b* have high temporal variance in the last components

CSP(5)

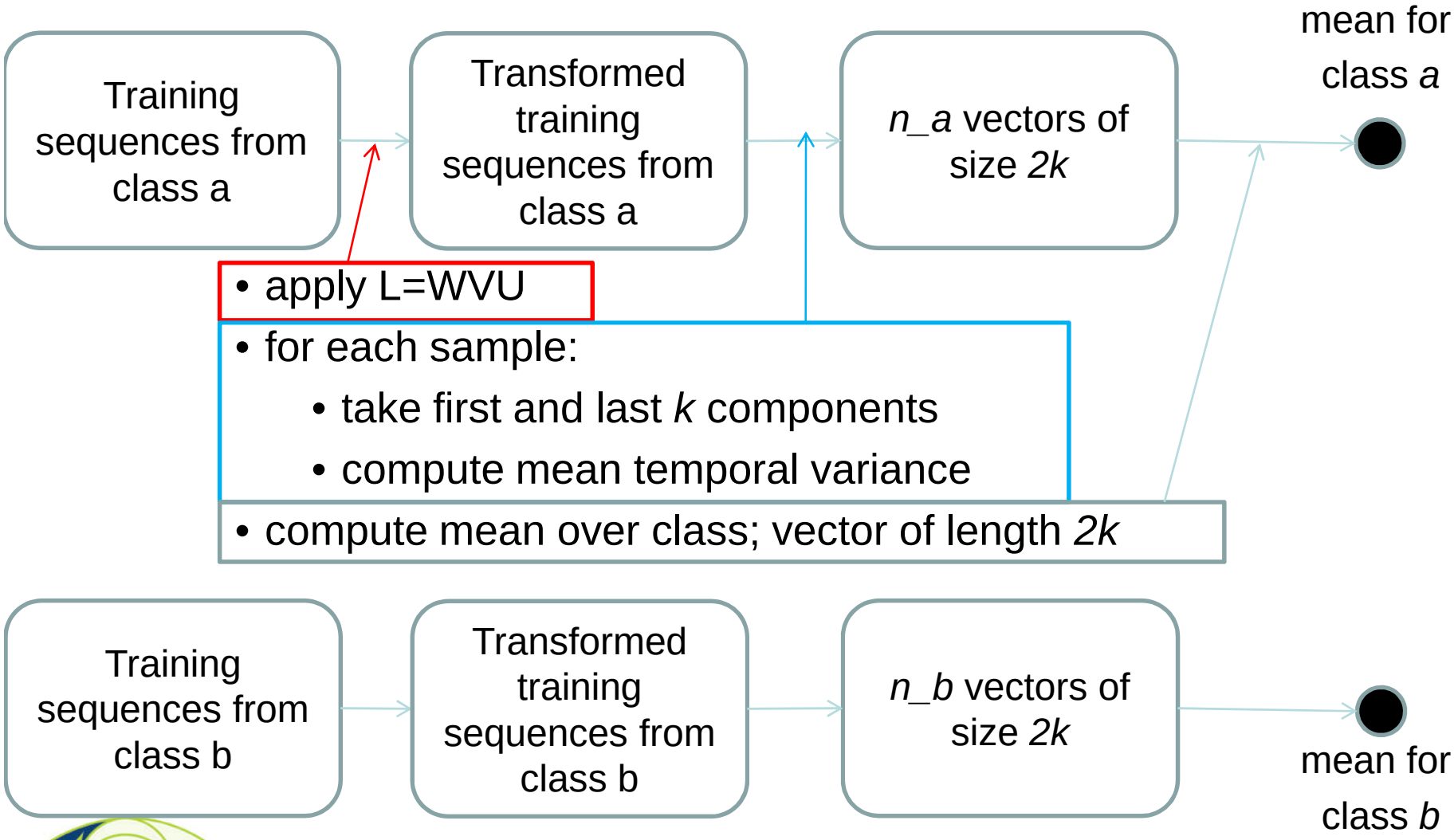


Weizmann data set



