Karlsruhe Institute of Technology

Recognizing Micro-Expressions & Spontaneous Expressions

Presentation by Matthias Sperber

Institute for Anthropomatics, Computer Vision for Human-Computer Interaction Lab, FIPA Group

Motivation

- Recognize micro-expressions
- Distinguish spontaneous vs. posed expressions
- Useful for..
  - Police & surveillance
  - Doctors
  - Psychology researchers
  - Teachers, business negotiators, …?
  - →In short: **lie detection** using facial expressions

Introduction: Facial expressions

- Facial expressions caused by certain emotions
- 6 basic types of facial expressions (according to Ekman):
  - Disgust
  - Anger
  - Fear
  - Happiness
  - Sadness
  - Surprise

Video: http://www.youtube.com/watch?v=A_XyYxpWIS0

Introduction: Micro-Expressions

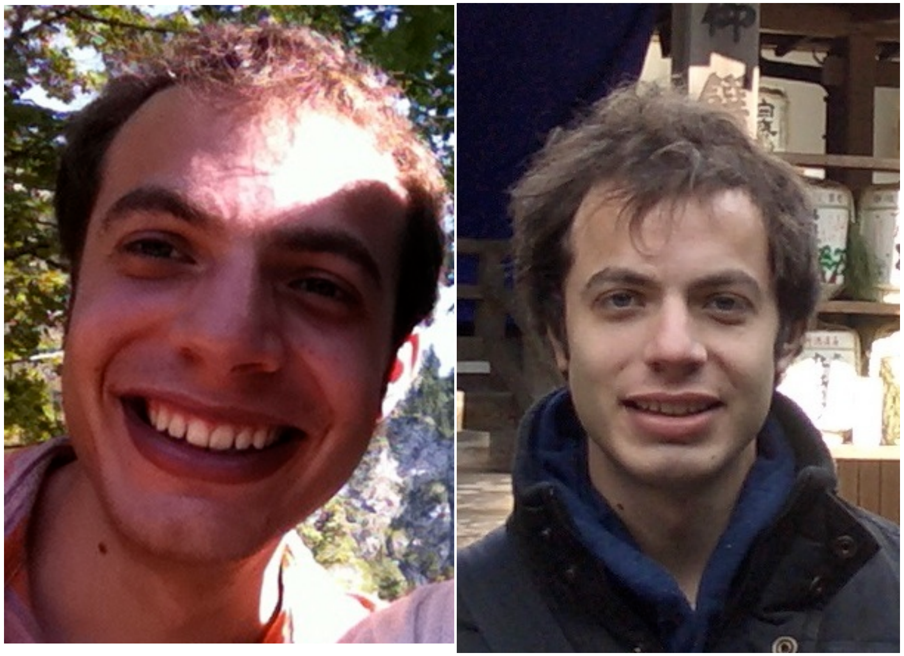
- What are micro-expressions?
  - Very short expressions (1/3 ~ 1/25 seconds)
  - Involuntary (concealed or repressed expressions)
  - Humans are very bad at seeing them
  - Can be learned easily (to some extent)
    - Trained humans: 47% accuracy (untrained: ~25%)
  - Discovery:
    - Hospital patient with secret suicide intentions fools her doctor
    - Video recordings reveals micro-expressions of concealed anguish, quickly covered up by a smile
    - Could be avoided with automatic method to detect micro-expressions!

Introduction: Micro-expressions

Video: http://www.youtube.com/watch?v=4S4xmIkfq6c

Introduction: Posed vs. Spontaneous Expressions

- Recently: research shifting from **posed** expressions to **spontaneous** expressions
- Both differ quite strongly
  - E.g.: Posed smiles: only movement around mouth, real smiles also around eyes



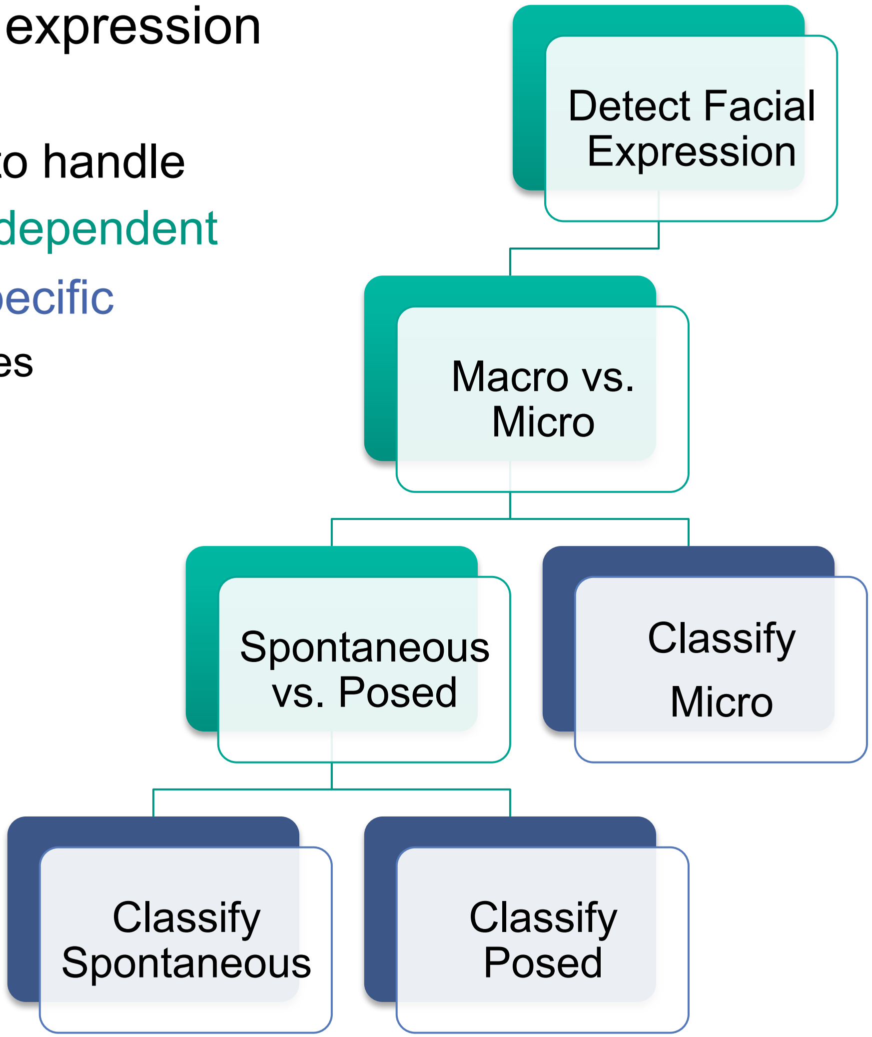
Institute for Anthropomatics, Computer Vision for Human-Computer Interaction Lab, FIPA Group

