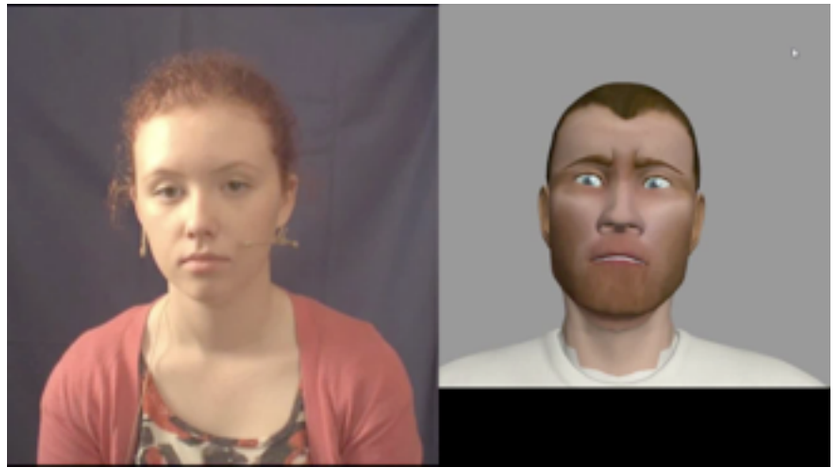
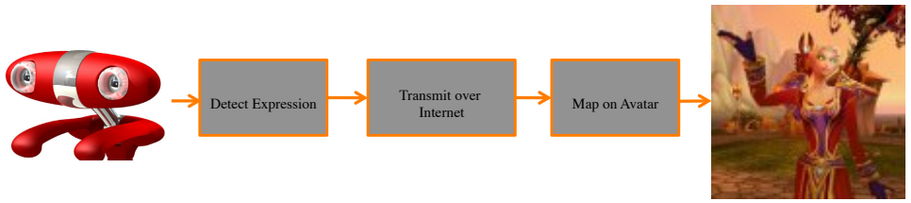
Automatic Facial Expression Recognition

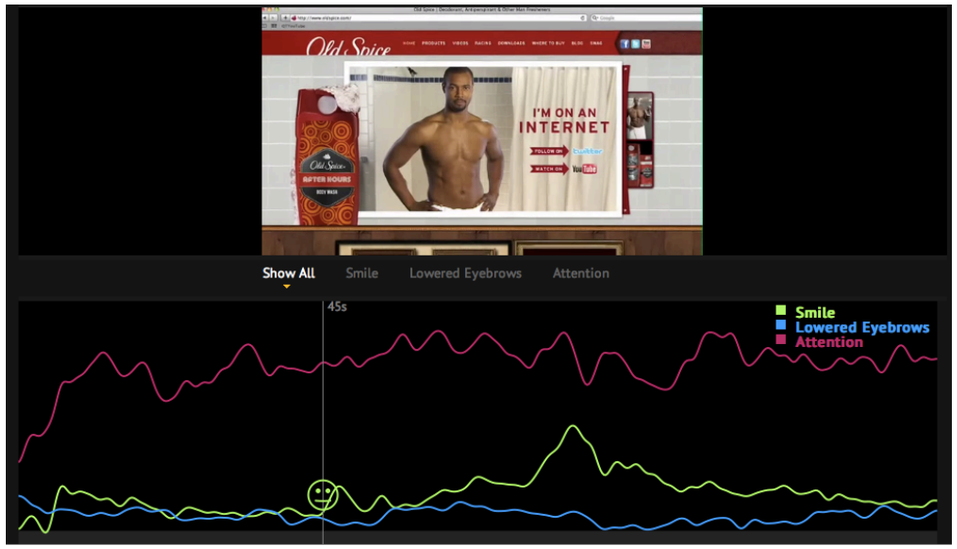
A practical tutorial

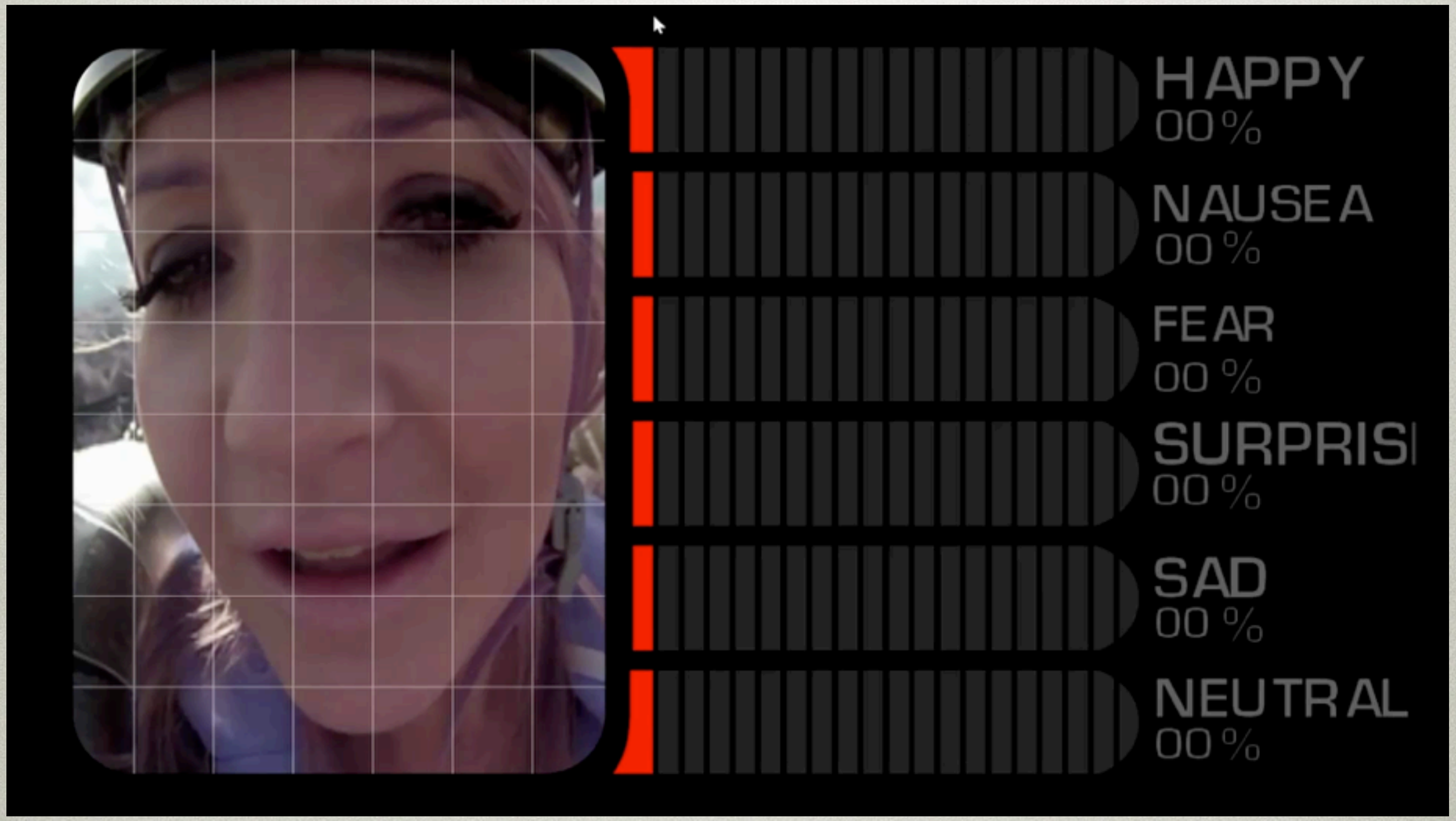
by Michel Valstar

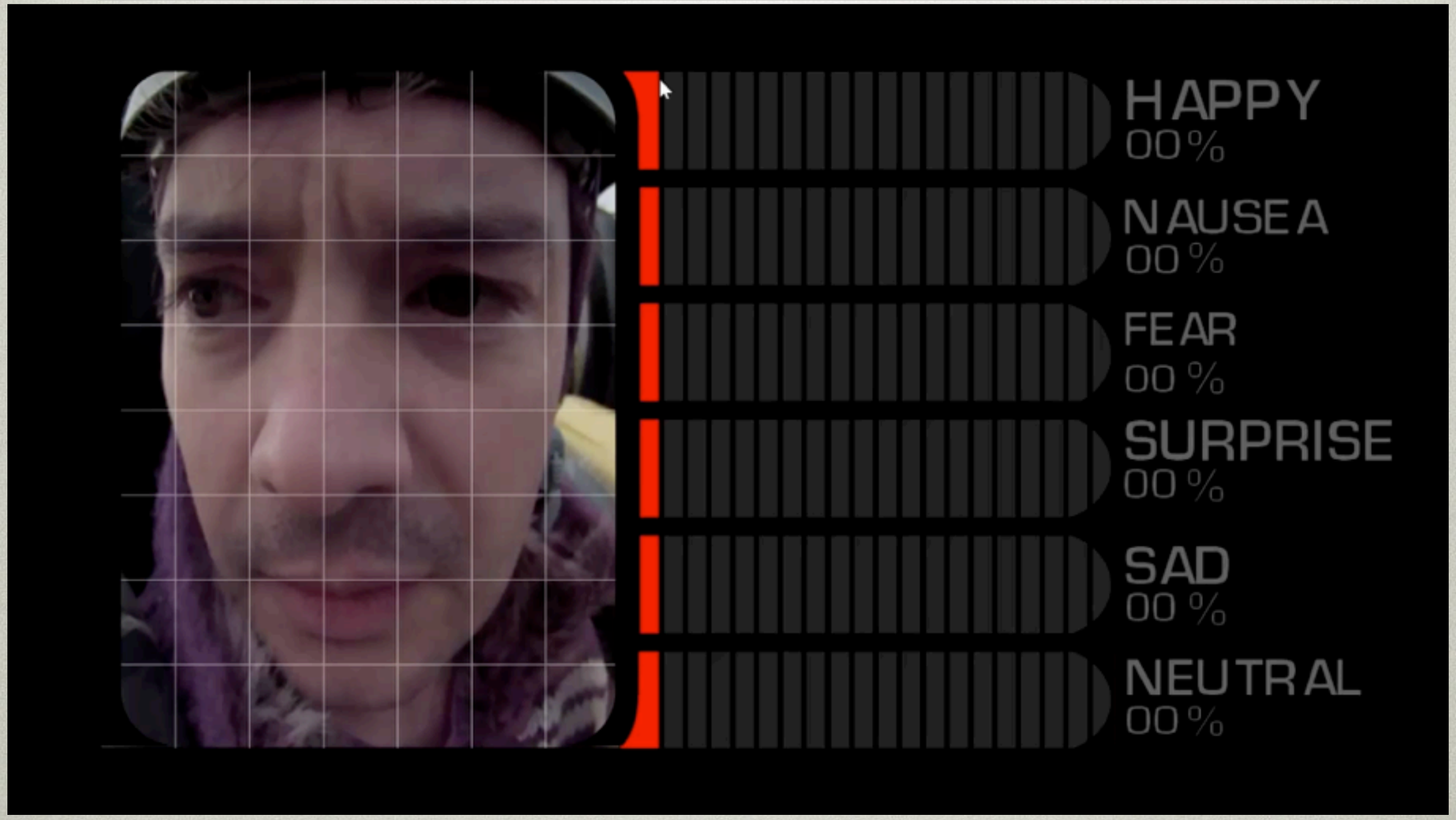
Opportunities





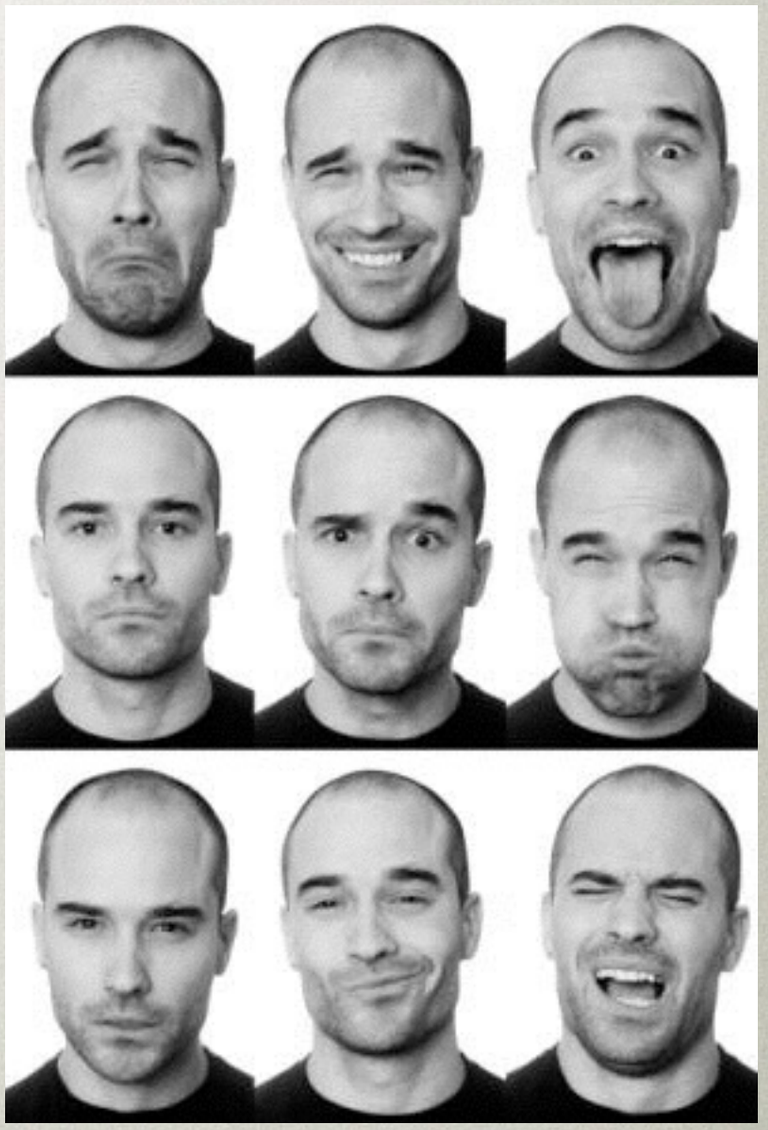



HAPPY
00%
NAUSEA
00%
FEAR
00%
SURPRISE
00%
SAD
00%
NEUTRAL
00%


HAPPY
00%
NAUSEA
00%
FEAR
00%
SURPRISE
00%
SAD
00%
NEUTRAL
00%

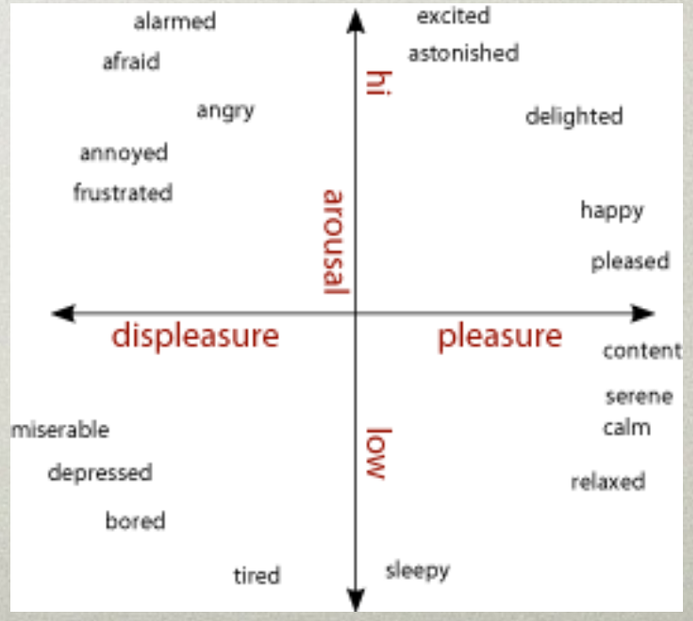
What are Expressions?

- Facial Displays are faces with a different appearance than neutral
- Facial Expressions are facial displays that express emotion
- Facial Displays can also express other messages, e.g. social signals

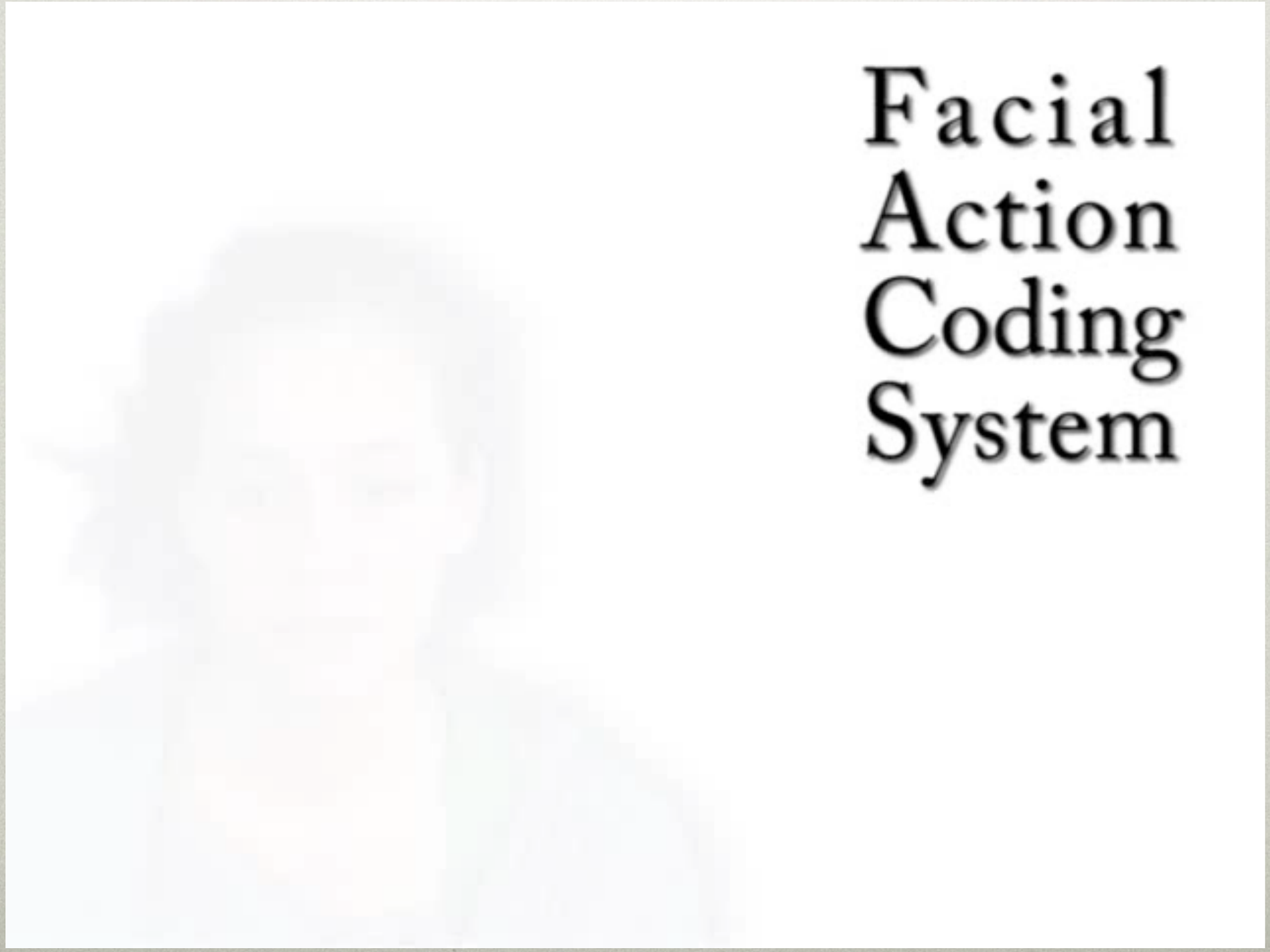


Emotion Theory

- There are a number of different theories on how humans perceive and feel emotions.
- The most commonly used models are the six basic emotions, and Russell's Valence/Arousal/Dominance (VAD) model.
- Both are message judgment models and inherently subjective.