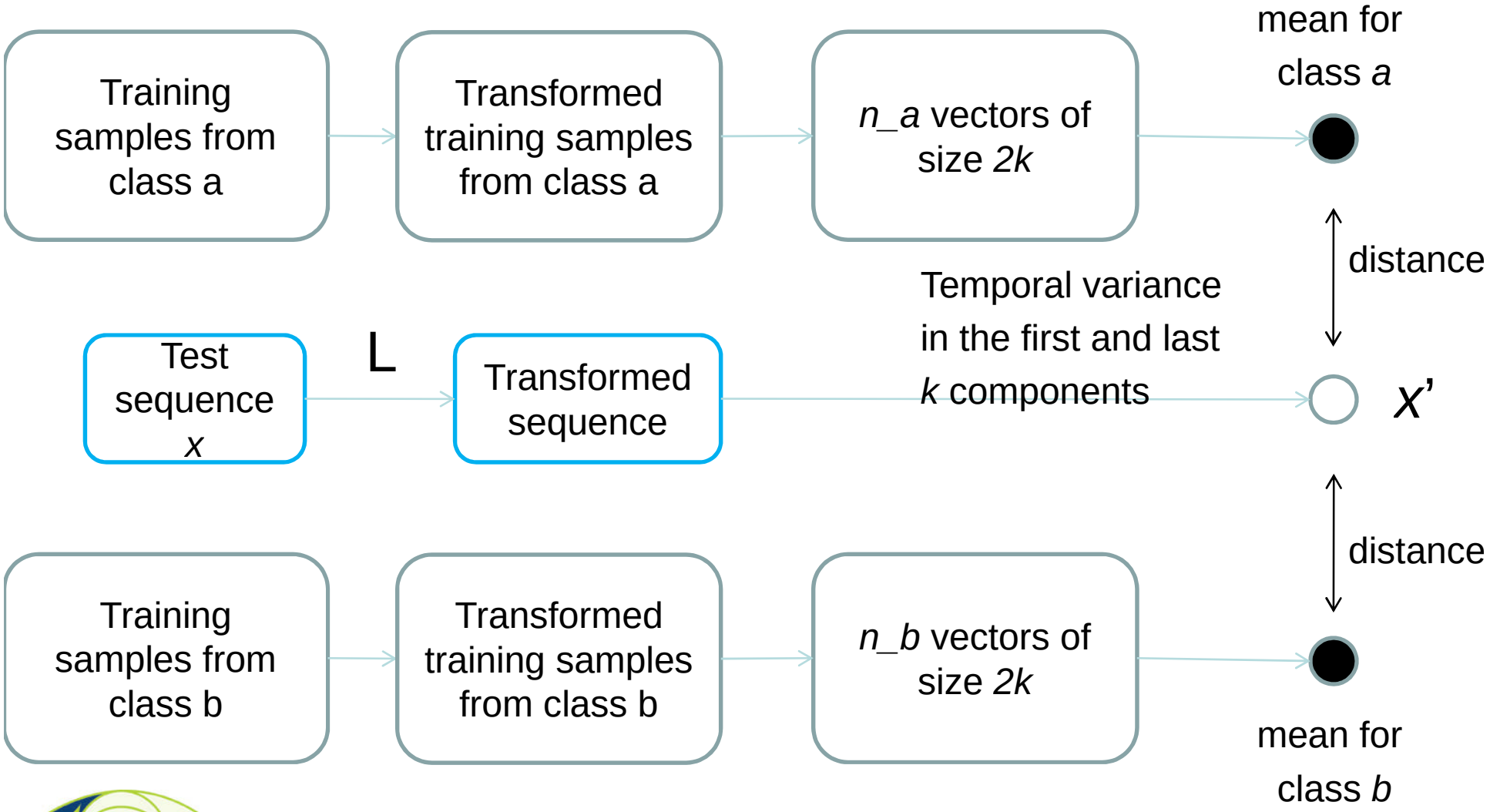
BCI data set

CSP based classification



Human Media Interaction

CSP based classification



CSP based classification (2)

- Discriminating function for class *a* and *b:*

$$g_{a,b}(x) = \frac{\|\bar{b} - x'\| - \|\bar{a} - x'\|}{\|\bar{b} - x'\| + \|\bar{a} - x'\|}$$

- Voting function for class *a:*

$$g_a(x) = \sum_{a \neq b} g_{a,b}(x)$$

Classification of *x:* $Cl(x) = \mathrm{argmax}_a \, g_a(x)$

Evaluation (1)

- *k*=5: this means that every action sequence is reduced to 10-dim. vector
- Leave one person out cross validation
- Performance data set: 95.56%
  - skip of Daria → jumping
  - skip of Ido → running
  - jump of Eli & run of Shahar → walking
  - wave1 of Lyova → wave2

Evaluation (2)

- Leave more persons out cross validation

Nr. training persons	Performance
1	64.72%
2	77.82%
3	81.83%
4	84.60%
5	86.63%
6	89.01%
7	91.39%
8	95.56%

Evaluation

- Divide data set in training, validation and test.
- Models are trained on train set.
- Hyper parameters are determined on validation set.
- Chose best model, including hyper parameters based on performance on validation set.
- Asses performance on test set. This gives the best indication of the performance in “real life”!
- It is not allowed that any information from the test set leaks into the training.

Evaluation methods

- Precision and recall
- Error; percentage of misclassified samples in the test set
- Confusion matrix
- ROC Curve

Confusion Matrix

	Predicted class	
True class	Yes	No
Yes	TP True positive	FN False negative
No	FP False positive	TN True negative

Many valuable statistics, such as error, precision and recall, can be derived from the confusion matrix

Concluding remarks

- For human behavior analysis in a naturalistic setting;
  - annotated data is of utmost importance but hard to get.
  - Uncontrolled environments and context
- In ML there is no “Free Lunch”. Best ML approach is determined by the problem under consideration.

More information

- Video lectures:
  - Introduction to Machine Learning
  - Machine Learning in Vision
- Books:
  - E. Alpaydin: Introduction to Machine Learning
  - C.M. Bishop: Pattern Recognition and Machine Learning
- Journals:
  - Machine Learning
  - Journal of Machine Learning


ZIJN ER NOG VRAGEN?

Case 4:
Visual Focuses of Attention

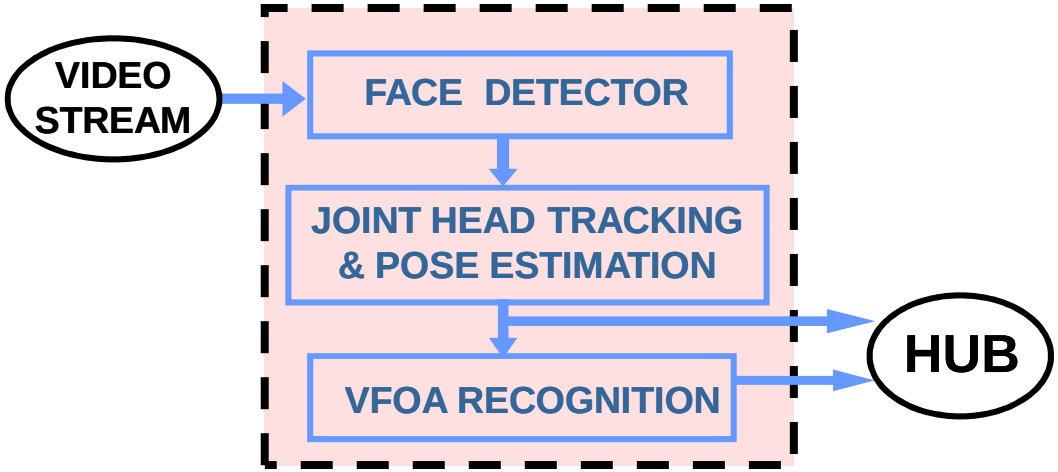
Challenge

Can we build a model which determines the head of a person and where the person is looking at.