FACS



Expression Recognition Tasks

- Discrete Expressions of Emotion (e.g. 6 basic emotions)
- FACS Action Units: Facial muscle actions
- Dimensional Affect (Valence/Arousal/Dominance)
- Higher-order behaviour (e.g. depression, engagement)

State of the Art

FERA 2013

- AU Sub-Challenge: Frame by frame detection of 12 Action Units

  87 Training videos

  71 Test videos

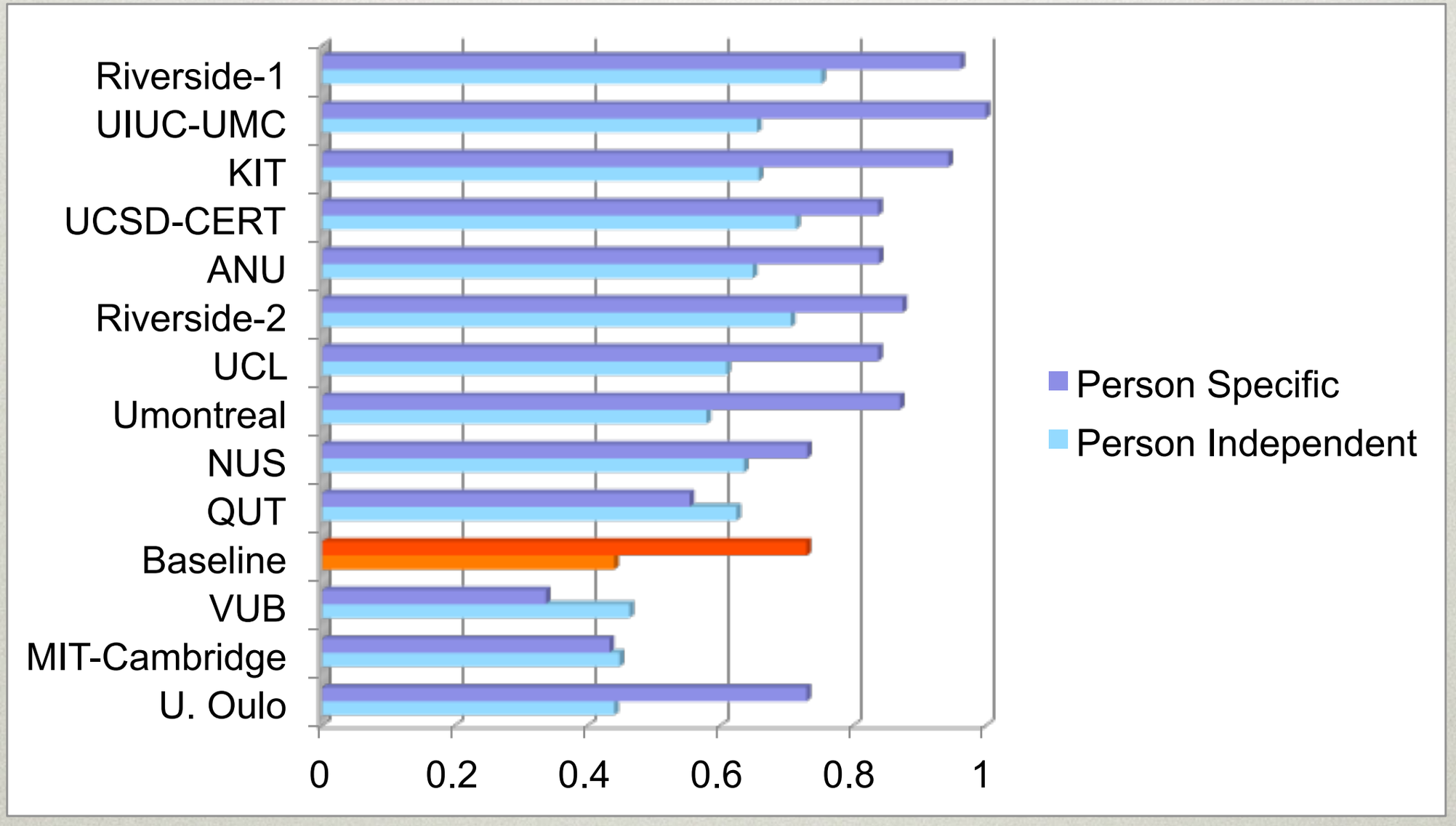
- Emotion Sub-Challenge: Event detection of 5 discrete emotions (Anger, Fear, Joy, Relief, Sadness)

  155 Training videos

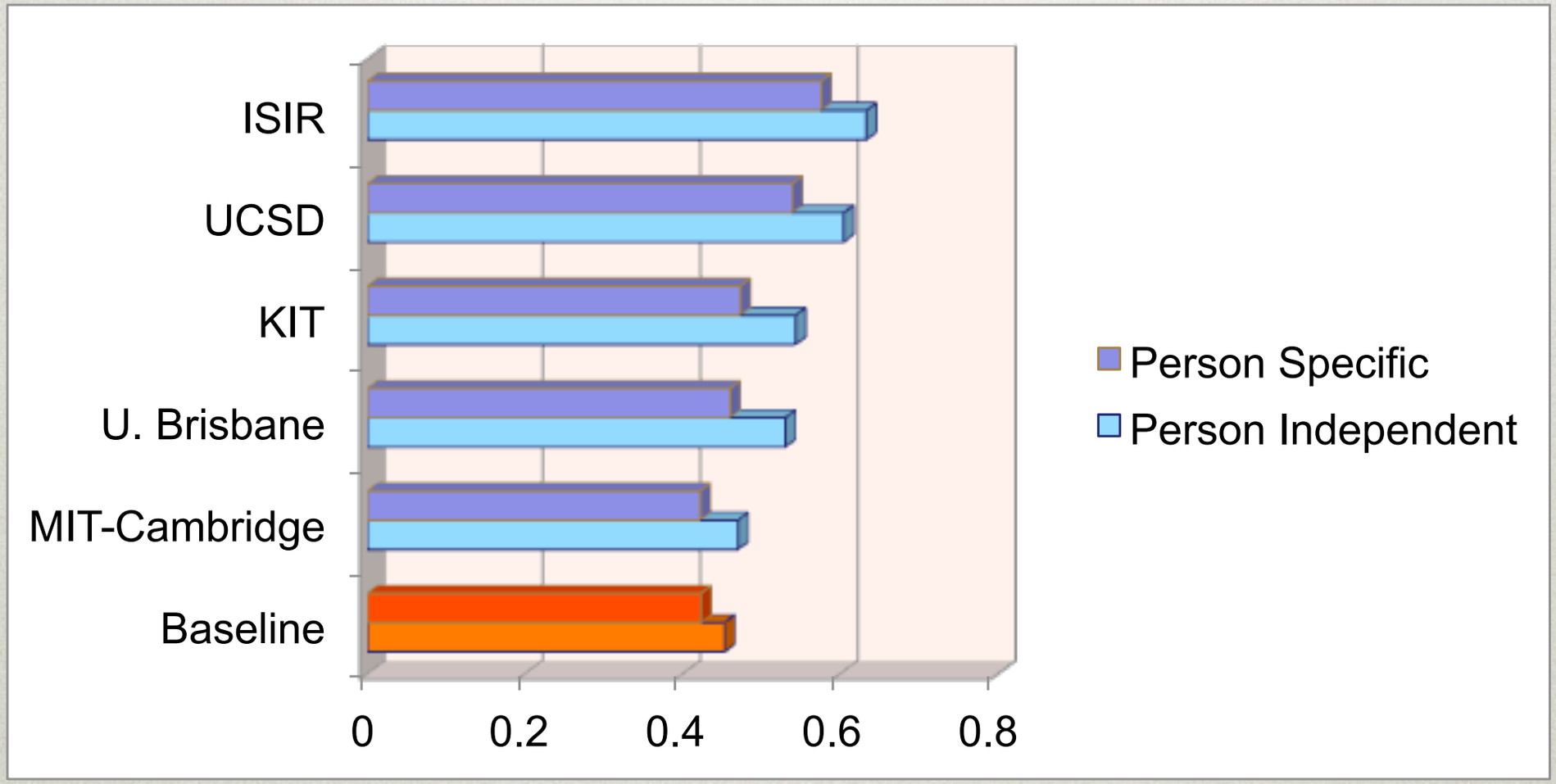
  134 Test videos

Michel Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer, 'Meta-Analysis of the First Facial Expression Recognition Challenge', IEEE Transactions on Systems, Man, and Cybernetics-B, vol. 42(4), pp. 966-979, 2012.

Discrete Emotions



Action Units



FERA 2015: Occurrence

Occurrence sub-challenge results
(weighted F1 score over SEMAINE and BP4D)

AUs	EPFL-LTS5	Rainbow Group	VicarVision	Appearance baseline	Geometric baseline
Weighted Mean	0.499	0.481	0.465	0.400	0.444

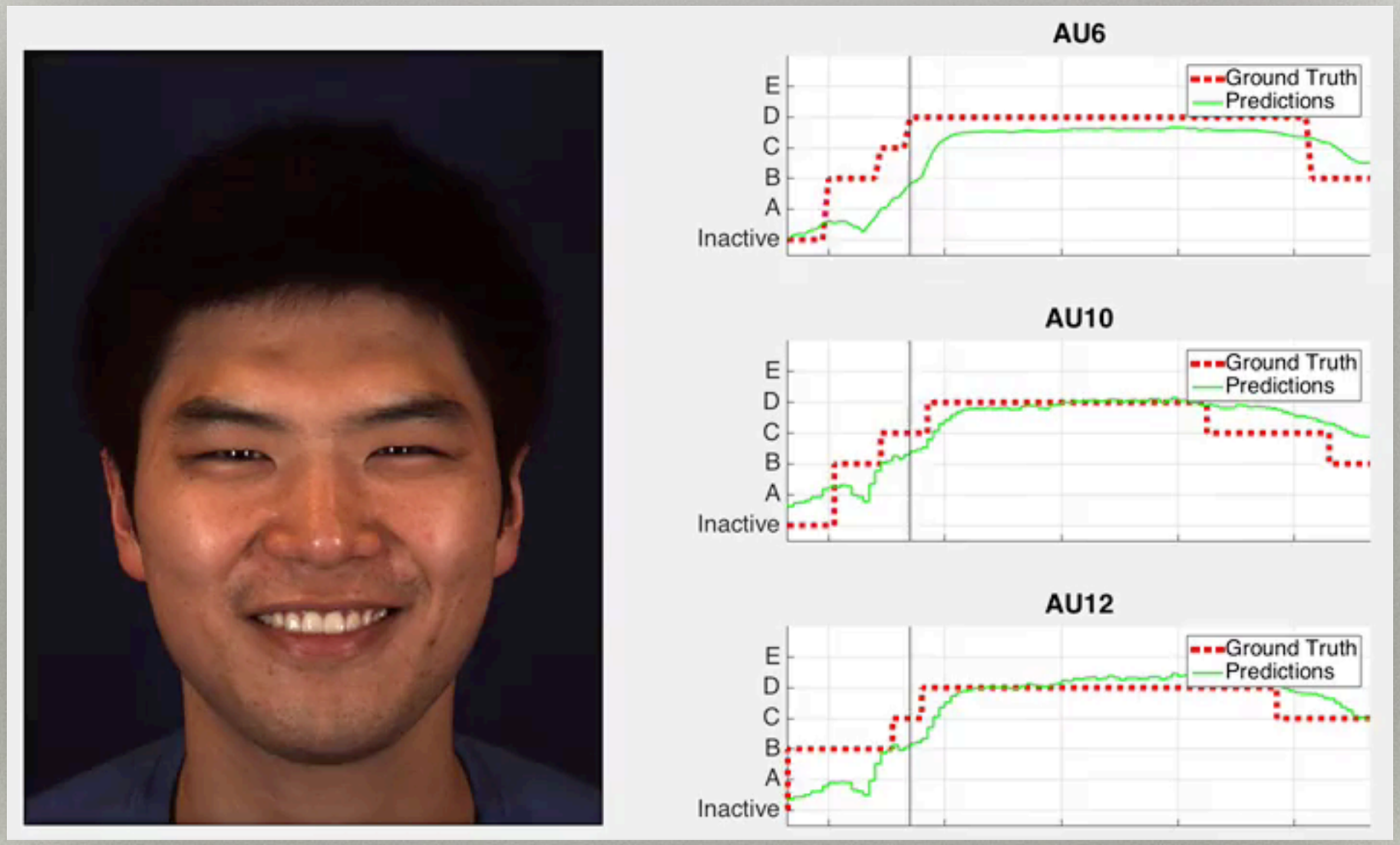
AU12
Ground Truth
Predictions
Active
Inactive
AU17
Ground Truth
Predictions
Active
Inactive
AU23
Ground Truth
Predictions
Active
Inactive

EPFL-LTS5

FERA 2015: Intensity

Fully automatic intensity sub-challenge results (ICC score, BP4D database)

AUs	ISIR	Rainbow Group	KIT	VicarVision	LaBRI	Appearance baseline	Geometric baseline
AU6	0.787	0.719	0.678	0.664	0.720	0.622	0.670
AU10	0.802	0.718	0.731	0.734	0.720	0.656	0.732
AU12	0.861	0.828	0.826	0.788	0.784	0.767	0.780
AU14	0.711	0.546	0.533	0.549	0.277	0.389	0.586
AU17	0.443	0.377	0.308	0.329	0.268	0.168	0.144
Mean	0.721	0.638	0.615	0.613	0.554	0.520	0.582


AU6
Ground Truth
Predictions
E
D
C
B
A
Inactive
AU10
Ground Truth
Predictions
E
D
C
B
A
Inactive
AU12
Ground Truth
Predictions
E
D
C
B
A
Inactive

Audio-Visual Dimensional Affect

AVEC 2012

4 affective dimensions: *Activation, Expectation, Power, Valence*

2-8 labellers for all sessions

68.4% > 5 raters 82% > 2 raters

Fully Continuous Sub-Challenge (FCSC):

Continuous time (prediction for every video frame)

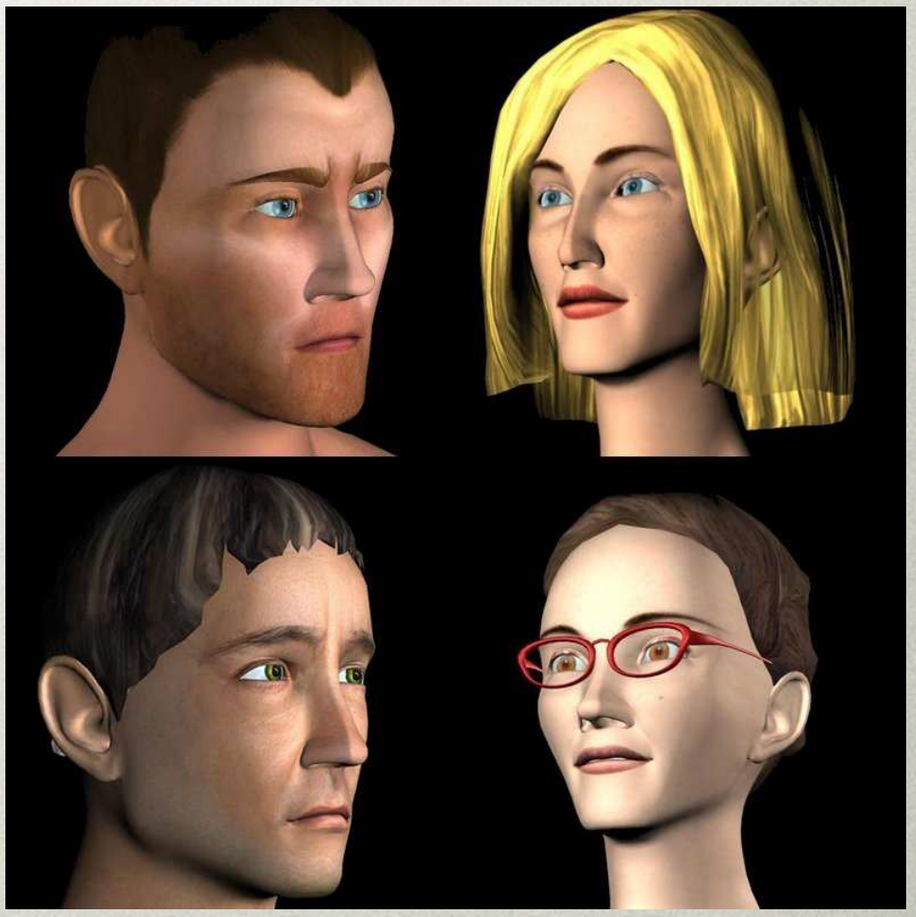
Continuous in value (real number, range -1 to +1, expectancy from 0 to 100)

Björn Schuller, Michel F. Valstar, Florian Eyben, Roddy Cowie, and Maja Pantic, 'AVEC 2012 – The Continuous Audio/Visual Emotion Challenge', proc. ACM Int'l Conf. Multimodal Interaction, pp. 449-456, October 2012

SEMAINE Database

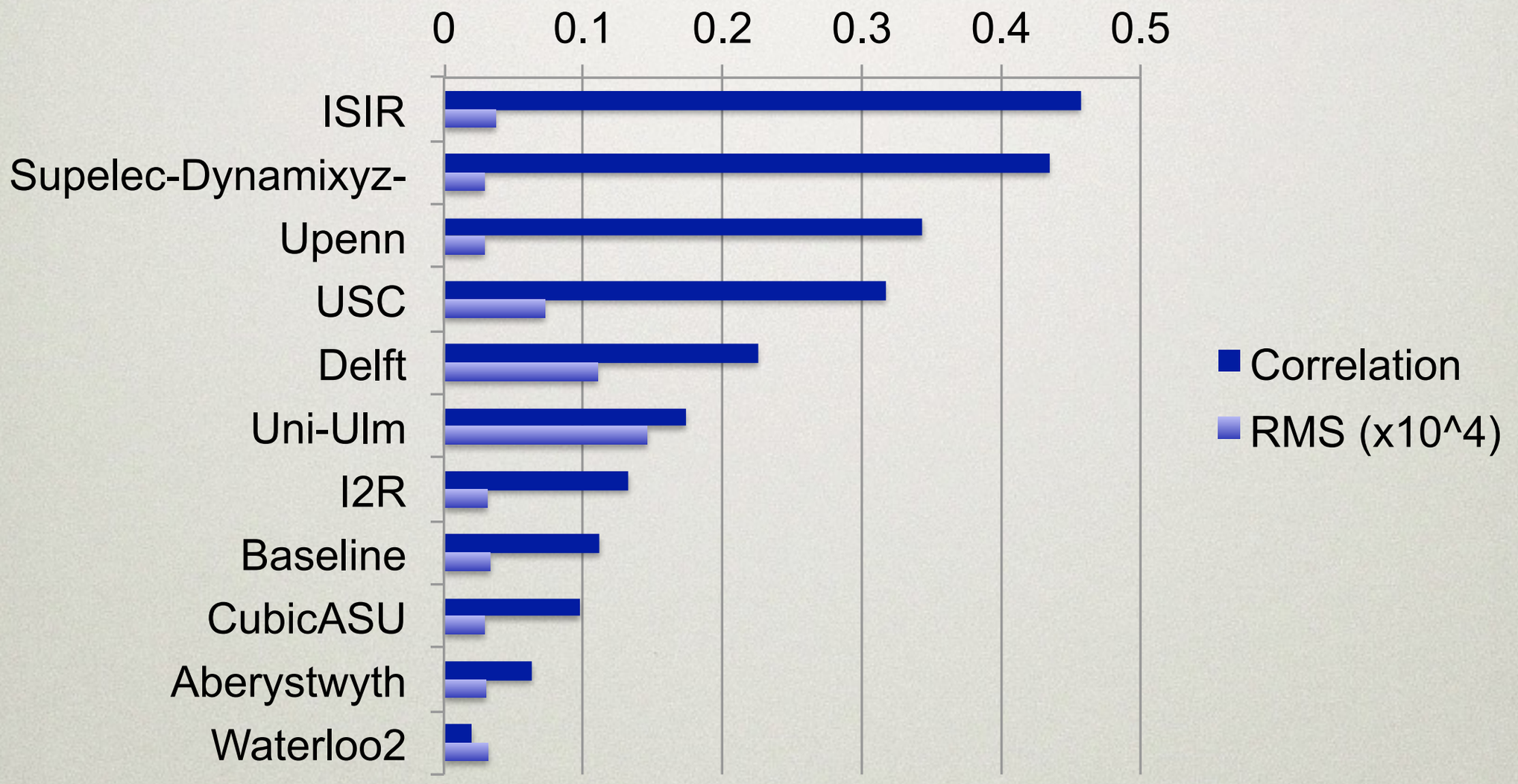
High-quality recordings of dyadic interactions between user and emotionally stereotyped operator

- 959 interactions
- 150 participants
- Database blocks:
  - Solid-SAL
  - Semi-automatic SAL
  - Automatic SAL

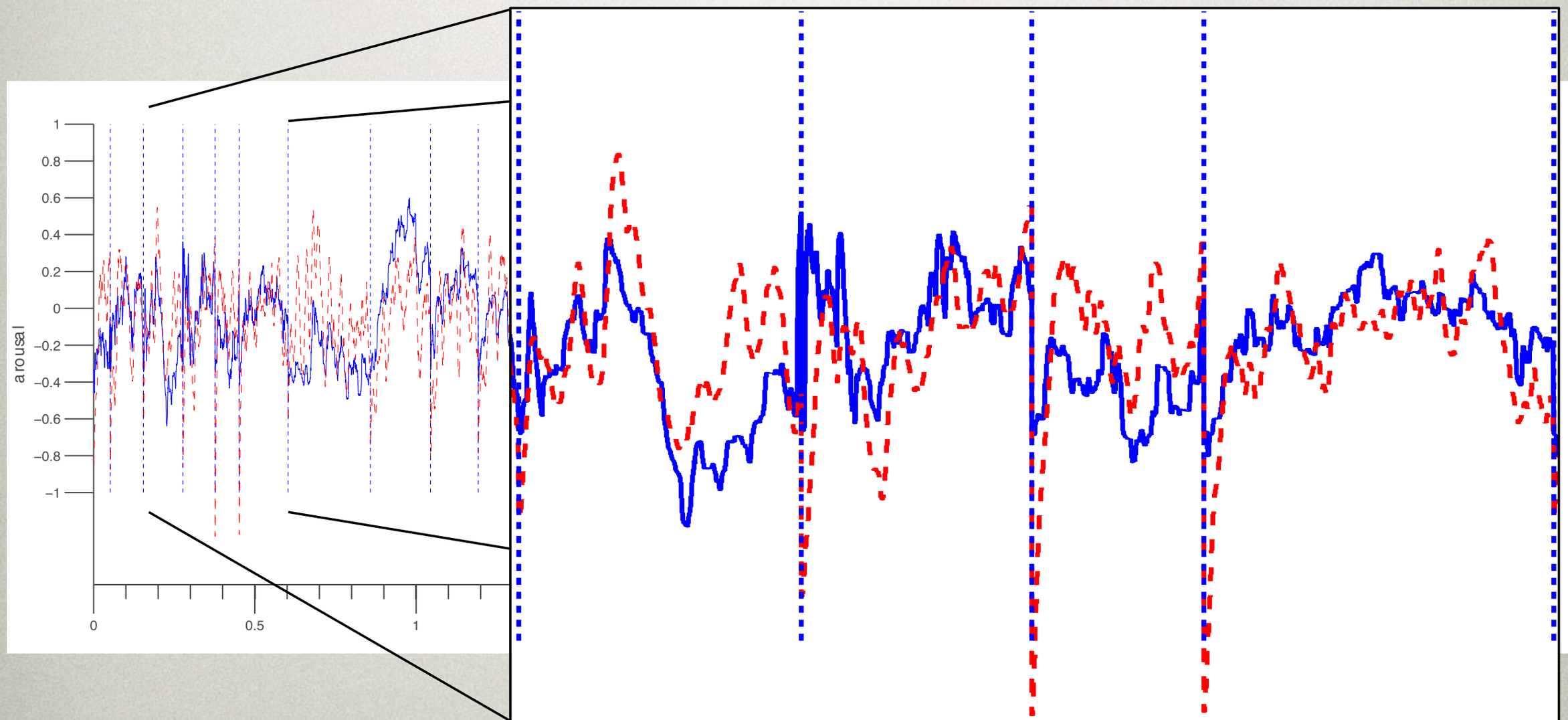


Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder, 'The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent', IEEE Transactions on Affective Computing, vol. 3, pp 5-17, 2012

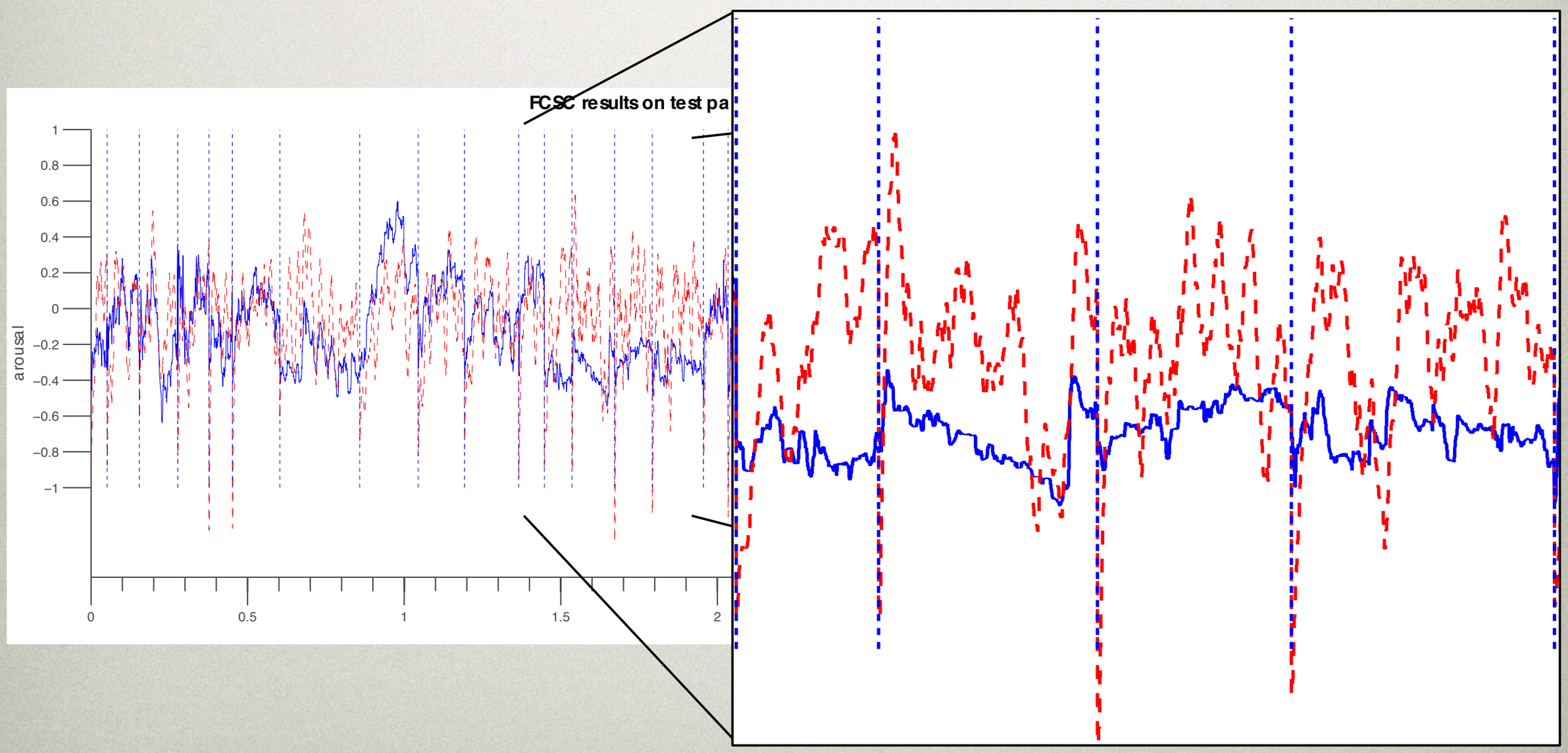
Fully Continuous Sub-Challenge



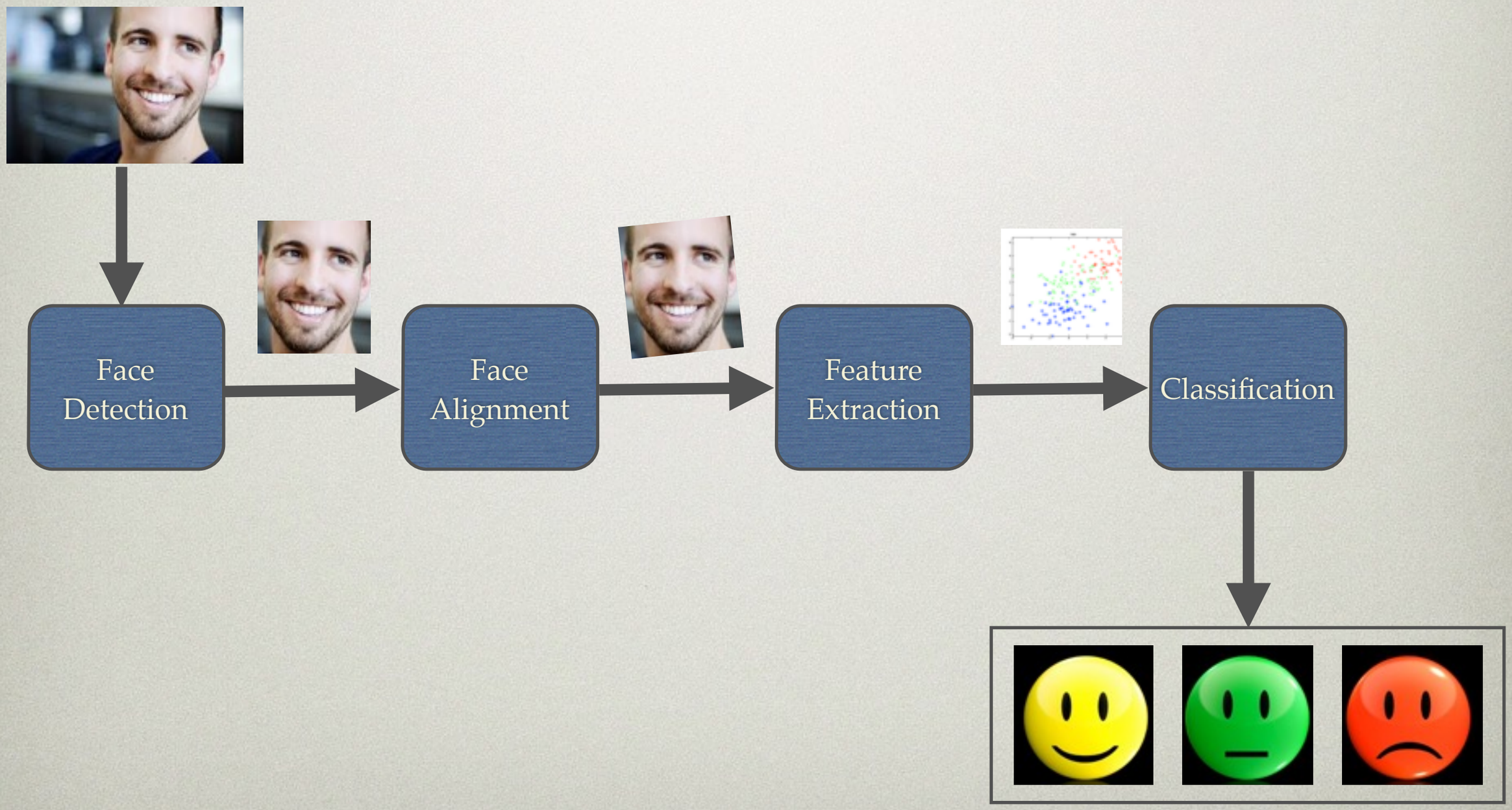
Visualisation of Best FCSC Result



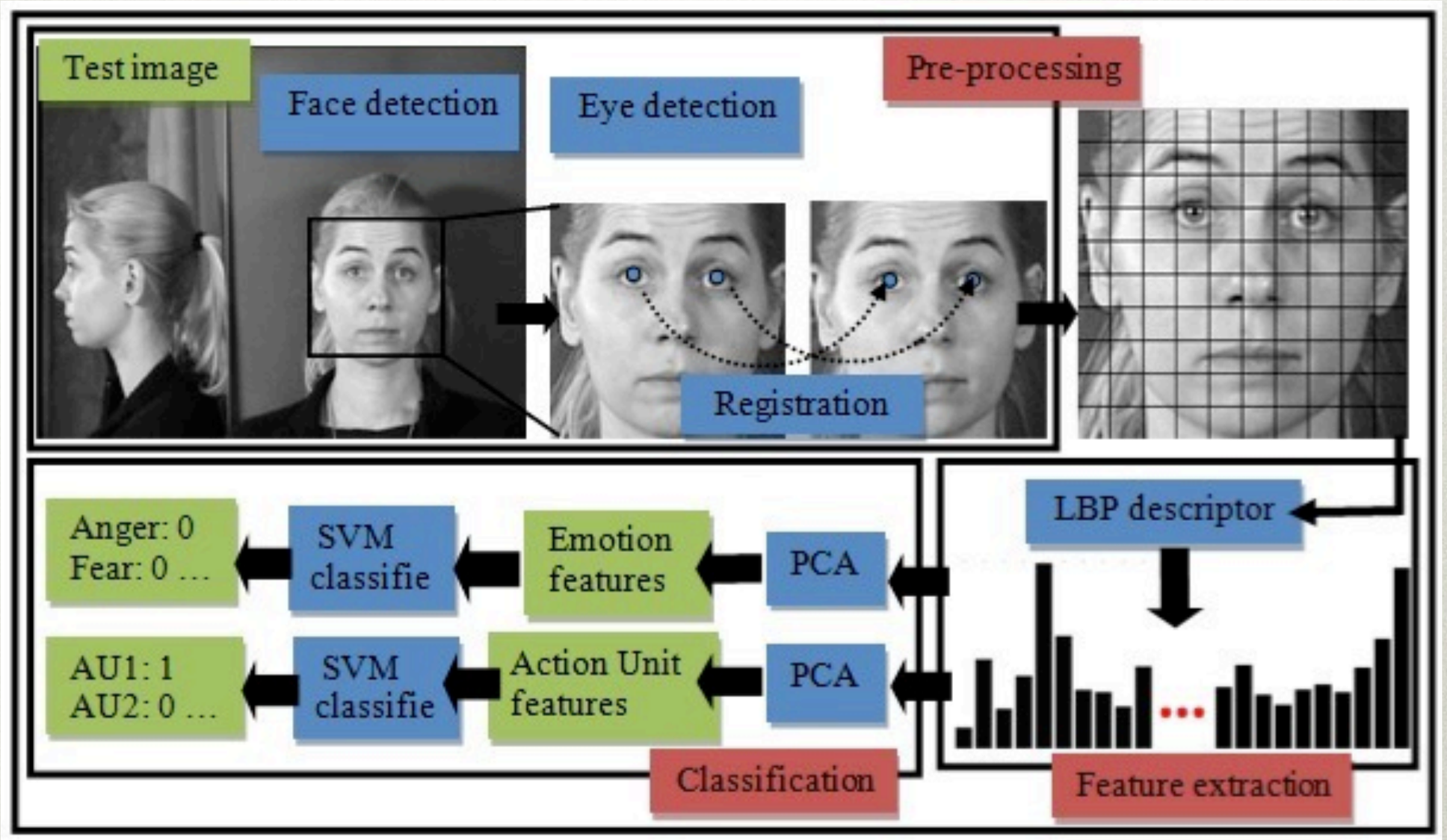
And Some Poor Results



The Basic Pipeline



Basic System



Pre-Processing

- Goals:

  1. To detect location of the face
  2. To remove variation of head pose
     or to allow using mode-specific models
  3. To remove variations between people
  4. To remove variations within one person

Face Detection

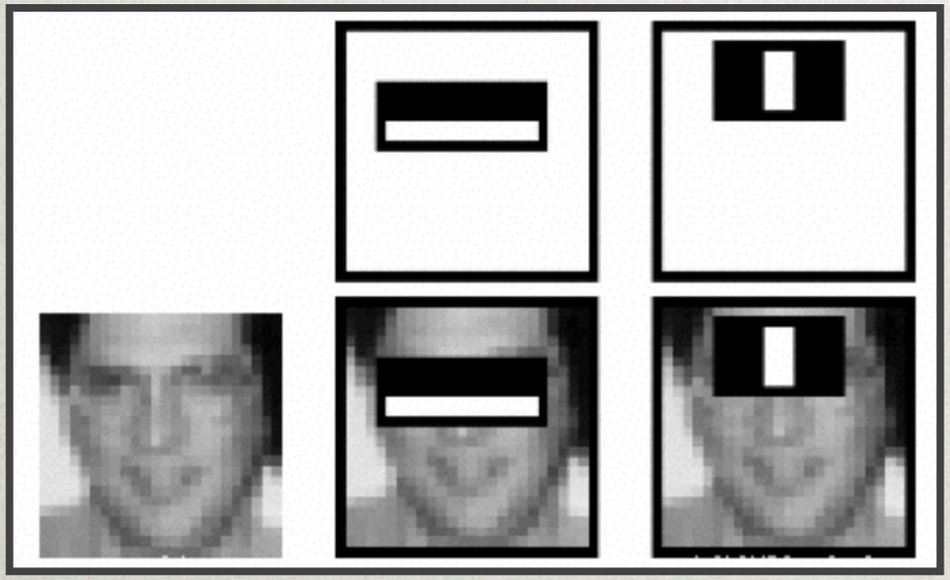
- Goals:

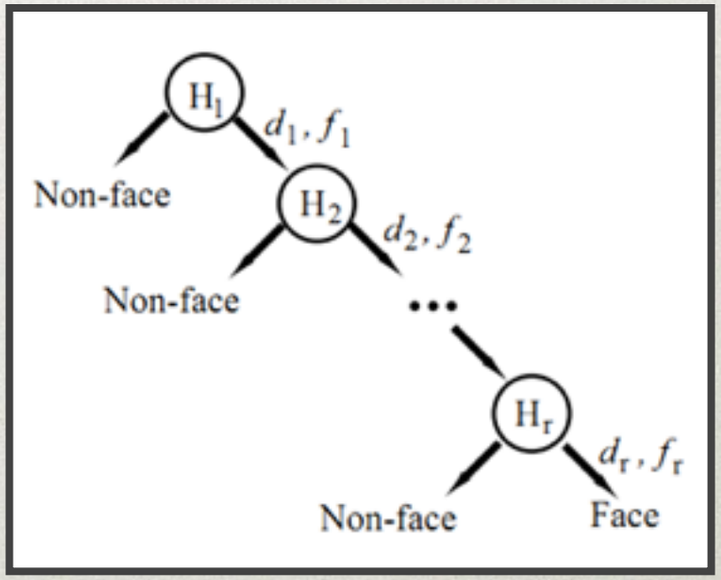
    1. To detect location of the face (x,y location)
    2. To detect extend of the face (height, width)

Viola & Jones

- Goals:

    1. To detect location of the face (x,y location)
    2. To detect extend of the face (height, width)





Viola & Jones (2003). *'Fast multi-view face detection'*, Technical report, Mitsubishi Electric Research Laboratory

Zhu & Rhamanan

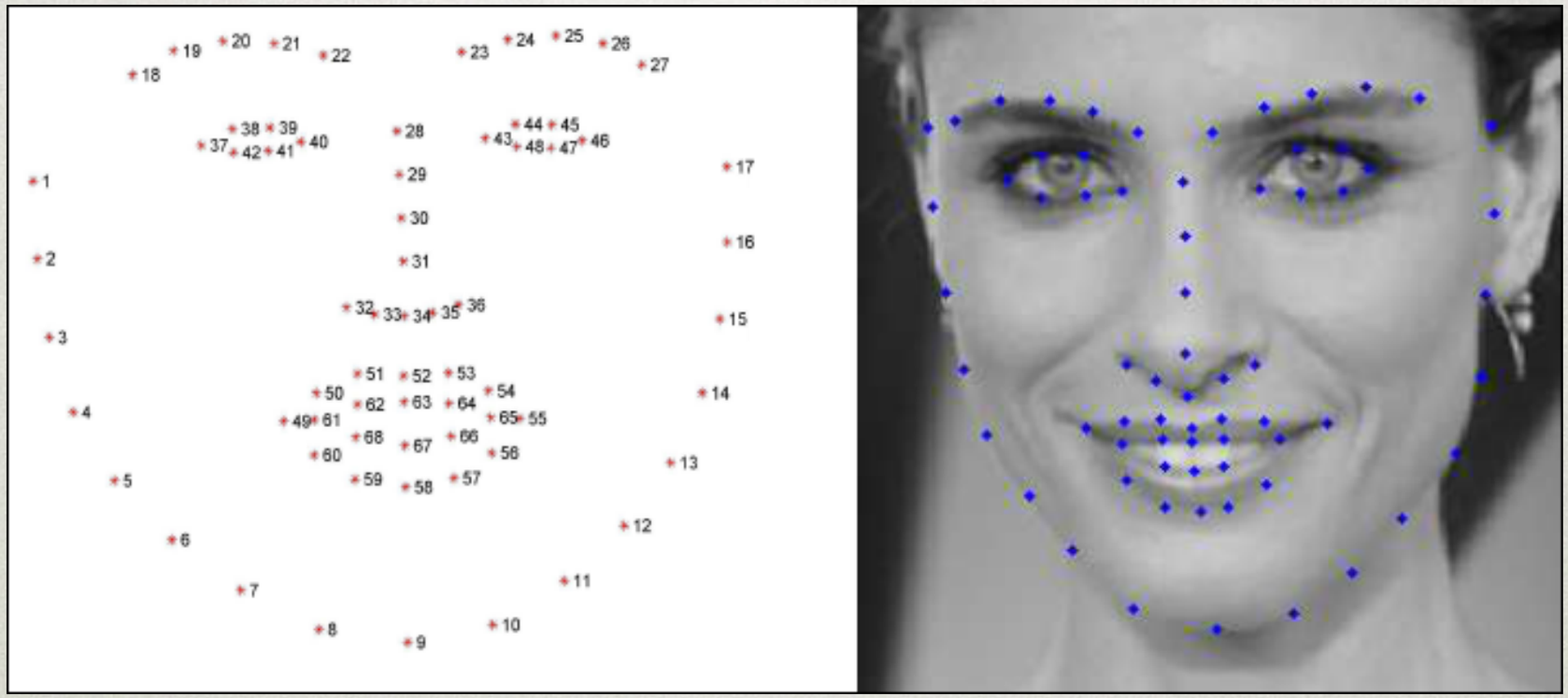
- Goals:

  1. To detect location of the face (x,y location)

  2. To detect extend of the face (height, width)

Zhu & Ramanan (2013). *'Face detection, pose estimation and landmark localization in the wild'* Computer Vision and Pattern Recognition 2013

Face Alignment

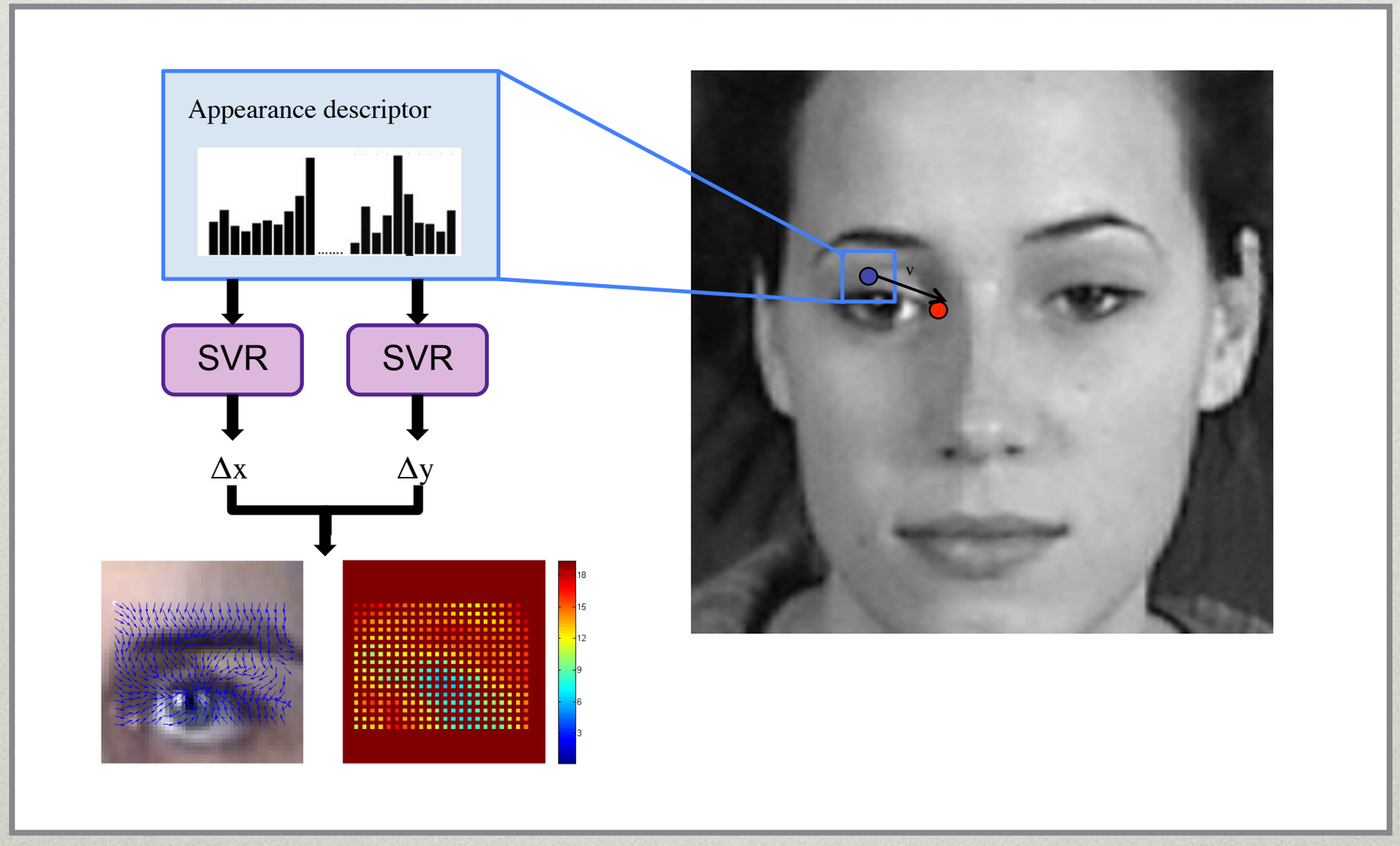


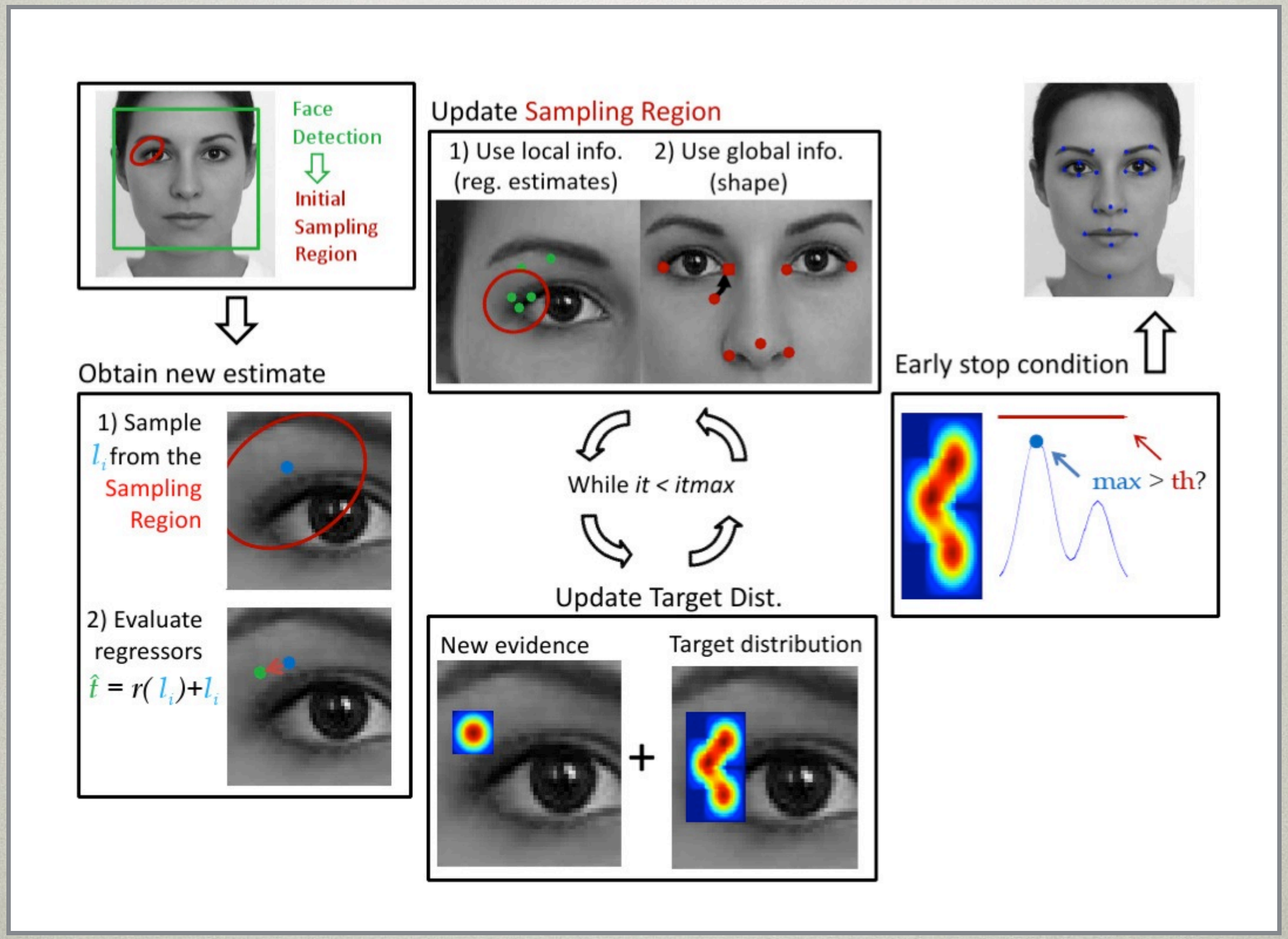
High variation in face images caused due to