VFoA recognition system

work of Elisa Ricci @ Idiap

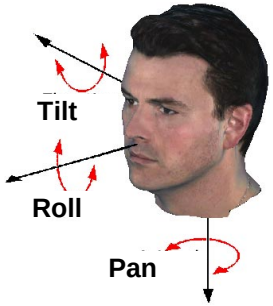
- VFoA (*who is looking at whom or what*) is defined by eye gaze
- Eye gaze is difficult to infer: head pose is used as an indication of VFoA
- IDIAP system for VFoA recognition
  - Fully automatic
  - Running close to realtime
  - Handling faces at low/mid resolution



Human Media Interaction

Head tracking and pose estimation

- Goal: provide as outputs the location of the head and the head pose (pan/tilt/roll angles).
- *Joint* head tracking and pose estimation in a Bayesian framework
  - *Mixed state particle filter:* the state X contains both continuous and discrete variables.
  - Automatic *initialization* with a multiview face detector.
  - *Exemplar based tracking*: a reference models for each possible pose.
  - Reference models *learned* offline with a *discriminative approach* from the PRIMA-POINTING database.

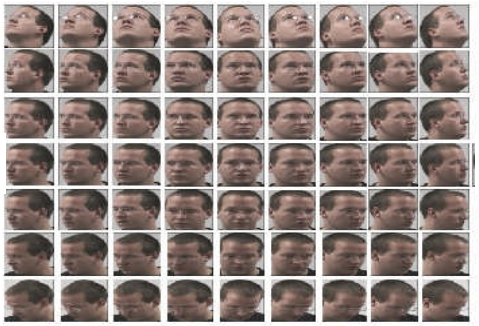


$$X = (S, \theta, r)$$

S: head location and size

θ : out-plane rotations

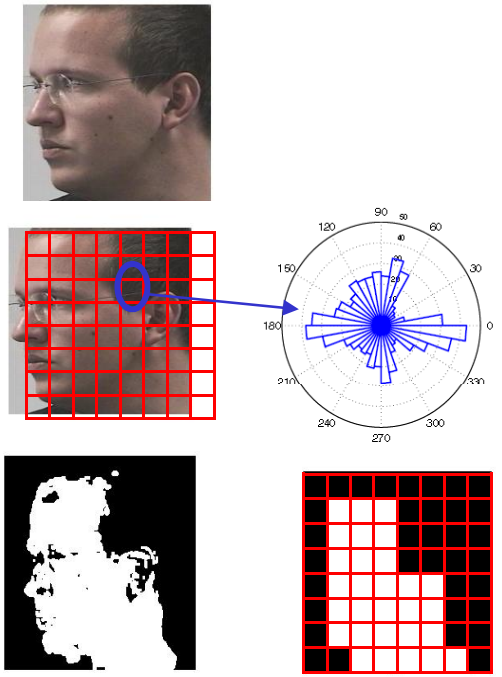
r: in-plane rotations



Human Media Interaction

Head tracking and pose estimation

- Observation model
  - $Y = (Y^{text}, Y^{col})$: independent features for textures and color
  - Features are computed with integral images
  - Y^{text}: concatention of Edge Orientation Histograms (EOH)
  - Y^{col}: binary masks extracted by a gaussian skin color model learned offline and adapted online
- Likelihood function



EOH

Skin color mask

$$p(Y \mid X = (S,\theta,r)) = p_{text}(Y^{text}(S) \mid \theta,r)\, p_{col}(Y^{col} \mid \theta,r)$$

$$= \underbrace{e^{-\lambda_{text} D_{text}(Y^{text}(S_i), R^{text}(\theta,r))}}_{\text{Texture likelihood}} \underbrace{e^{-\lambda_{col} D_{col}(Y^{col}(S_i), R^{col}(\theta,r))}}_{\text{Color likelihood}}$$