- Different head poses
- Different facial expressions

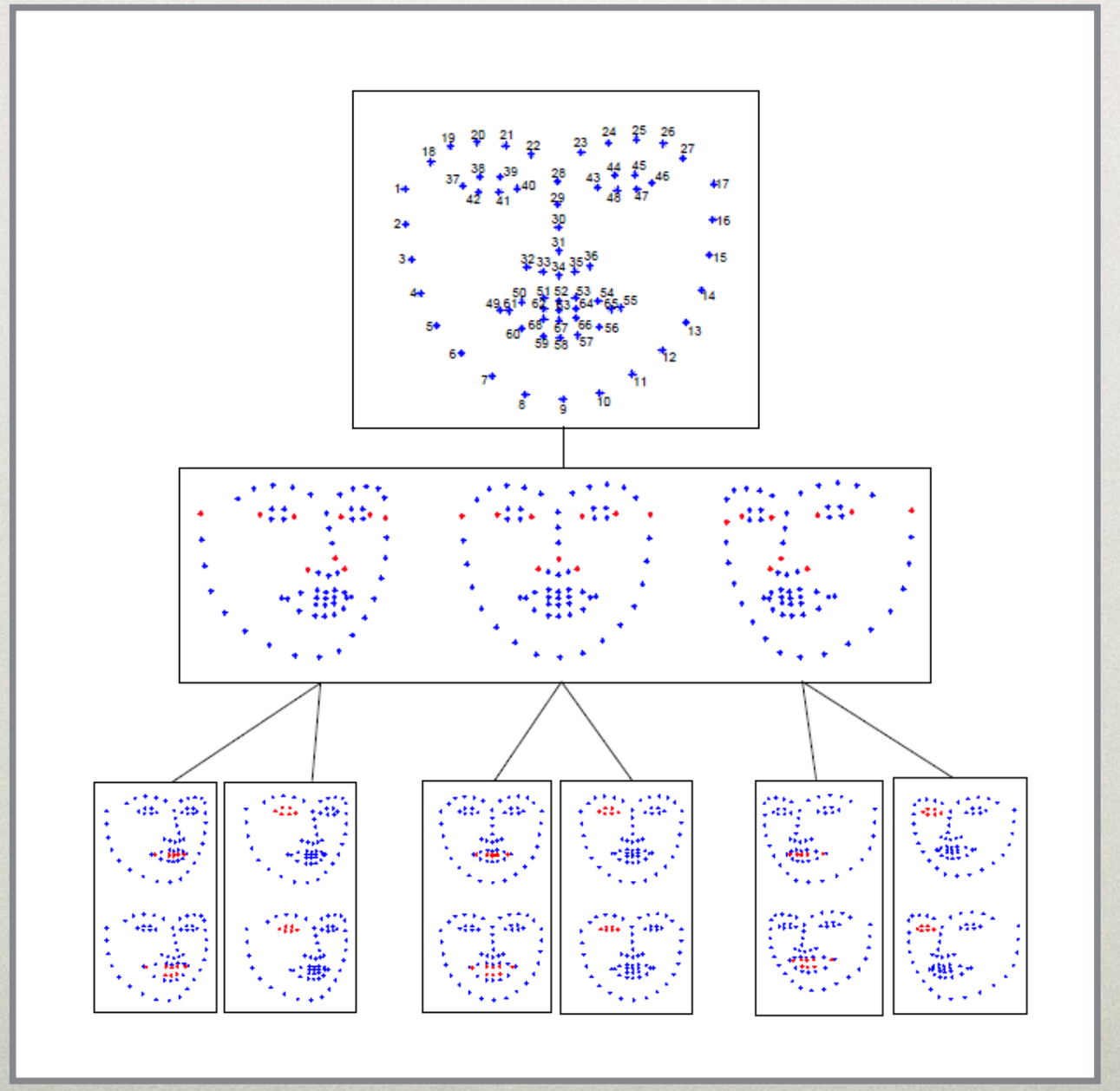
Results in non linear changes in the shape and appearance of facial points

Regression-Based Approaches

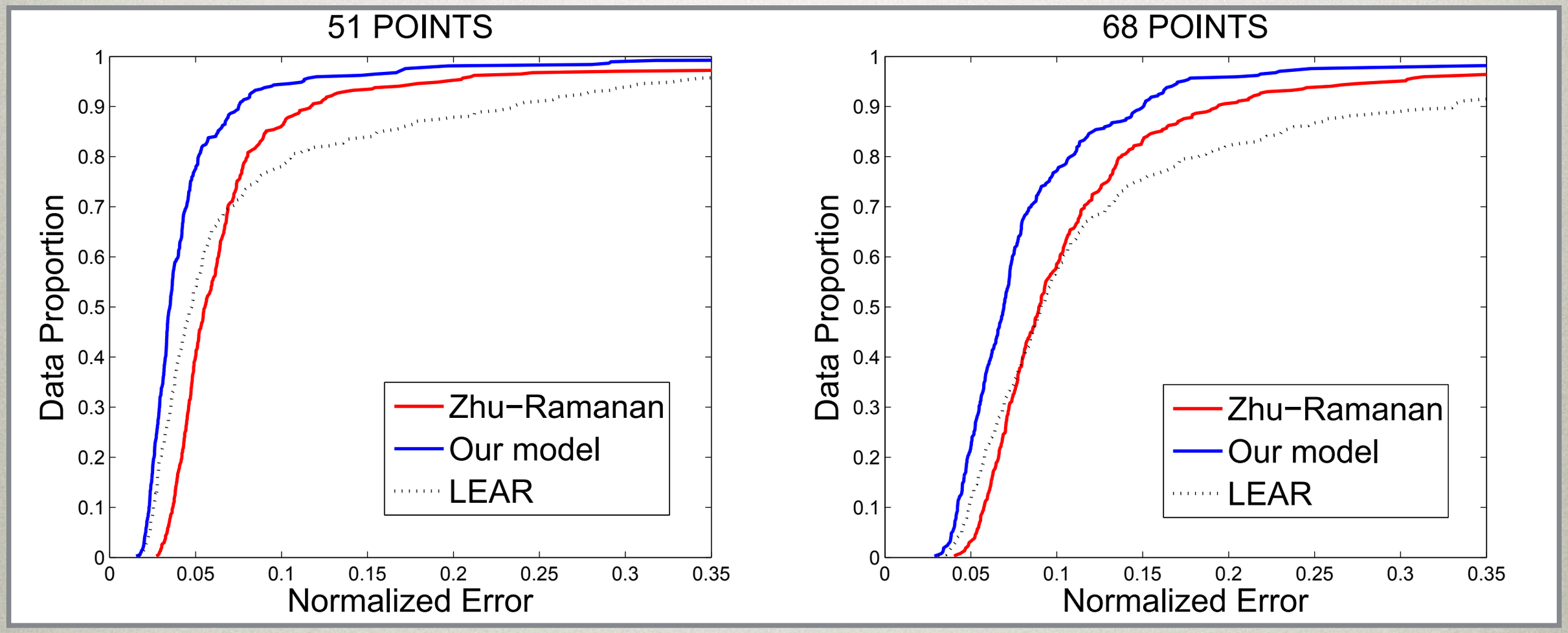



Face Detection
Initial Sampling Region
Update Sampling Region
1) Use local info. (reg. estimates)
2) Use global info. (shape)
Obtain new estimate
1) Sample l_i from the Sampling Region
2) Evaluate regressors
$\hat{t} = r(l_i) + l_i$
While it < itmax
Update Target Dist.
New evidence
Target distribution
Early stop condition
max > th?

Mode-Specific Solution

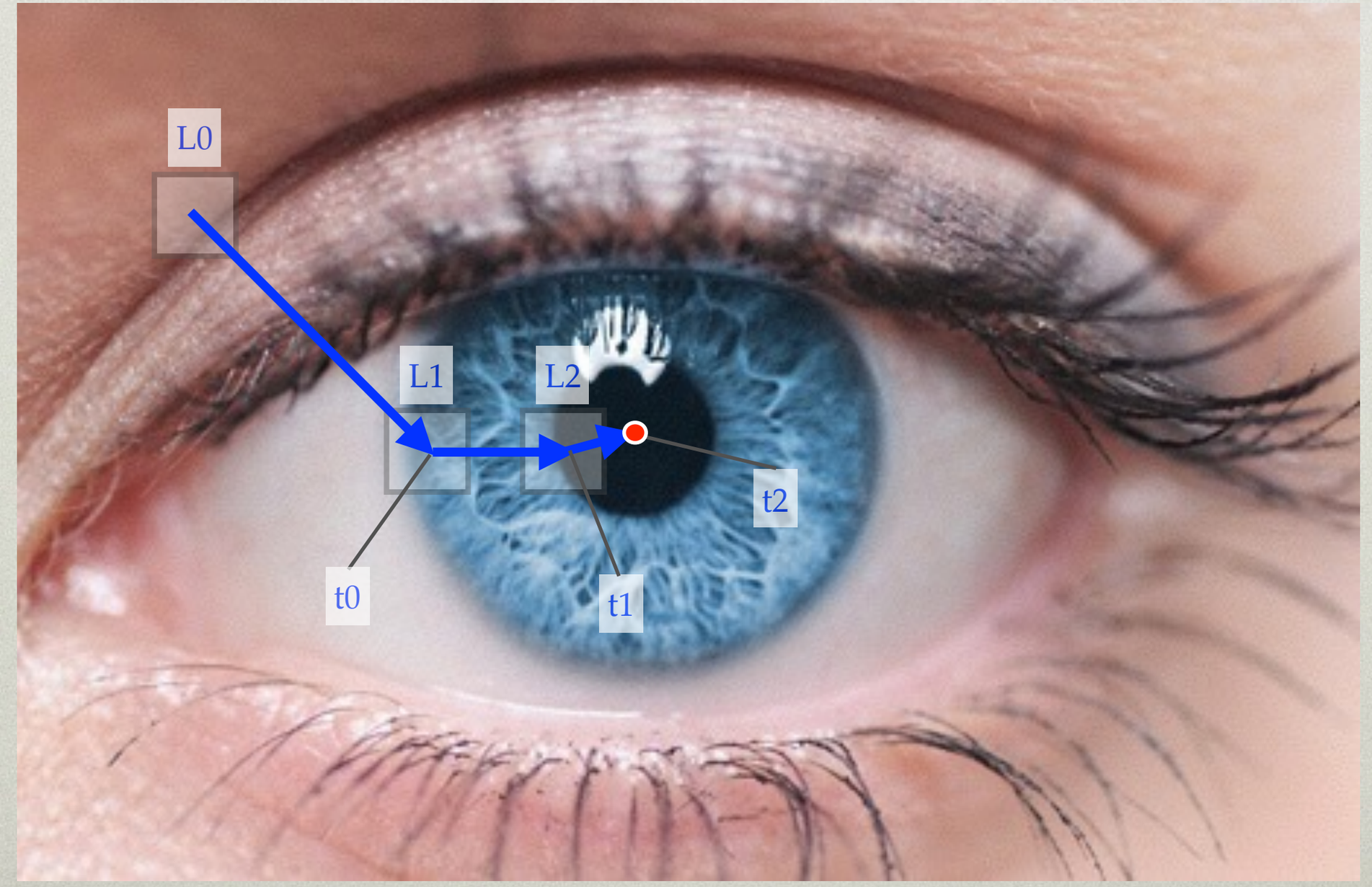


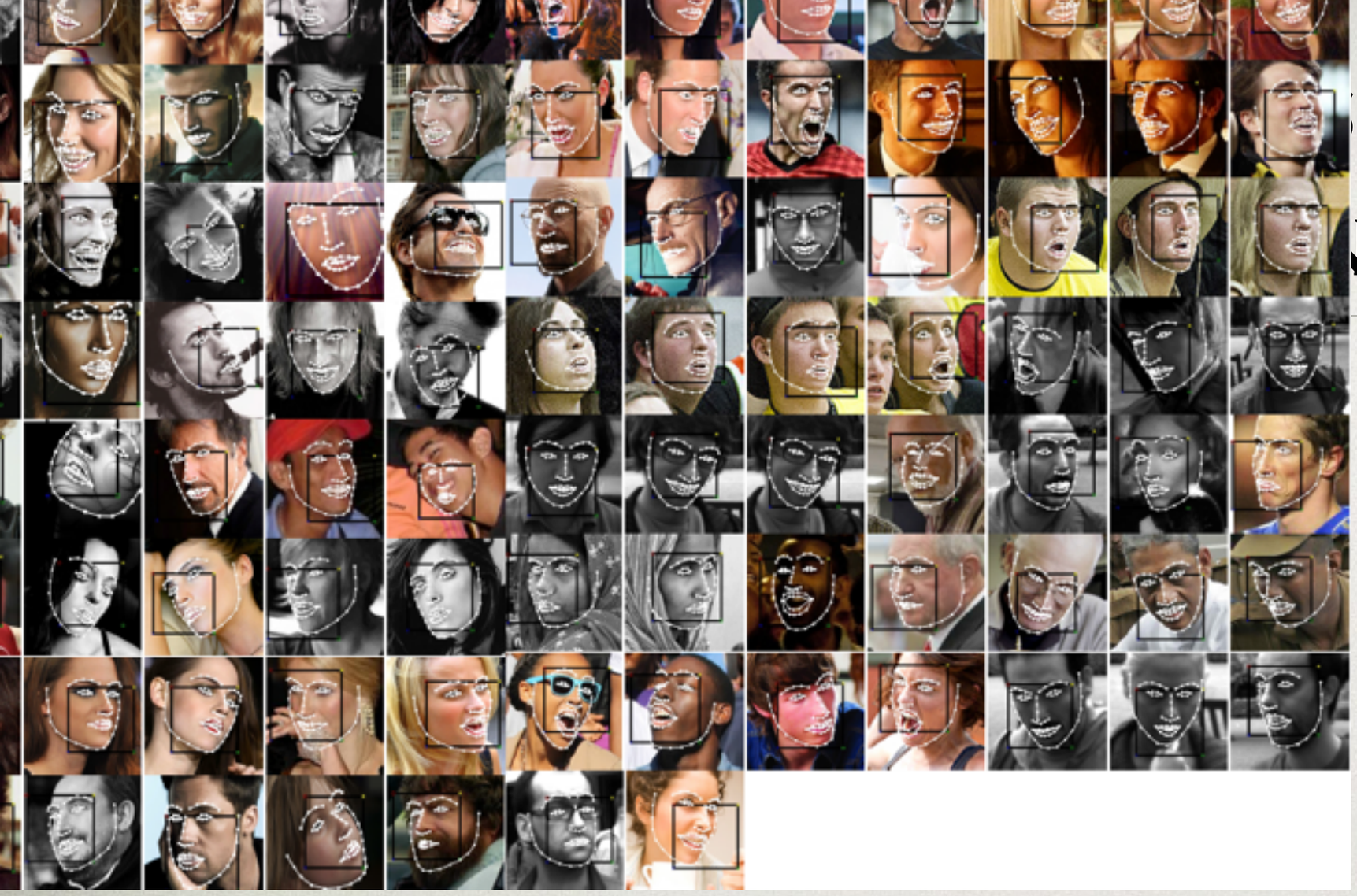
Results



Shashank Jaiswal, Timur Almaev, Michel F. Valstar, 'Guided Unsupervised Learning of Mode Specific Models for Facial Point Detection in the Wild', ICCV workshops 2013

Cascaded Regression (SDM)





CADED

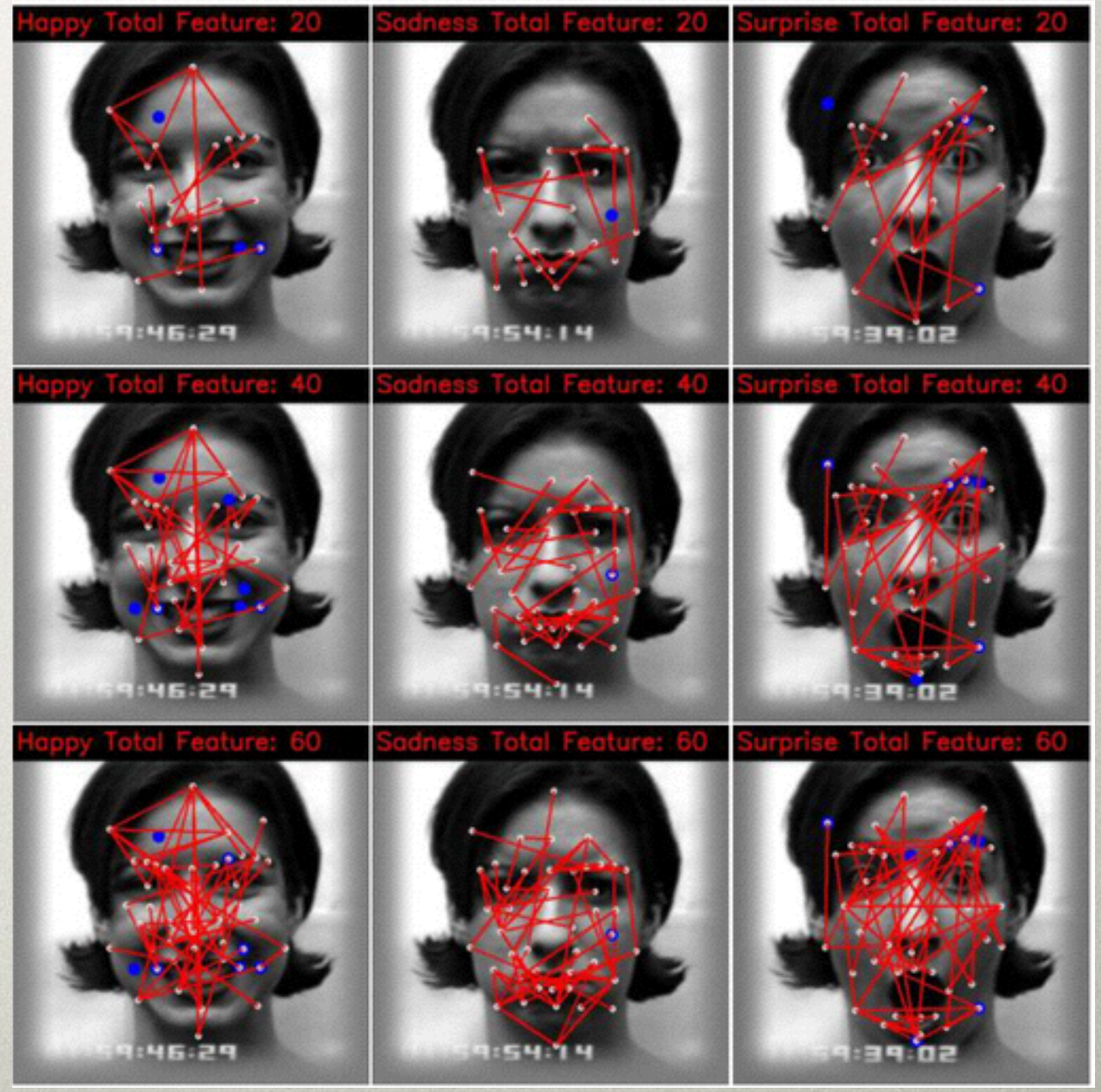
N

Features

- Types of features:
  1. Geometric Features
  2. Appearance Descriptors
  3. Hybrid

Geometric Features

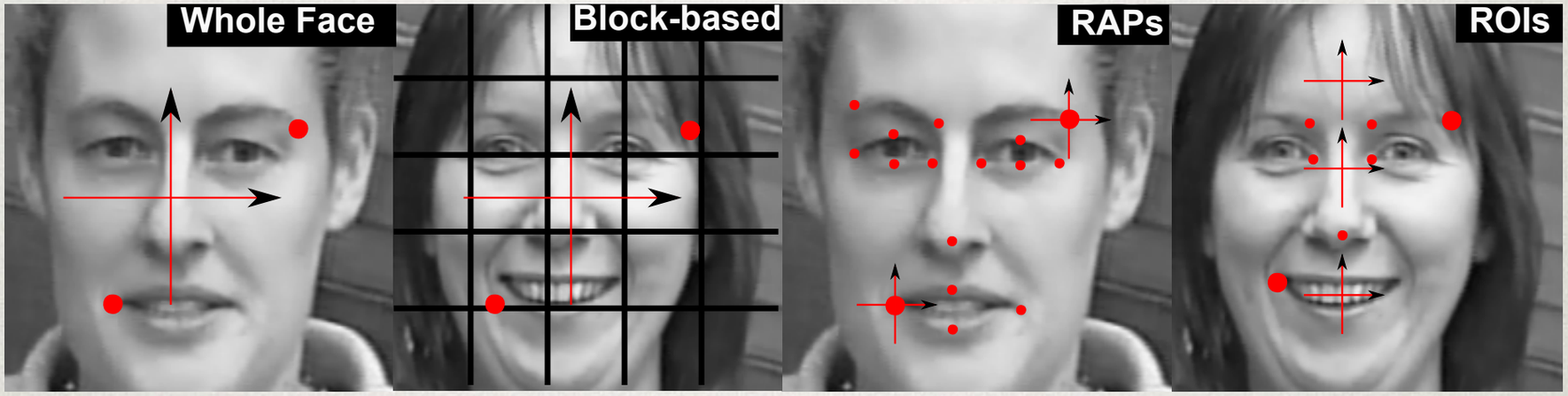
- Expert rules
- Simple rules
- Temporal extensions
- Features of pairs / triplets of points

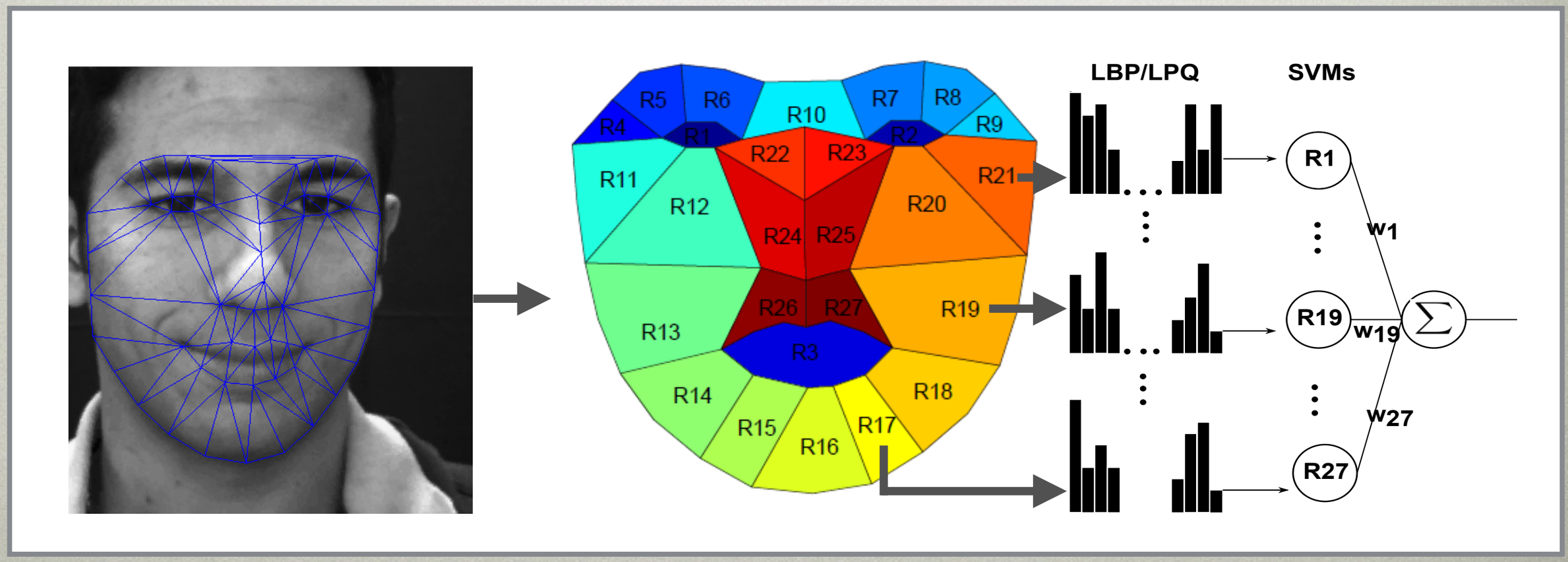


Appearance Descriptors

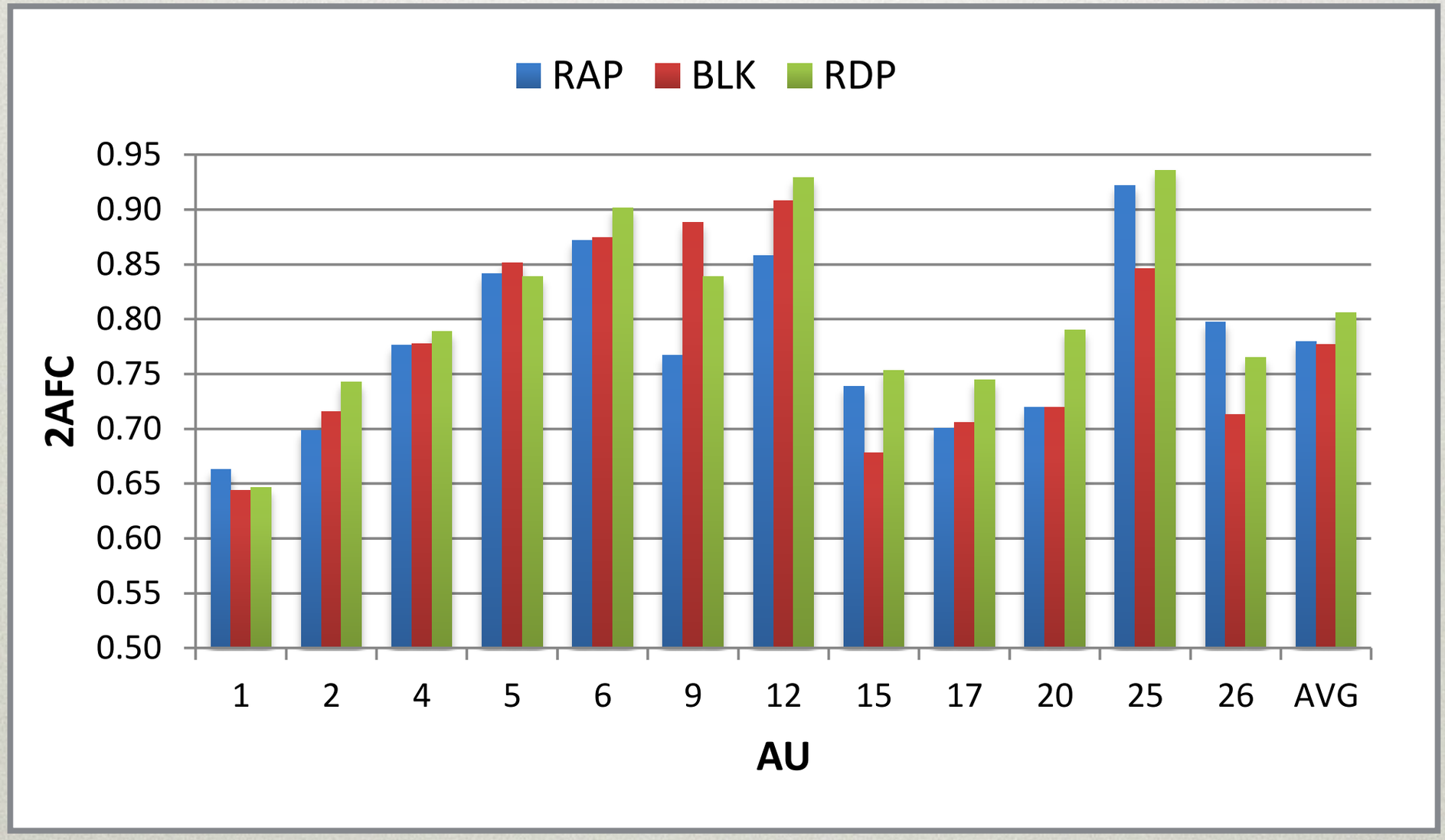
- Commonly used appearance descriptors:
  1. Gabor
  2. Local Binary Patterns
  3. HOG/SIFT

Where to Extract?





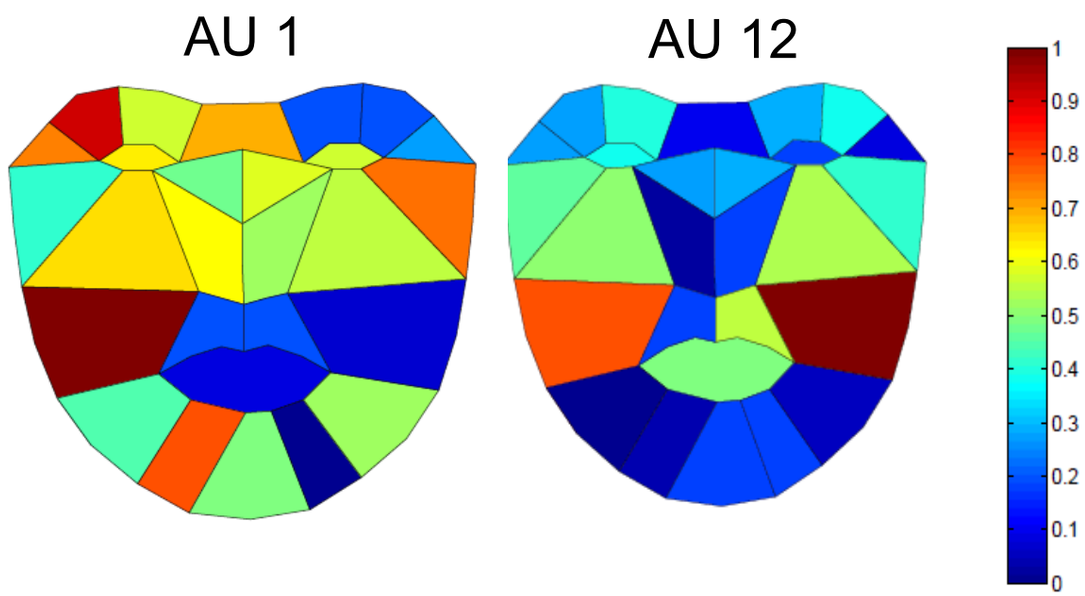
RDP Results



Jiang, Martinez, Valstar, and Pantic (2014), *'Decision Level Fusion of Domain Specific Regions for Facial Action Recognition'*, Proc. Int'l Conf. Pattern Recognition (ICPR)

ANNs vs SVM fusion

Importance of regions



SVM fusion vs ANN (DISFA)

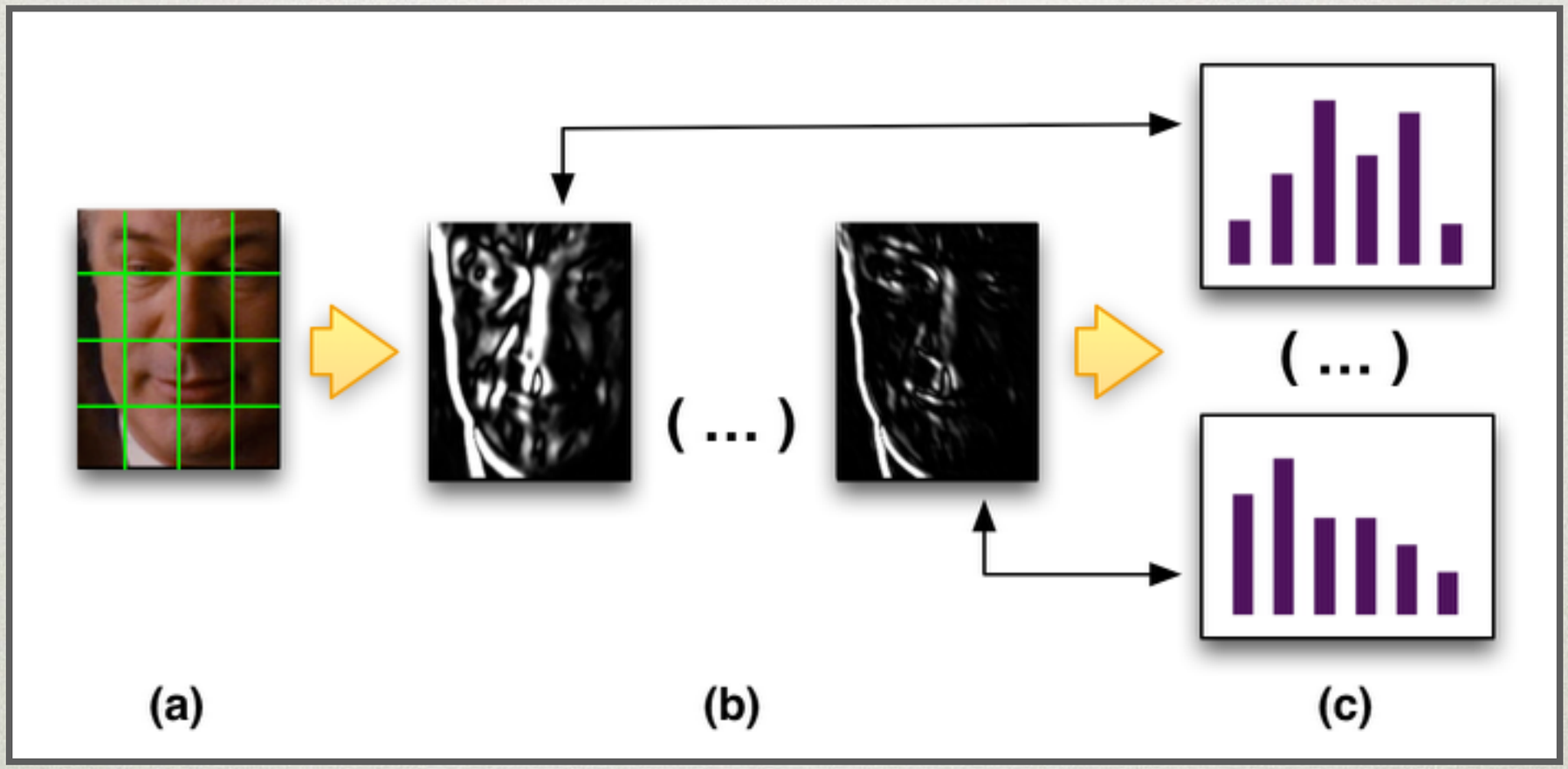
AU	Jiang et al. [5]	Our method
1	0.64	**0.81**
2	**0.72**	0.71
4	0.53	**0.65**
6	0.72	**0.82**
7	**0.68**	0.64
10	**0.66**	0.60
12	0.74	**0.79**
15	0.53	**0.67**
17	0.70	**0.73**
18	**0.75**	0.72
25	0.57	**0.60**
26	0.53	**0.59**
Mean	0.65	**0.69**

Jaiswal, Martinez, Valstar (2015), *'Learning to combine local models for Facial Action Unit detection'*, Proc. IEEE Int'l Conf. Face and Gesture Recognition (FG)

Temporal Features

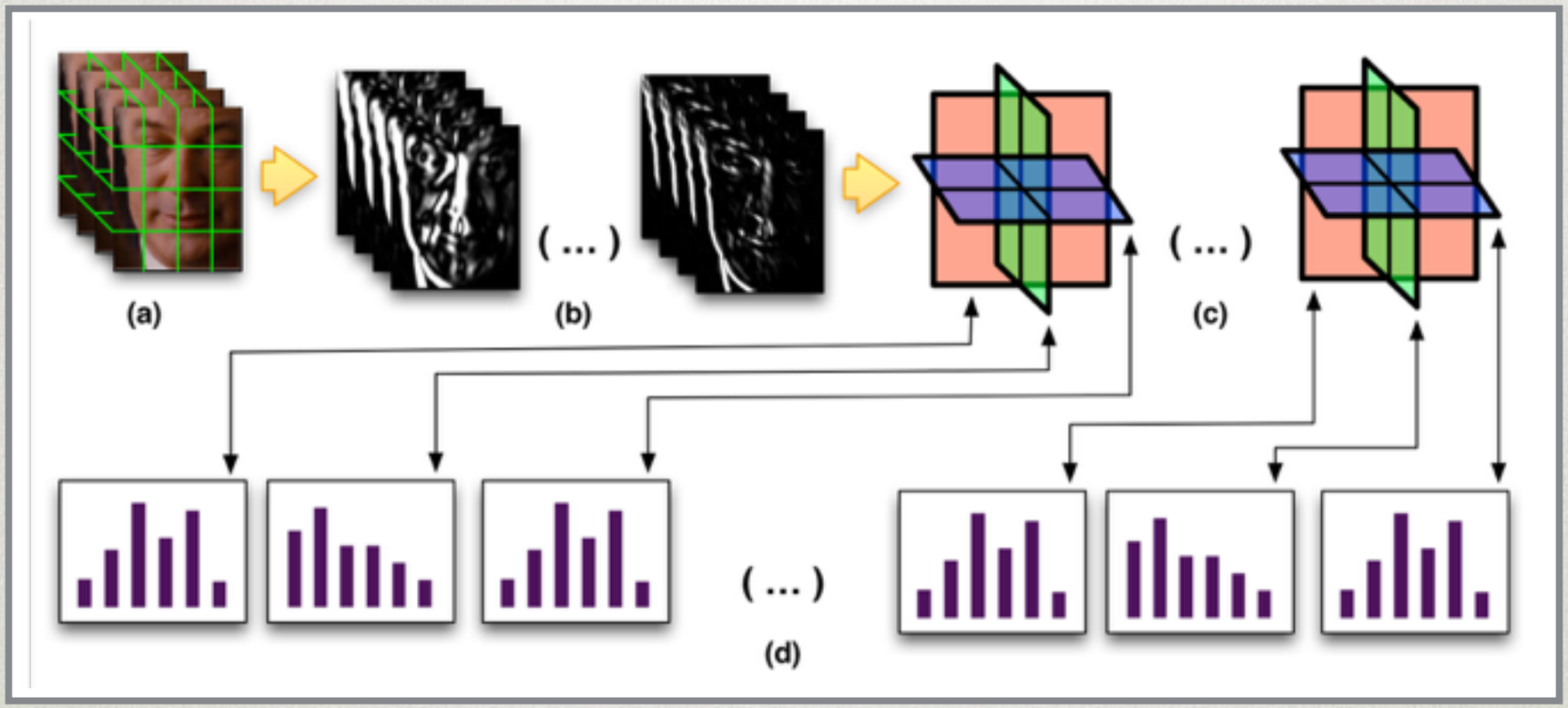
- Appearance Features describe static face appearance
- But expression recognition is largely action recognition!

Local Gabor Binary Patterns



- Gabor Filters applied to create Gabor Pictures
- Local Binary Patterns applied on Gabor Pictures
- Attains higher accuracy than either applied separately

LGBP-TOP



- To model changes in appearance Volume Local Binary Patterns were introduced
- Proved problematic due to large feature dimensionality
- LBP-Three Orthogonal Planes calculates LBP in xy, xt, and yt planes only
- We extended LGBP to LGBP-TOP in a similar fashion

LGBP-TOP results

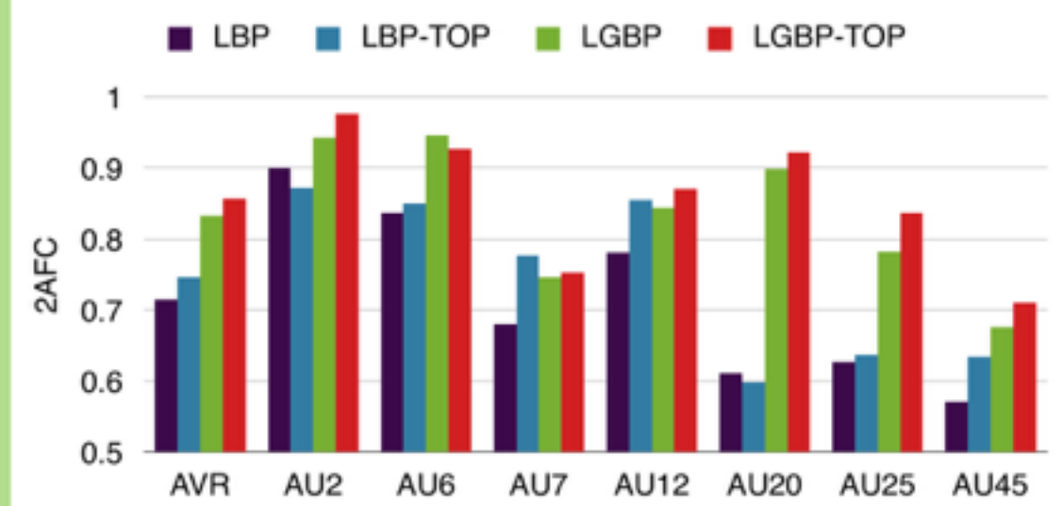


Figure 2: MMI database performance

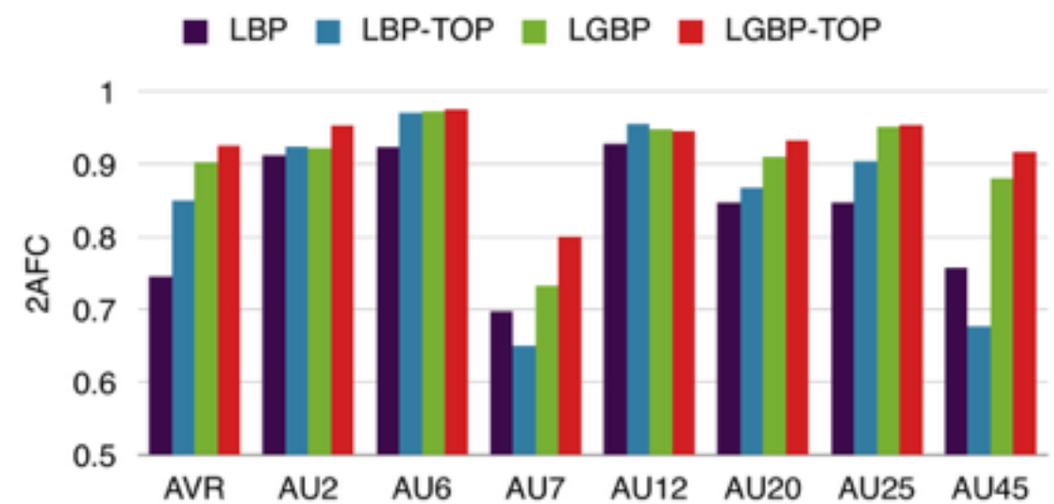


Figure 3: Cohn-Kanade database performance

Almaev & Valstar (2013), *'Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition'*, Proc. Affective Computing and Intelligent Interaction (ACII'13)

Robust to Alignment?