Human Media Interaction

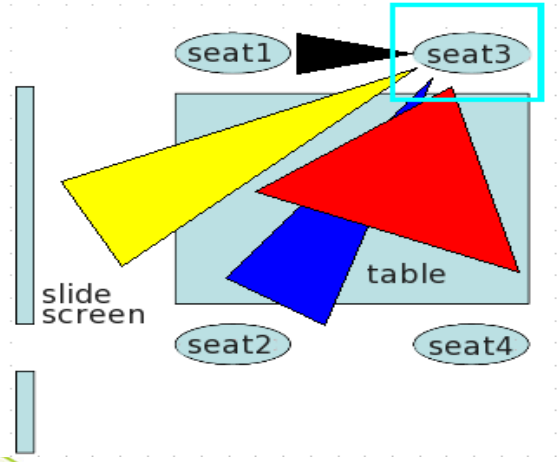
VFoA recognition

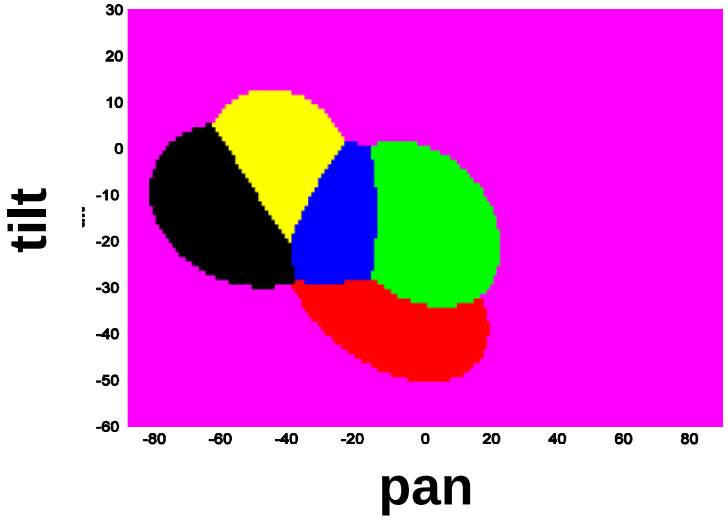
- Gaussian distribution models to estimate for a given person the VFOA f^t given the head pose θ^t (using the mean of the particle distribution)
  - Gaussian distribution for regular target

$$p(\theta^t \mid f^t = T_k) = N\left(\theta^t; \mu_k, \Sigma_k\right)$$

  - Uniform distribution for unfocused

$$p(\theta^t \mid f^t = unfocused) = u$$





VFoA recognition

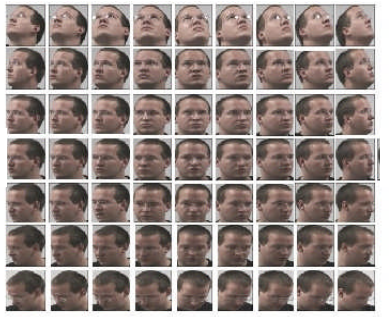
- Representation of the head pose distribution only by its mean value implies a loss of information
- Exponential distribution models

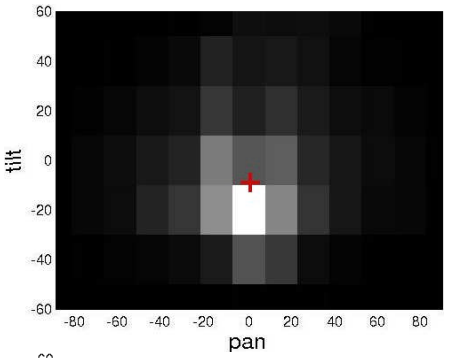
$$p(\theta^t \mid f^t = T_k) = \lambda e^{-\lambda \rho\left(\pi_k^t(\theta), \Pi_k^t\right)}$$