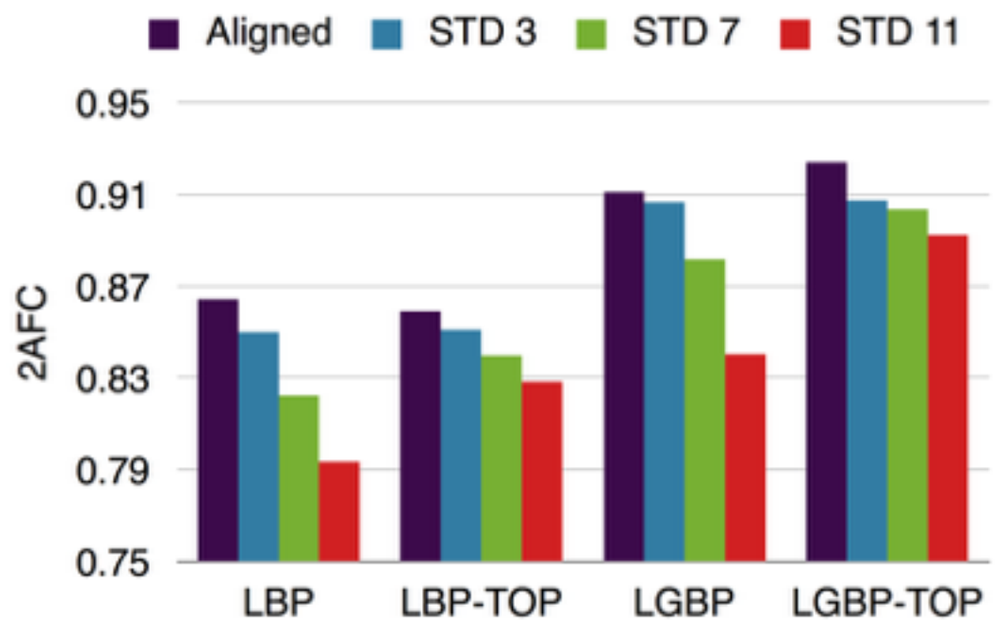


Figure 4: Alignment sensitivity analysis

- When introducing artificial rotation noise, TOP features are most resilient.

Machine Learning for Expression Recognition

- Very large number of methods proposed
  - SVM
  - HMM
  - ANN
  - CRF
  - Gaussian Processes
  - RVM
- Architectures depend on the problem

Class Types

	Multiclass	Multiple Independent Binary	Regression
Discrete Emotions	✓		
Action Units (Occurrence)		✓	
Action Units (Intensity)		✓	✓
Action Units (Temporal Phases)		✓	
Dimensional Affect			✓
Higher-order	✓	✓	✓

AC/SSP

- Affective Computing:

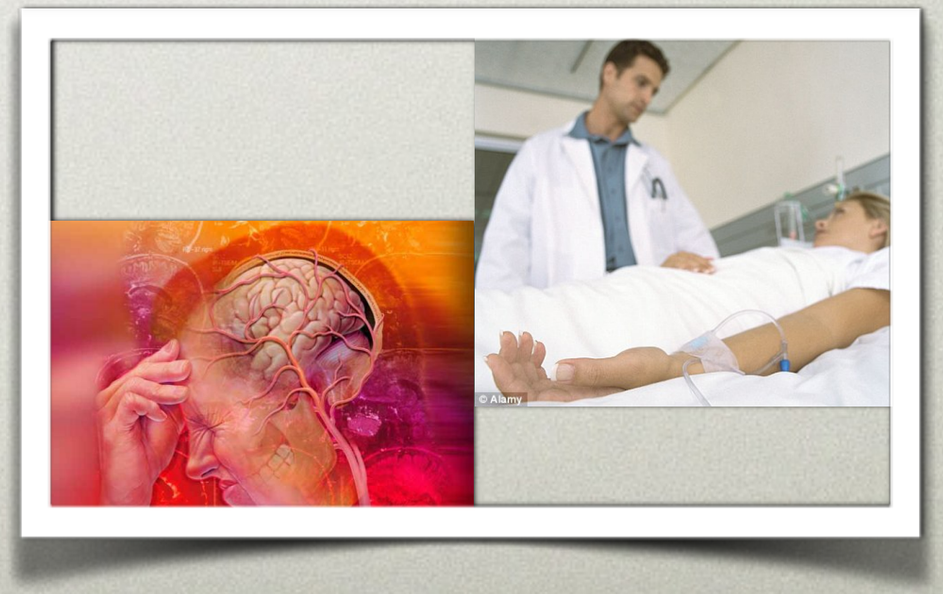
> "…[is] computing that relates to, arises from, or deliberately influences emotion or other affective phenomena." (Rosalind Picard)

- Social Signal Processing:

> "…[is a] research and technological domain that aims at providing computers with the ability to sense and understand human social signals." (Vinciarelli et al.)

Proposition

- Great opportunities for the application of existing AC/SSP methods in medicine
- Novel diagnosis, monitoring, and treatment options will have massive societal impact
- Feed-back to AC/SSP with new data and challenges to overcome

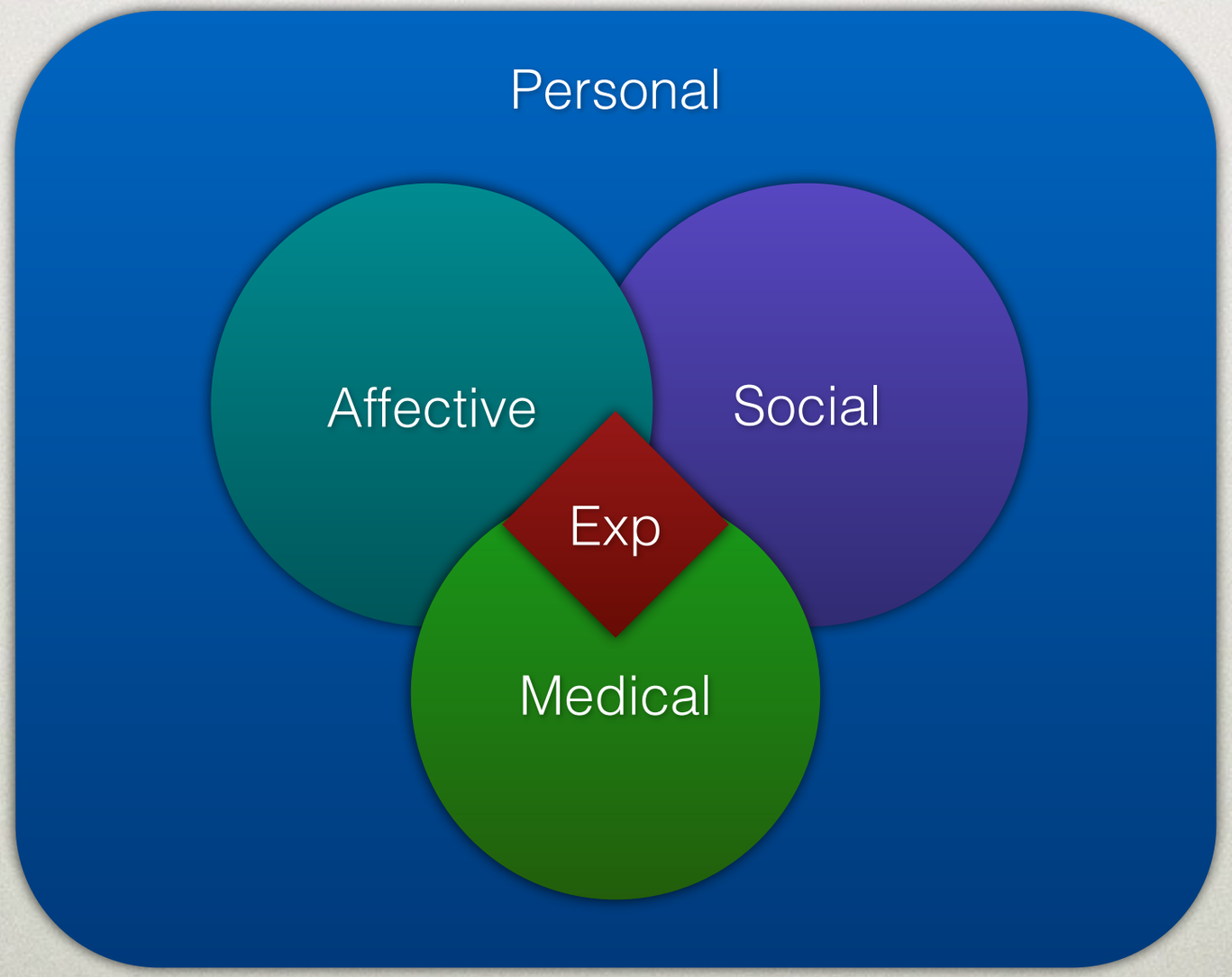


Behaviomedics

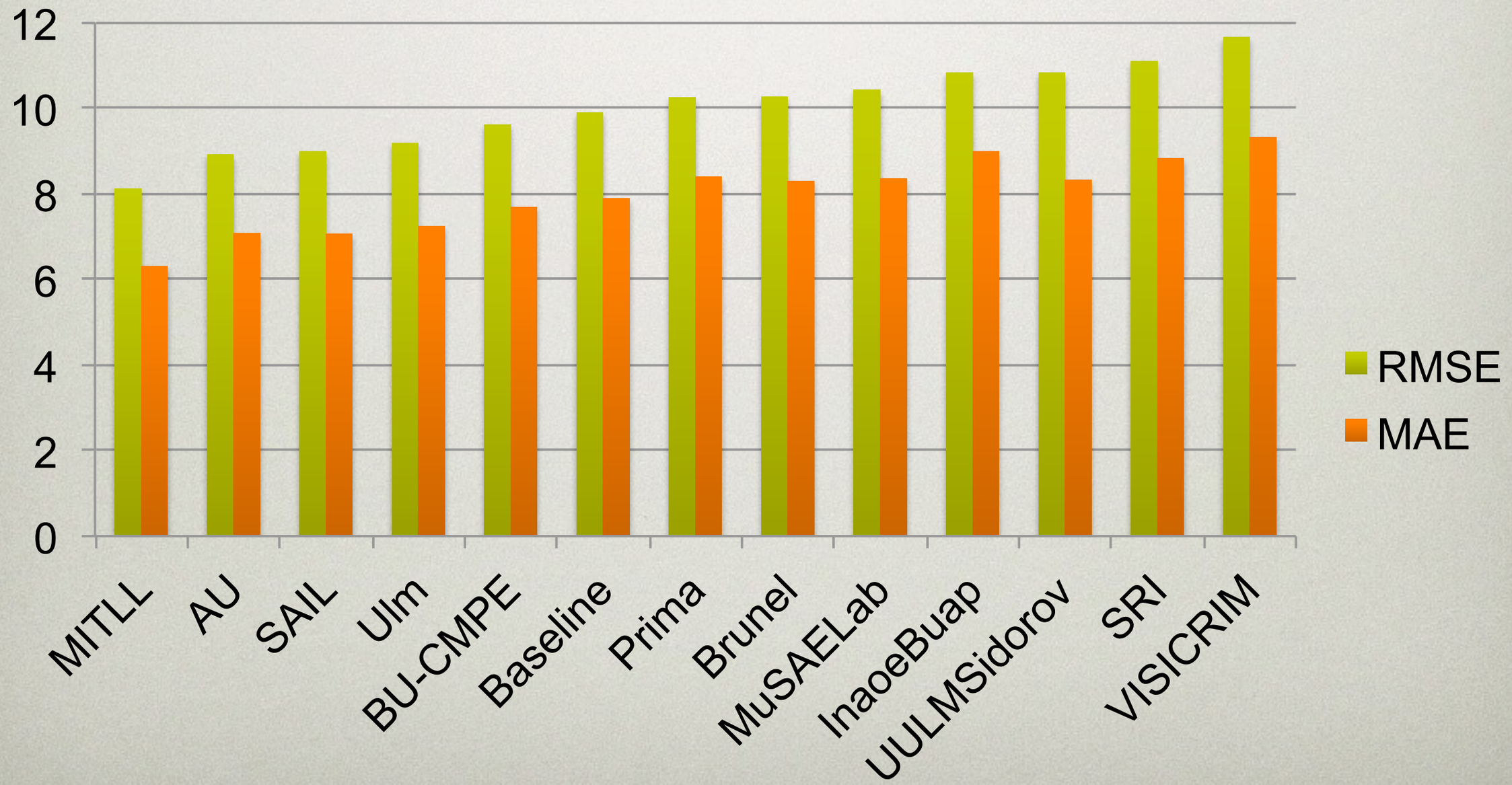
Behaviomedics - The application of automatic analysis and synthesis of affective and social signals to aid objective diagnosis, monitoring, and treatment of medical conditions that alter one's affective and socially expressive behaviour.

Valstar (2014), *'Automatic Behaviour Understanding in Medicine'*, Proc. ACM Int'l Conf. Multimodal Interaction, Roadmapping the Future of Multimodal Interaction (ICMI'14)

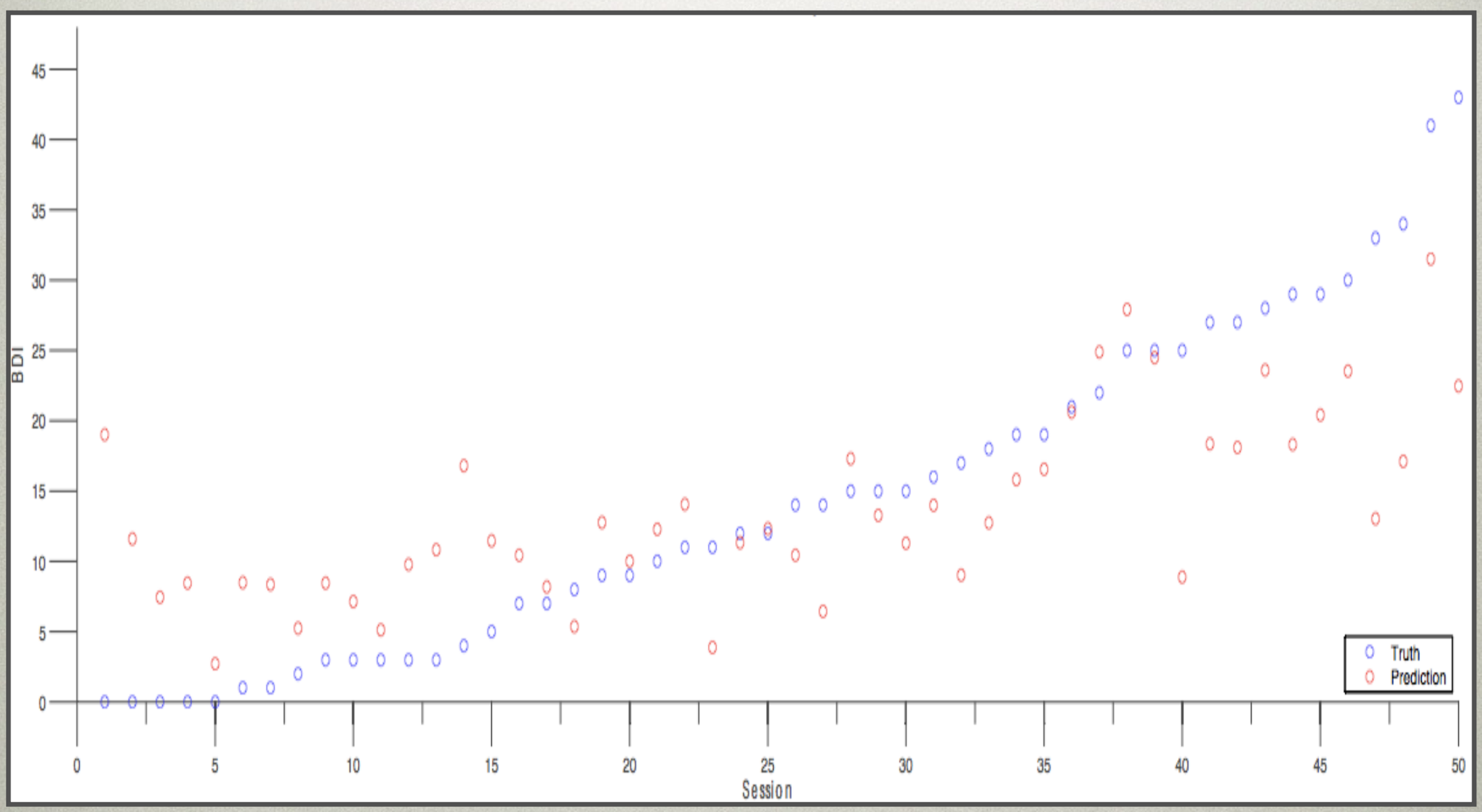
Inference of Meaning



Depression Recognition



Depression Recognition



SPARSE KERNEL METHODS

- Kernel function decisions:

$$y(\hat{\boldsymbol{x}}) = \boldsymbol{k}(\hat{\boldsymbol{x}})^T (\boldsymbol{K} + \lambda \boldsymbol{I}_N)^{-1} \boldsymbol{t}$$

$$k_n(\boldsymbol{x}) = k(\boldsymbol{x}_n, \hat{\boldsymbol{x}})$$

Gram Matrix needs to store NxN training data kernel results

So all training examples must be stored too!

Test pattern computed against all N training examples

- Must be evaluated on all training examples during testing
- Must be evaluated on all pairs of patterns during training
  - Training takes a long time
  - Testing too
  - Memory intensive (both disk/RAM)
- Solution: sparse methods

SUPPORT VECTOR MACHINES

- Sparse (only Support Vectors defining decision boundaries are used in decisions)
- Non-linear by means of the kernel function
- Decision machine - it does *not* provide a posterior probability
  - Many people use Platt's sigmoid fitting to the decision value output y(x) to obtain pseudo-probability, often without understanding that this is not a probability

Vladimir Vapnik
NEC, U Colombia, Royal Holloway London

LINEARLY SEPARABLE SVM

- We will consider for now:
  - 2-class problems
  - linearly separable problems
- Two-class classification definition:

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + b$$

Other definitions:

$\boldsymbol{\phi}(\boldsymbol{x})$ = feature space

$t_n \in \{-1, 1\}$ = labels

$\hat{y} = sign(y(\hat{\boldsymbol{x}}))$ = decision function

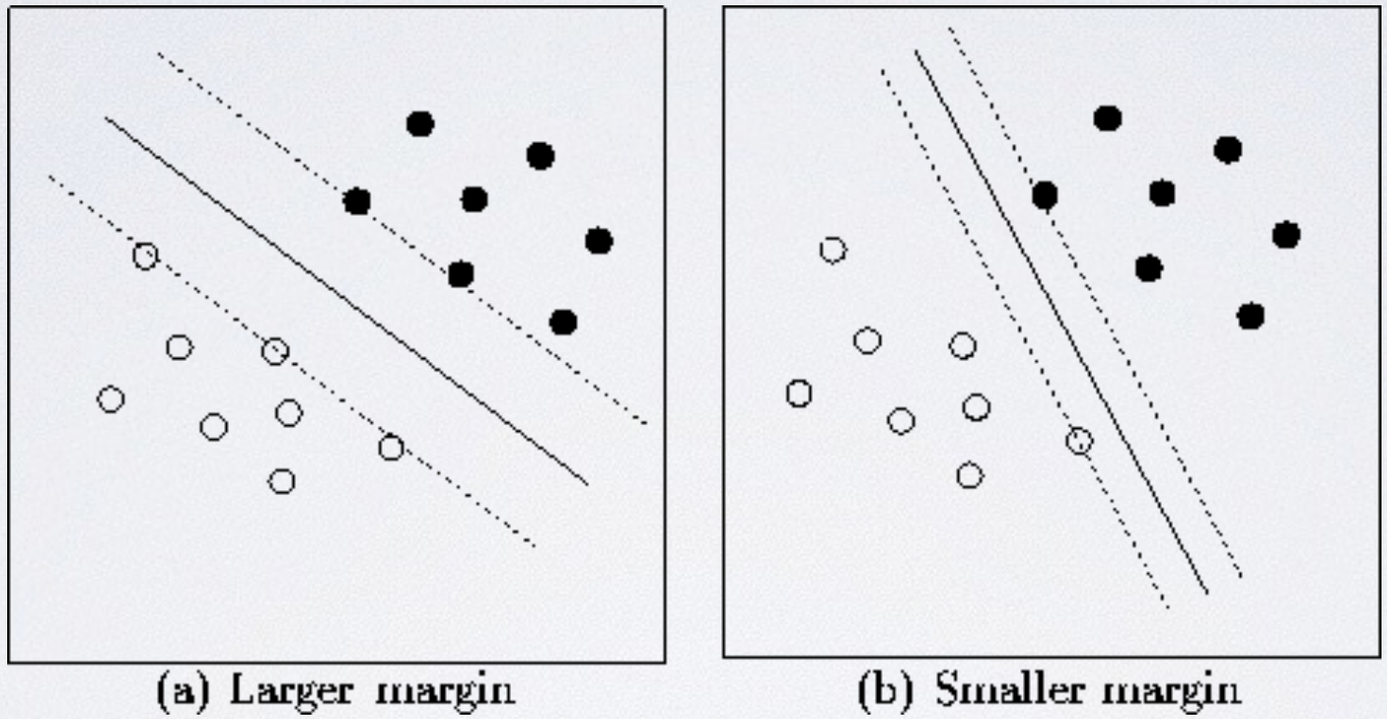
LINEARLY SEPARABLE SVM

- Linear separability implies formally:

$$\exists\{\tilde{\boldsymbol{w}}, \tilde{b}\} \mid \begin{cases} y(\boldsymbol{x}) > 0 & \forall t_n = 1 \\ y(\boldsymbol{x}) < 0 & \forall t_n = -1 \end{cases}$$

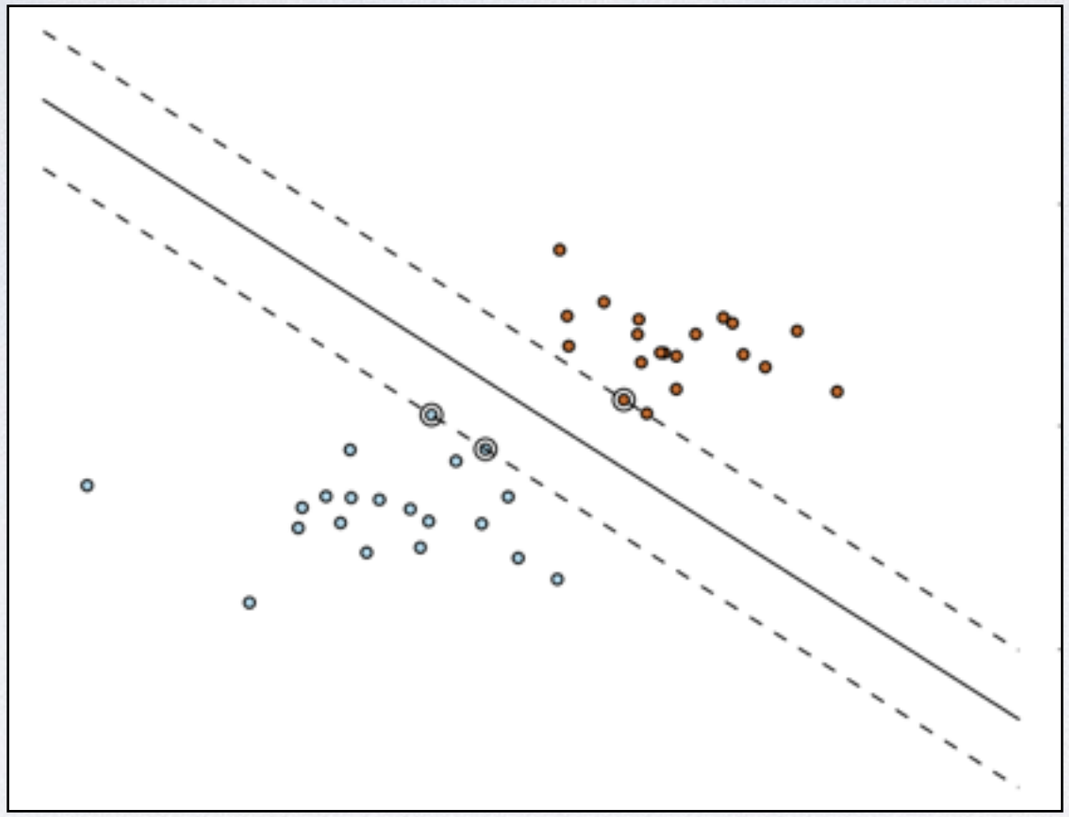
- In general, there are many possible solutions to this problem
- Satisfising solution (e.g. perceptron algorithm): finds a solution, not necessarily the *'best'*
- *Best* is that solution that promises maximum generalisibility

MAXIMUM MARGIN CLASSIFIERS



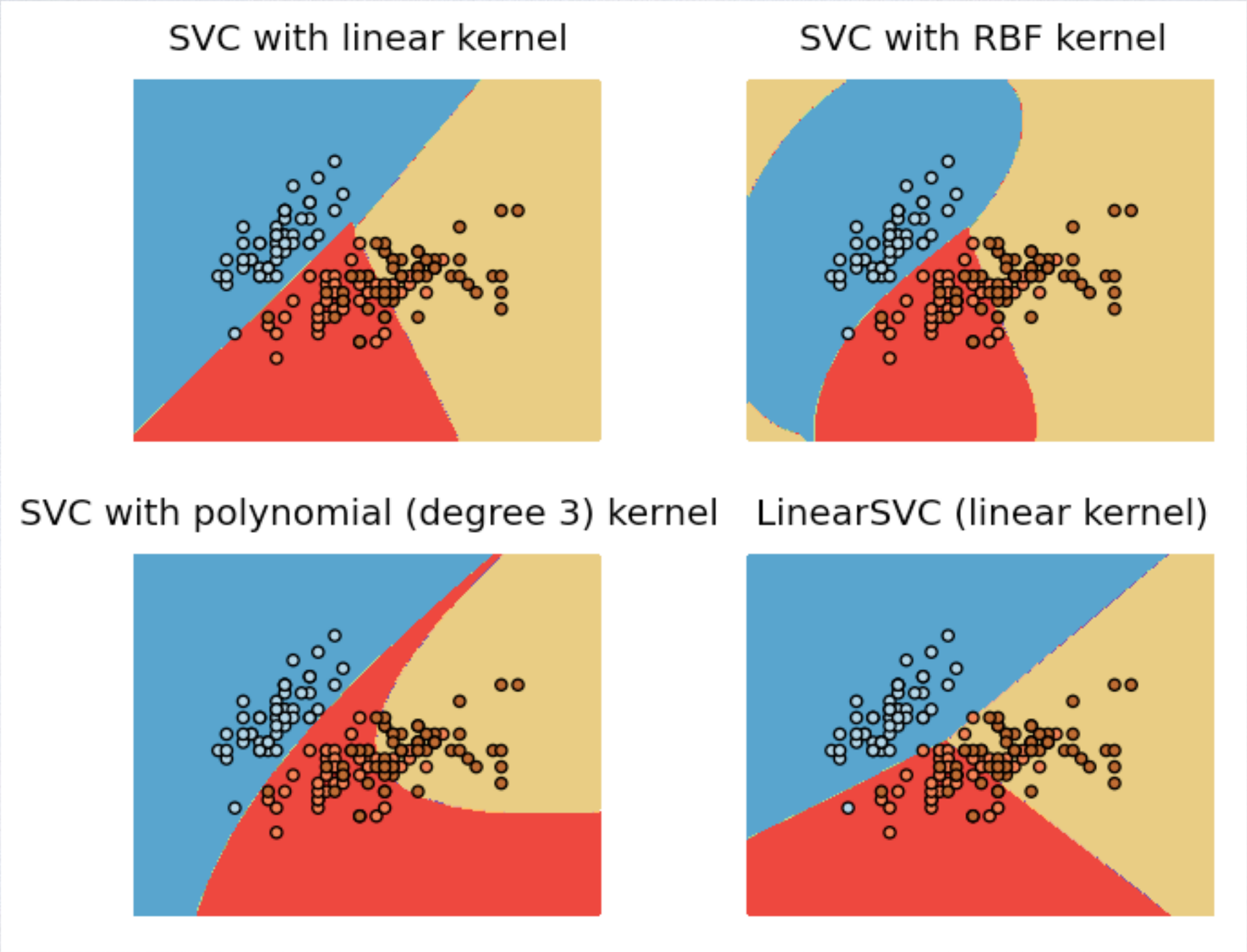
- Margin is defined as the minimum distance between decision boundary and any sample of a class
- SVMs seek a decision boundary that maximises the margin

MAXIMUM MARGIN CLASSIFIERS



- This turns out to be a solution where decision boundary is determined by nearest points only
- Minimal set of points spanning decision boundary sought
- These points are called Support Vectors

SVM KERNELS


SVC with linear kernel
SVC with RBF kernel
SVC with polynomial (degree 3) kernel
LinearSVC (linear kernel)

SVM KERNELS

- It is possible, but difficult, to handcraft domain-specific kernels
- Gram matrix K should be positive semidefinite
- Generally one of the following standard kernels is used:

Linear

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}'$$

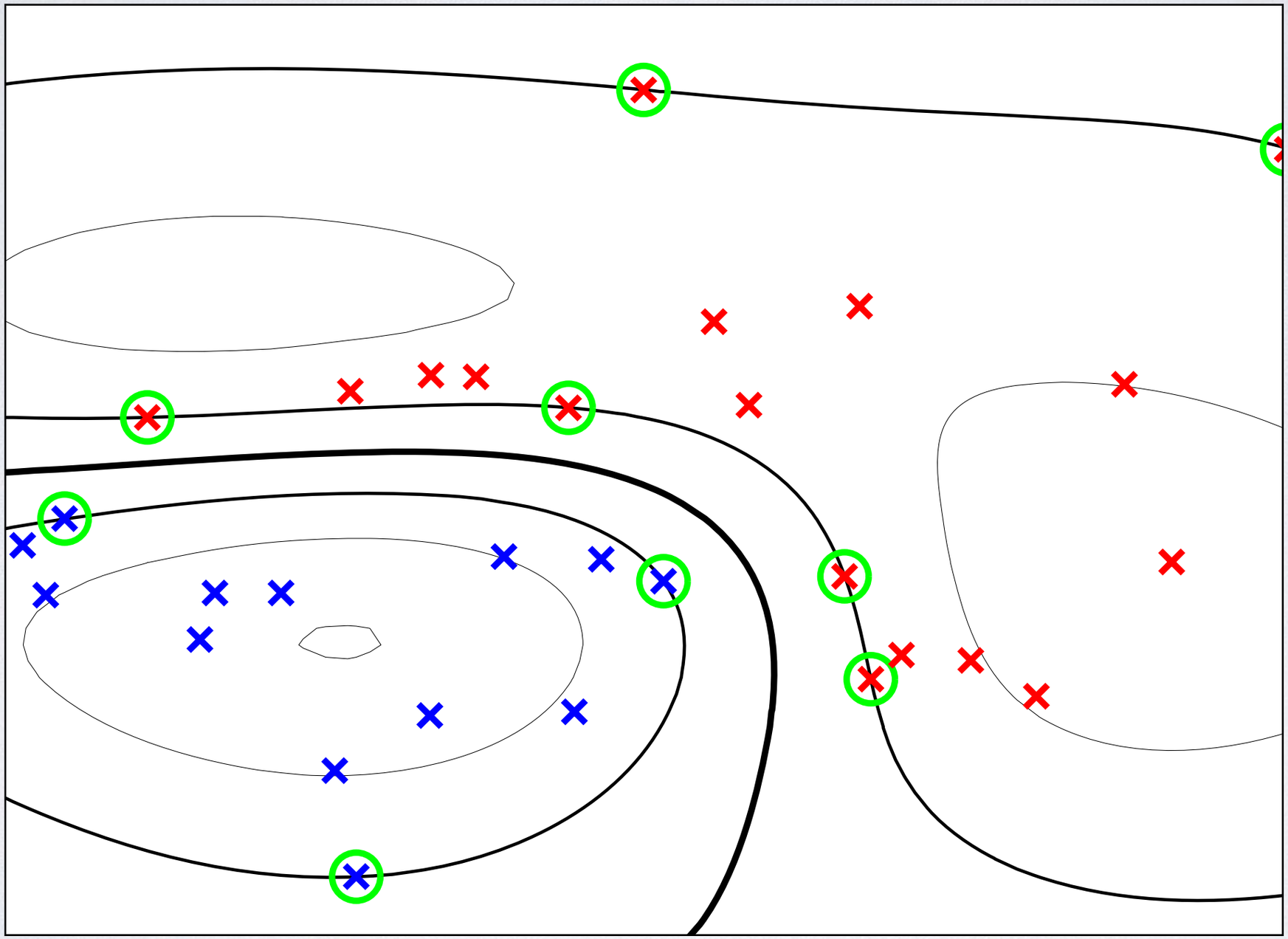
Polynomial

$$k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T \boldsymbol{x}' + c)^M$$

RBF (Gaussian)

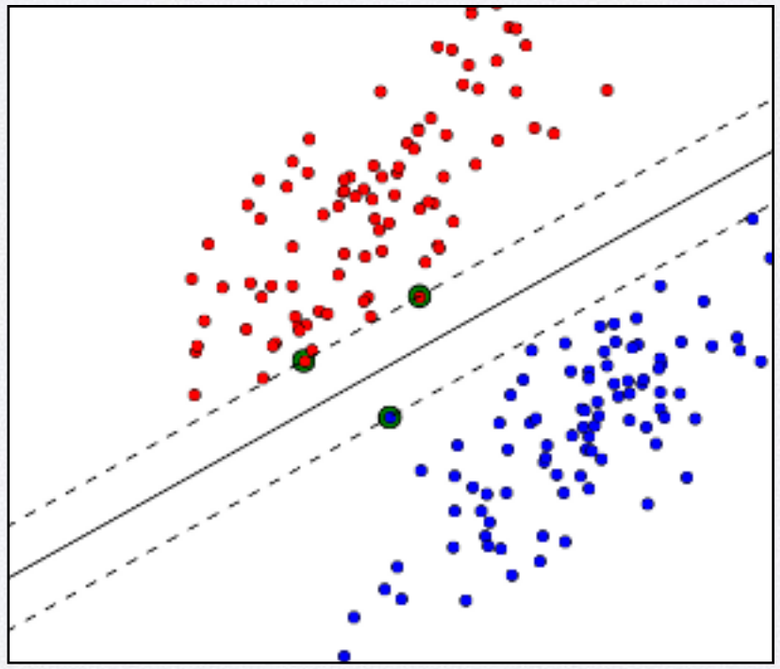
$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(\frac{-\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right)$$

GAUSSIAN KERNEL



SPARSE DESCRIPTIONS

- Small set of SVs means that our solution is now sparse
- Note that the decision boundary is now entirely defined by the points closest to another distribution - i.e. by the *atypical class examples*



SEQUENCE DATA

- Sequence data is data that comes in a particular order
- Opposite of independent, identical distributed (i.i.d.)
- Strong correlation between subsequent elements
  - DNA
  - Time series
    - Facial Expressions
    - Speech Recognition
    - Weather Prediction
  - Action planning

FACIAL EXPRESSIONS

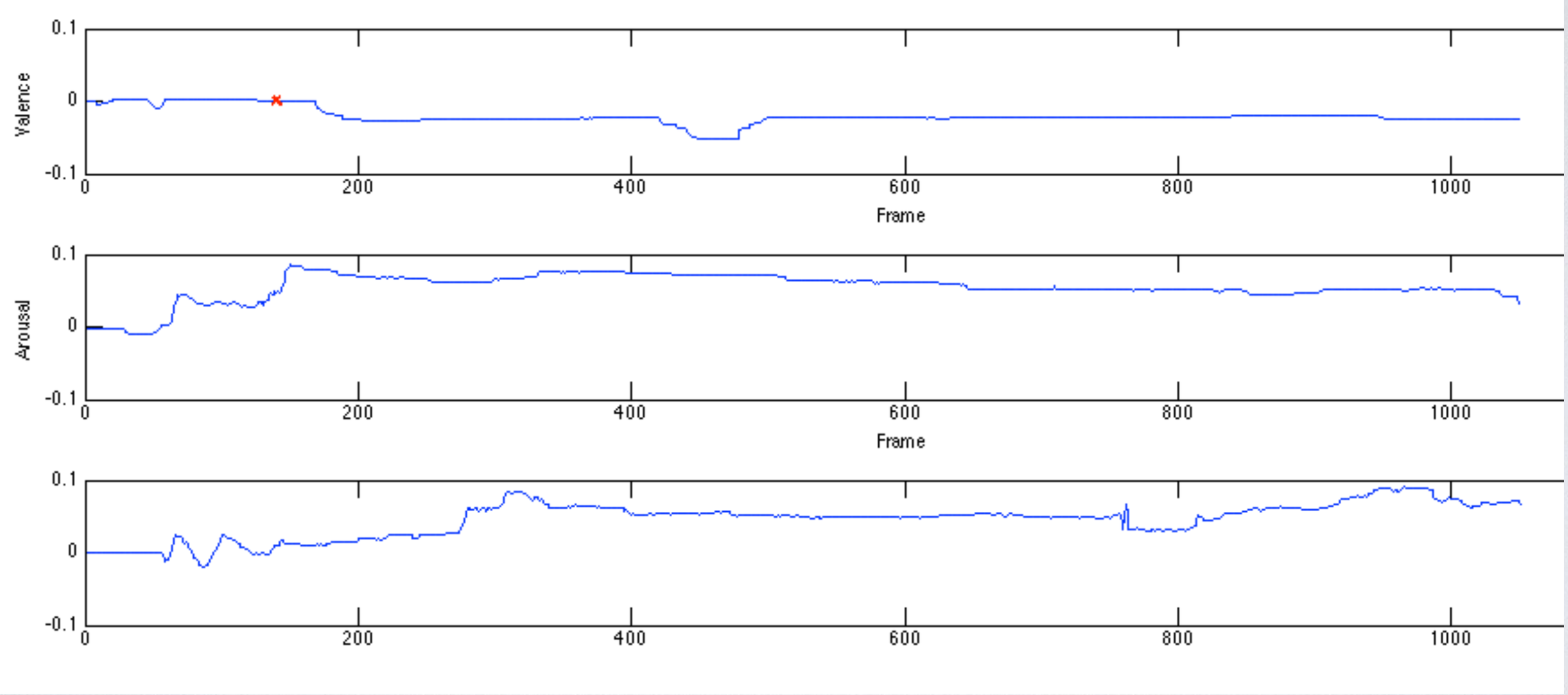
Facial Muscle Activation:

[0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0]

Temporal phases (Neutral-Onset-Apex-Offset):

[0 0 0 0 0 0 0 0 1 1 1 1 1 2 2 2 2 3 3 3 3 3 0 0 0 0]

DIMENSIONAL AFFECT

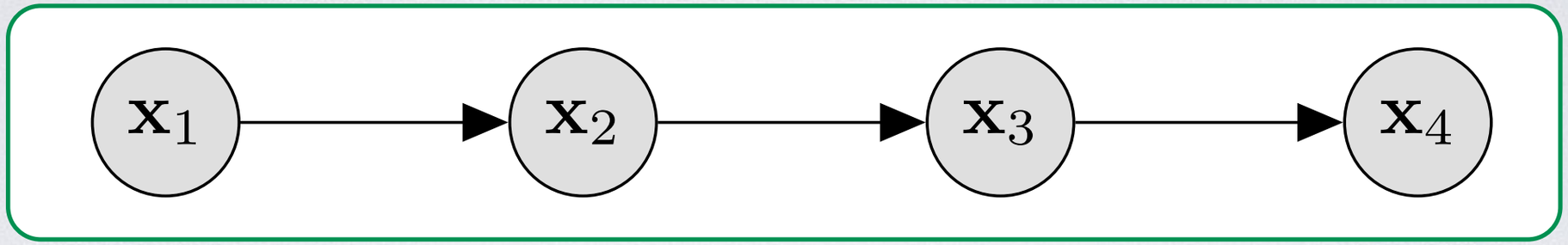


I.I.D. CASE

- Easiest way to treat sequential data is to ignore sequential aspect of subsequent patterns
- Data is then again simply treated independent, identically distributed (i.i.d.)
- Graph model of the observed variables is a set of disconnected nodes

Question: What is $p(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4)$?

1ST ORDER MARKOV MODEL



- When treating sequence data as a 1st order Markov chain, the joint distribution has factors that depend only on the previous observation:

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = p(\boldsymbol{x}_1) \prod_{n=2}^{N} p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1})$$

- d-separation then leads to:

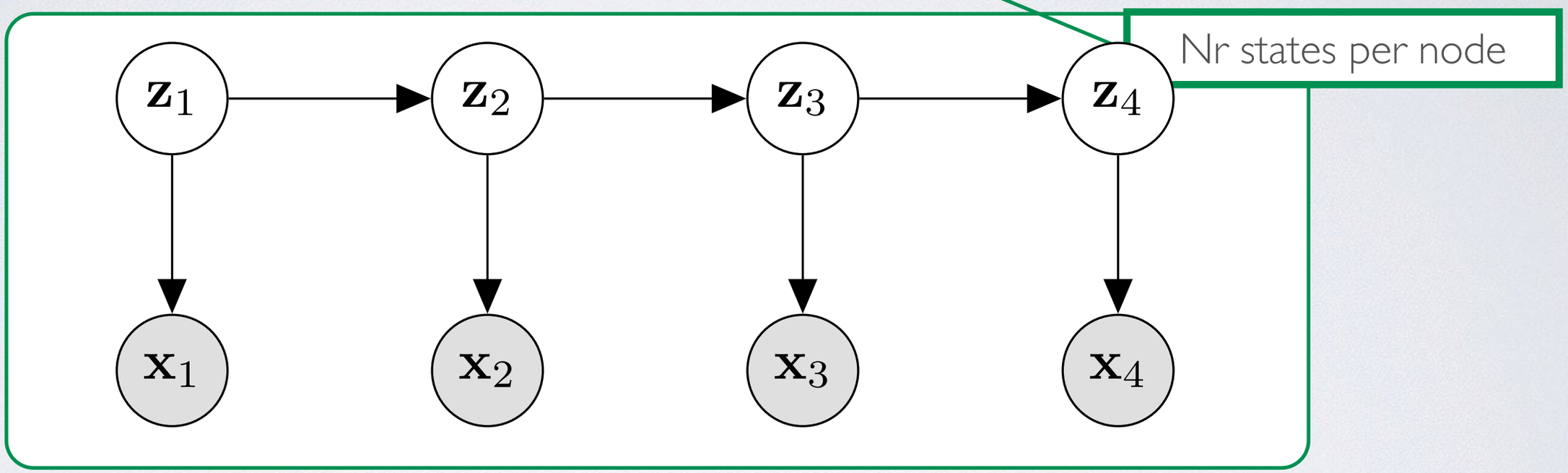
$$p(\boldsymbol{x}_n | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1})$$

STATE SPACE MODEL

- Number of parameters in higher order Markov Models increases exponentially as $K^{M-1}(K-1)$

Markov Model order

Nr states per node

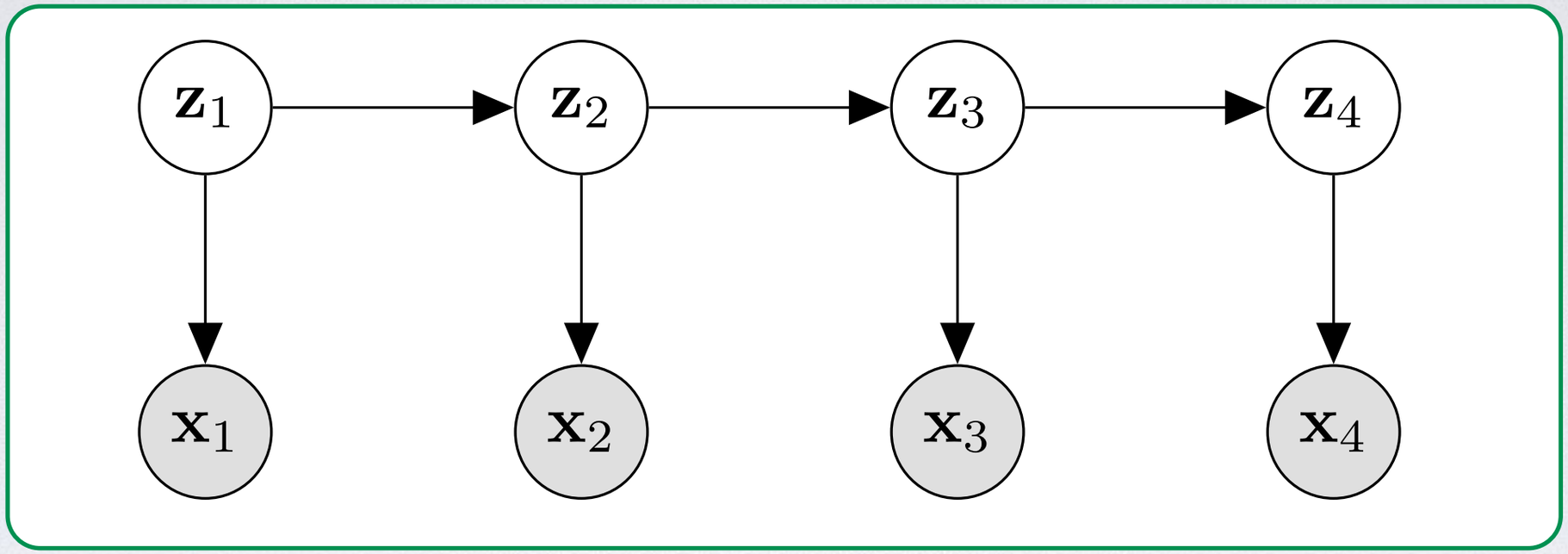


- Solution is a state space model in which a hidden variable is responsible for observations
- Hidden variables are directly sequentially linked, not observables

SS MODEL VARIANTS

- If $\boldsymbol{z}_n$ are discrete variables the state space model is called a Hidden Markov Model (HMM)
- If both $\boldsymbol{z}_n$ and $\boldsymbol{x}_n$ are Gaussian with a linear-Gaussian dependence of the conditional distributions on their parents, we have a Linear Dynamical System (LDS)