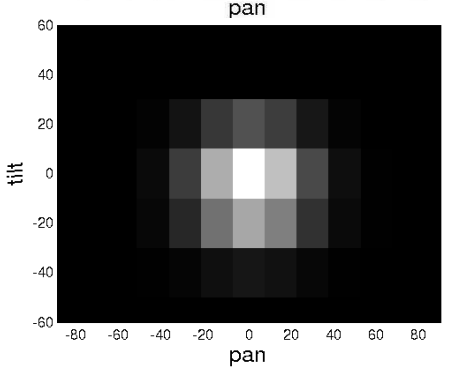
$\rho\left(\pi_k^t(\theta), \Pi_k^t\right)$ distance between distributions

$$\Pi_k^t(\theta = l) = \frac{N\left(l\ ; \mu_k, \Sigma_k^t\right)}{\sum_{l'} N\left(l'; \mu_k, \Sigma_k^t\right)}$$

pdf modeling focusing at target k



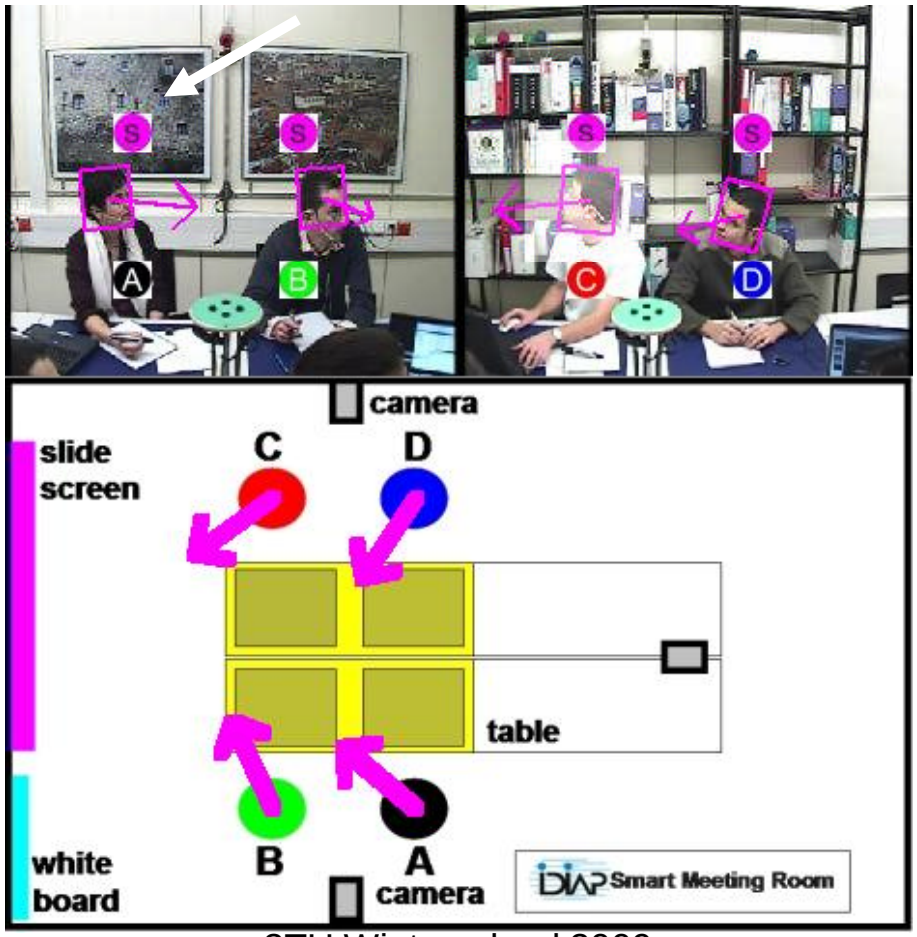




Human Media Interaction

VFoA recognition

recognized VFoA (e.g. looks at slide)



My Research Interests

- Applied Machine Learning:
  - Vision based Human Behavior Computing:
    - Multi Modal Laughter Recognition (MLMI 2008)
    - Pose Estimation (AMDO 2008)
    - Action Recognition (FG 2008)
  - Brain Computer Interfacing
    - BCI as one of the Interaction Modalities for Games
    - Challenge: dealing with artifacts

Content

- Introduction to Machine Learning (ML)
- Some cases:
  - Head detection
  - Upper Body Detection
  - Action Recognition
  - Visual Focus of Attention
- Evaluation
- Some concluding remarks