MIXTURE DISTRIBUTION



- At each point in time, nodes for $\boldsymbol{x}_n$ are a mixture distribution:

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n)$$

- Mixture distributions were also found in k-Means
- However, mixture components in HMMs are not independent but depend on choice of previous observations

TRANSITION PROBABILITIES

- Latent variables are discrete multinomial variables $\boldsymbol{z}_n$
- Describe which component of mixture is responsible for observation $\boldsymbol{x}_n$
- $\boldsymbol{z}_n$ depends on $\boldsymbol{z}_{n-1}$ only: $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})$
- Variables are discrete, so transition probabilities given as a table:

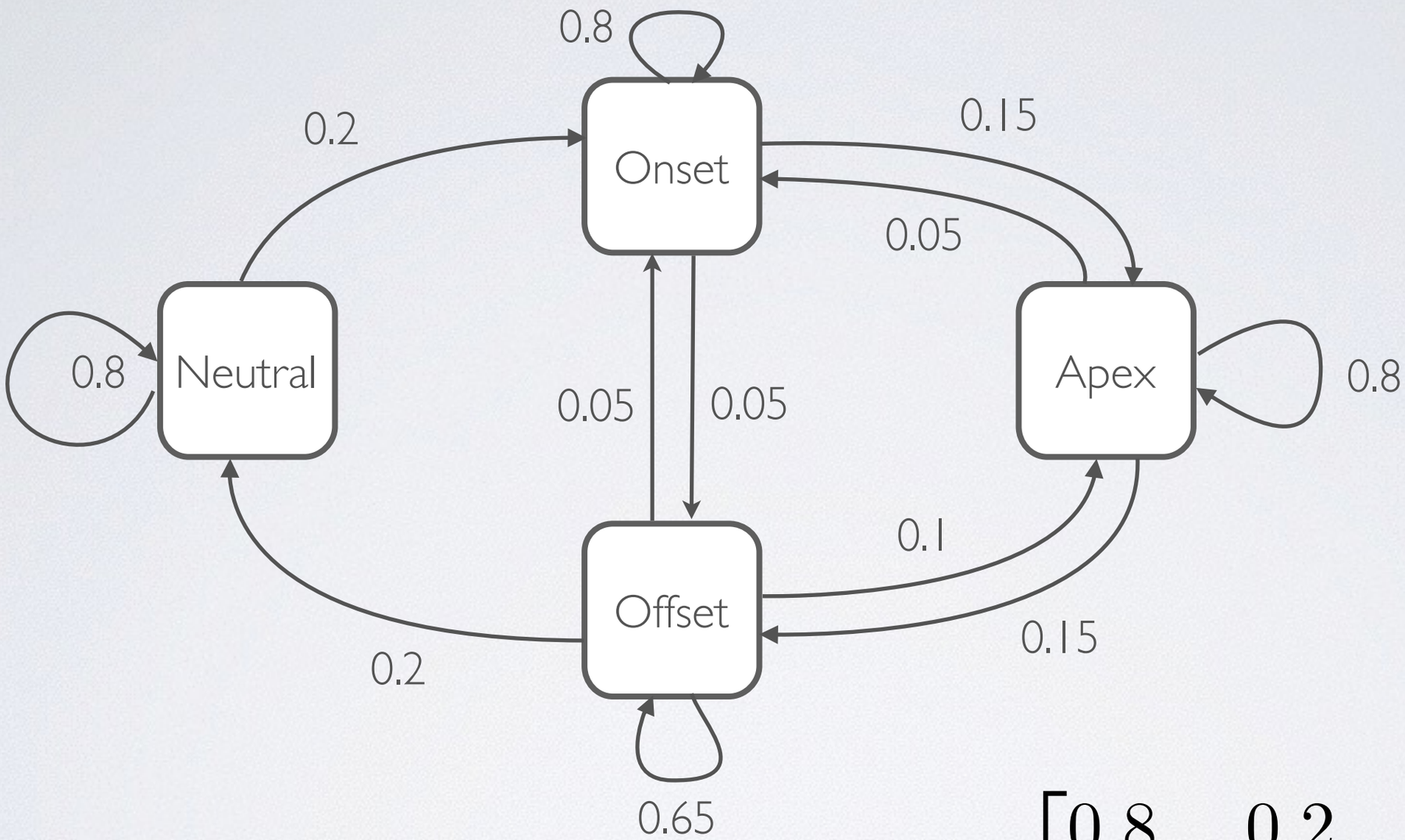
$$A = \begin{bmatrix} 0.8 & 0.2 & 0 & 0 \\ 0 & 0.8 & 0.15 & 0.05 \\ 0 & 0.05 & 0.8 & 0.15 \\ 0.2 & 0.05 & 0.1 & 0.65 \end{bmatrix}$$

Probability of transitioning from state 2 to 3

$$A_{jk} \equiv p(z_{nk} = 1 | z_{n-1,j} = 1)$$

=probability of transitioning from state j to k

EXAMPLE: EXPRESSIONS

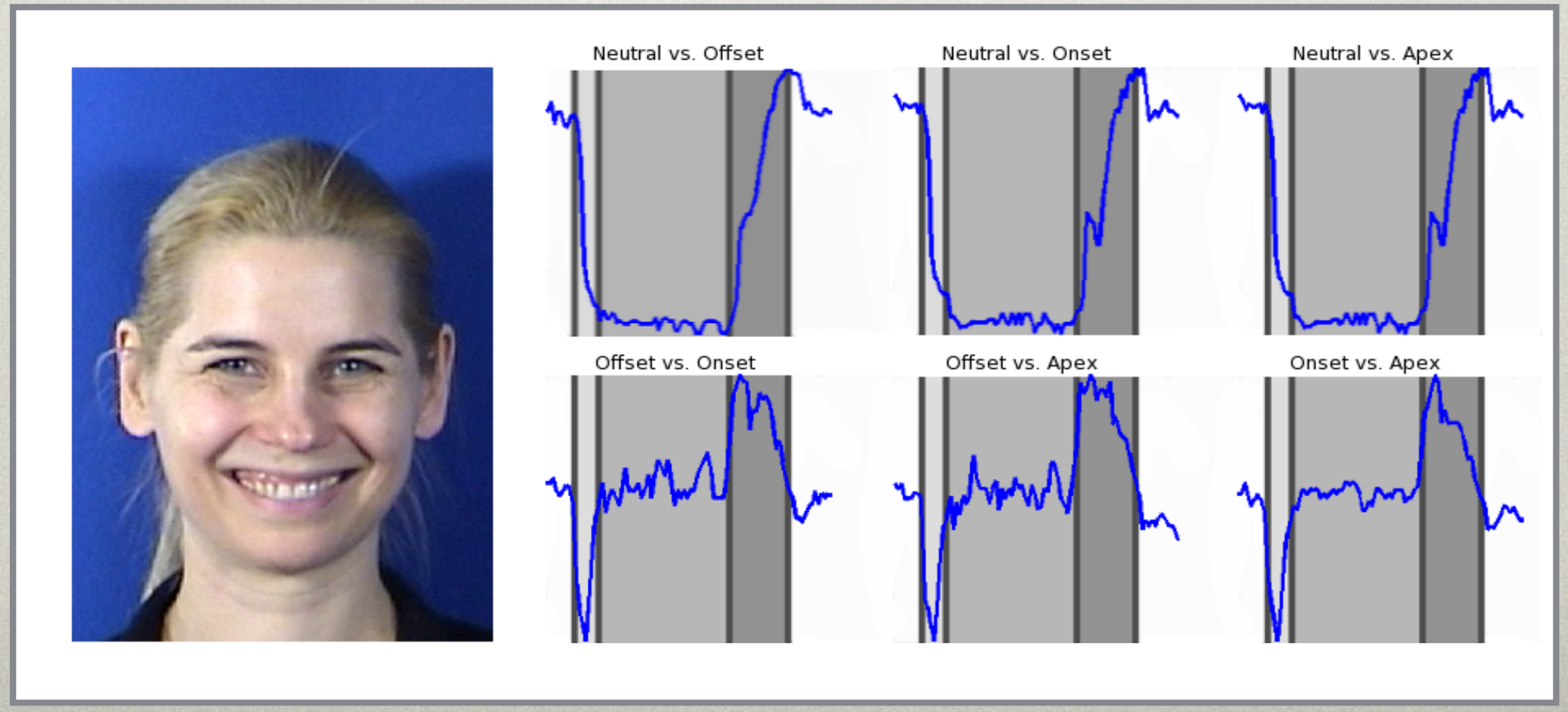


$$A = \begin{bmatrix} 0.8 & 0.2 & 0 & 0 \\ 0 & 0.8 & 0.15 & 0.05 \\ 0 & 0.05 & 0.8 & 0.15 \\ 0.2 & 0.05 & 0.1 & 0.65 \end{bmatrix}$$

Temporal Segment Detection

In temporal segment detection we wish to distinguish between the temporal phases neutral, onset, apex and offset.

We implemented a one-vs-one gentleSvm approach to temporal segment detection based on tracked facial points.



Hybrid SVM-HMM Classification

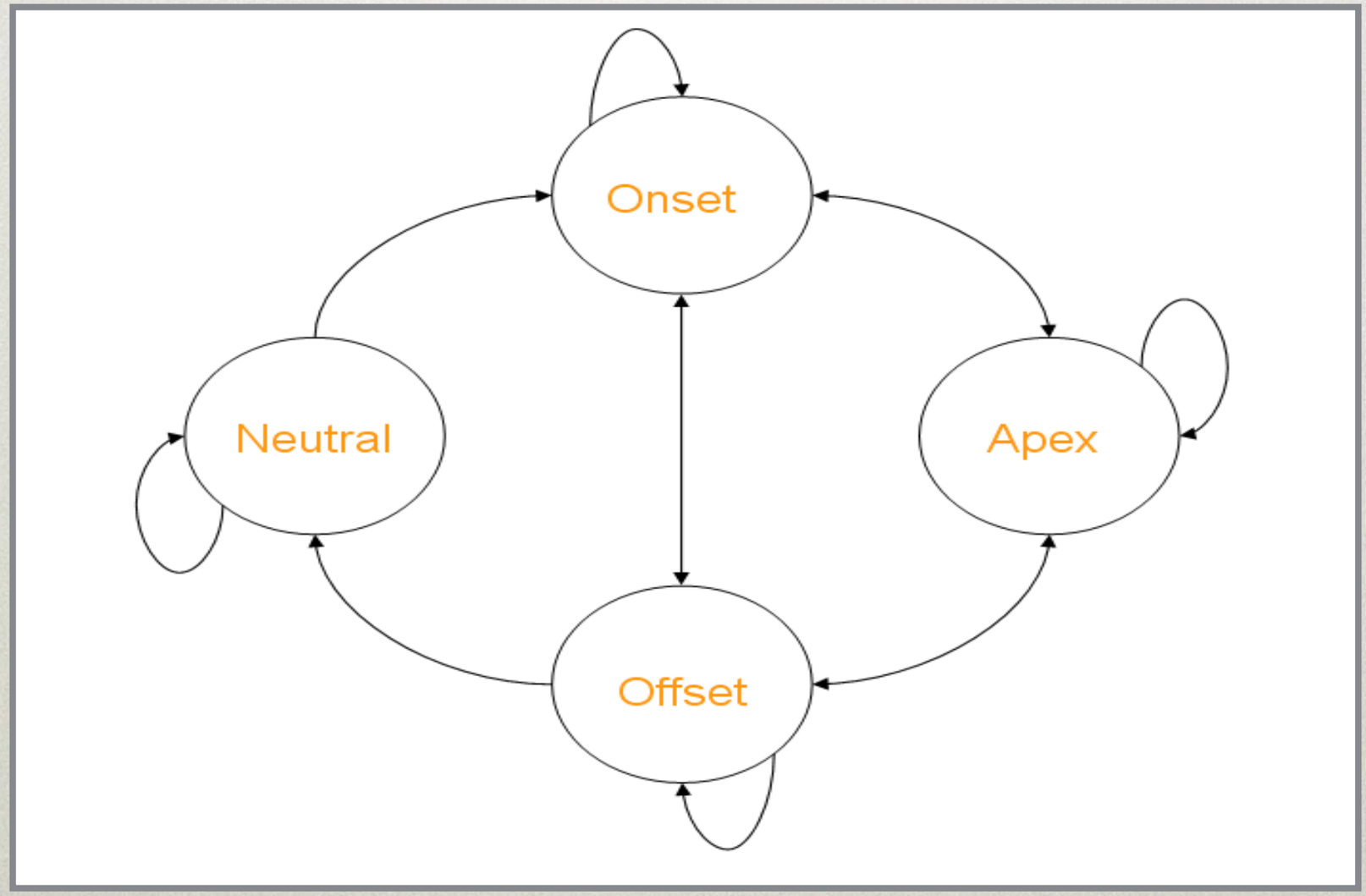
Hidden Markov Models are very well suited to model time and temporal relationships.

SVMs on the other hand discriminate between classes very well, but have no ability to model time at all.

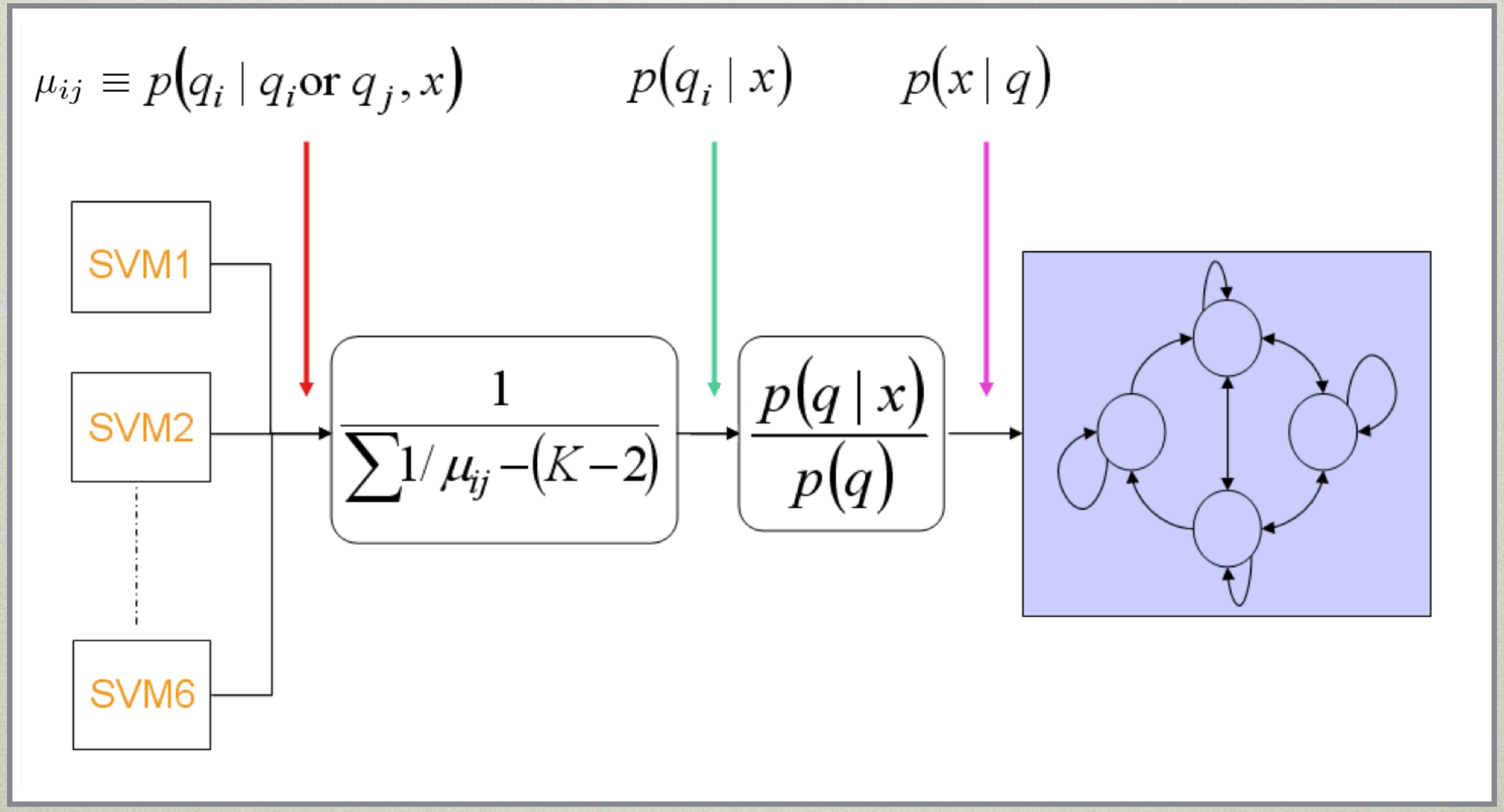
Therefore we propose to combine SVMs and HMMs, using the Platt-modified SVM output as emission probabilities for the HMM.

Valstar & Pantic (2011). *'Fully Automatic Recognition of the Temporal Phases of Facial Actions'*, Systems, Man and Cybernetics - Part B

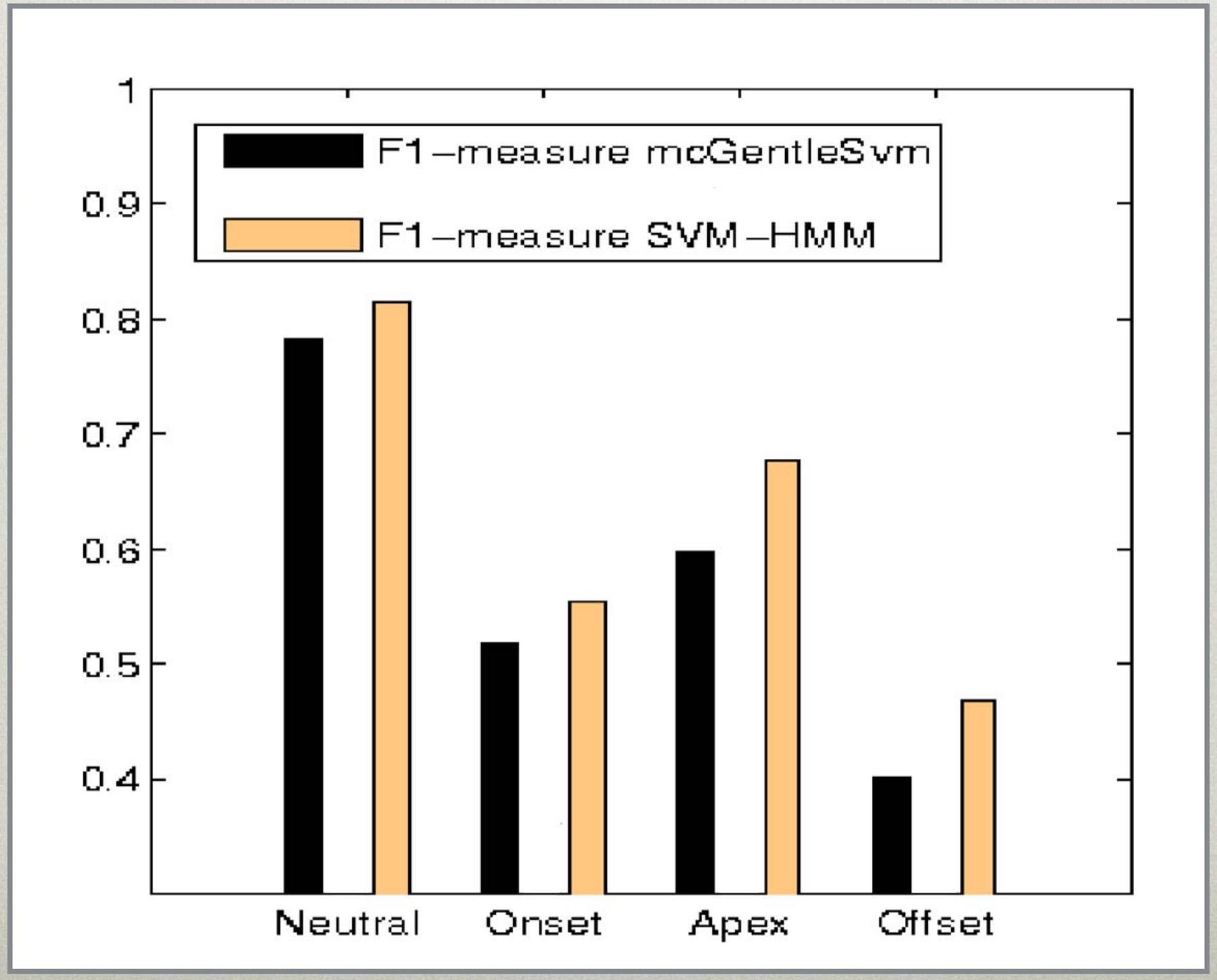
HMM State Transitions



Hybrid SVM-HMM



Results


1
0.9
0.8
0.7
0.6
0.5
0.4
F1–measure mcGentleSvm
F1–measure SVM–HMM
Neutral
Onset
Apex
Offset

Importance of dynamics

Knowledge of the onset-apex-offset phases of AUs has been used to distinguish between posed and spontaneous smiles and brow actions.

Knowing the timing of AUs would allow better facial expression synthesis.

Valstar, Pantic, Ambadar and Cohn (2006). '*Spontaneous vs. posed facial behavior: automatic analysis of brow actions*', ICMI'06
Valstar, Gunes and Pantic (2007). 'How to Distinguish Posed from Spontaneous Smiles using Geometric Features', ICMI'07

Improving Occurrence Detection

Knowledge of the onset-apex-offset phases of AUs can also be used to improve occurrence detection.

False-positive AU detections are often filtered out by the temporal phase modelling as it checks the validity of the sequence of temporal phases.

System	Cl. Rate	Recall	Precision	F1-measure
No temporal analysis	0.876	0.772	0.402	0.521
SVM-HMM	0.943	0.756	0.653	0.692

ACTC and API

- Tools available on a cloud computing service: http://www.cs.nott.ac.uk/ACTC2
- API to be released early 2014

Things to Look Out For

- Normalisation:
  - Head pose (Face registration)
  - Inter-personal differences
  - Lighting
  - Frame rate
- Occlusion

Design Choices

- Feature type:
  - Appearance
  - Geometric
- Classifiers
  - Static
  - Dynamic
- Resource consumption
  - Time
  - Memory

Learning and Evaluation

- Ensure Subject-independence
  - Also during parameter optimisation
- Often little data - use cross-validation
- Usually unbalanced data
  - Use AUC or 2AFC scores
  - NEVER use classification rate

Databases

Publicly available:

- MMI-Facial Expression DB
- Cohn-Kanade(+)
- SEMAINE
- McMaster PAIN database
- DISFA
- BP4D

On To Work!

- Download the code from:

  http://www.cs.nott.ac.uk/~mfv/files/package.zip

- Build a smile detector
- Follow the handout