Introduction

- Goal of this research:
  - Detect and classify micro-expressions
  - Distinguish posed from spontaneous expressions
  - Possibly outperform humans
- Challenges
  - Short duration of micro-expressions (limited # of frames)
  - How to collect realistic data of micro- and spontaneous expressions?
- Approach
  - Complete method including normalization, feature extraction, and classification
  - Use same method for different tasks, train on problem-specific data
  - Use a cascade-structured algorithm to subdivide tasks

A Generic FE Recognition Framework

- Subdivide task of facial expression recognition
  - Each subtask easier to handle
  - 3 tasks application-independent
  - 3 tasks application-specific
    - E.g. 6 basic FE types



Overview of Method

SvP

MicMac

..

select task-specific training data

Collect videos: Visual spectrum or near-infrared, onset to apex

Normalize:
- Spatial normalization: crop using eye-positions, ASM + LWM
- Temporal interpolation / normalization

Extract Features: Spatio-temporal local texture descriptors: LBP-TOP / CLBP-TOP

Classify: LIN | SVM | MKL | Random Forest | Fusion

Institute for Anthropomatics, Computer Vision for Human-Computer Interaction Lab, FIPA Group

Overview of Method

Collect videos: Visual spectrum or near-infrared, onset to apex

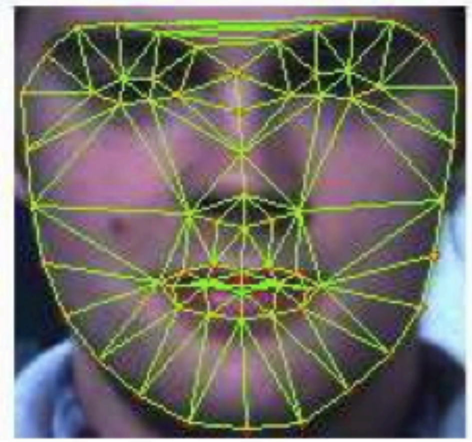
Normalize:
- Spatial normalization: crop using eye-positions, ASM + LWM
- Temporal interpolation / normalization

Extract Features: Spatio-temporal local texture descriptors: LBP-TOP / CLBP-TOP

Classify: LIN | SVM | MKL | Random Forest | Fusion

Active Shape Models (ASM)

- Statistical model for shape of object
  - Shape model (specifies allowable constellations of landmarks)
  - Profile model (templates for each landmark)
- Iteratively:
  - Use template matcher to move around landmarks
  - Adjust shape by calculating similarity transform



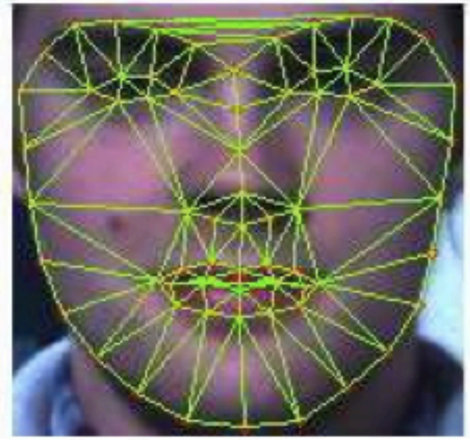
(a)

Local Weighted Means (LWM)

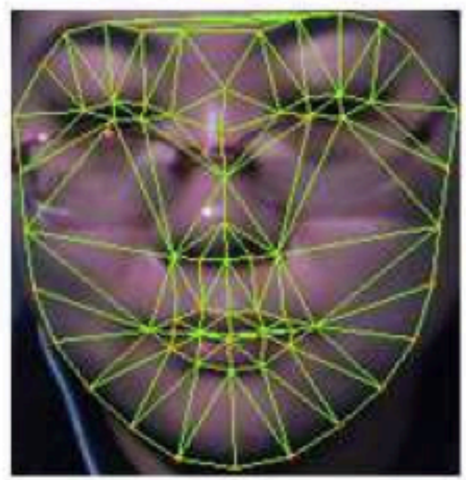
- Using ASM landmarks: compute transformation from first neutral frame to model face:

$$f:(x,y)\mapsto(x',y')$$

- Apply same transformation to all frames
- Effect: spatial normalization
  - Certain facial features always lie in same area on image
  - Muscle movement not affected



(a)



(b)

(c)

Overview of Method

Karlsruhe Institute of Technology

Collect videos: Visual spectrum or near-infrared, onset to apex

Normalize:
- Spatial normalization: crop using eye-positions, ASM + LWM
- Temporal interpolation / normalization

Extract Features: Spatio-temporal local texture descriptors: LBP-TOP / CLBP-TOP

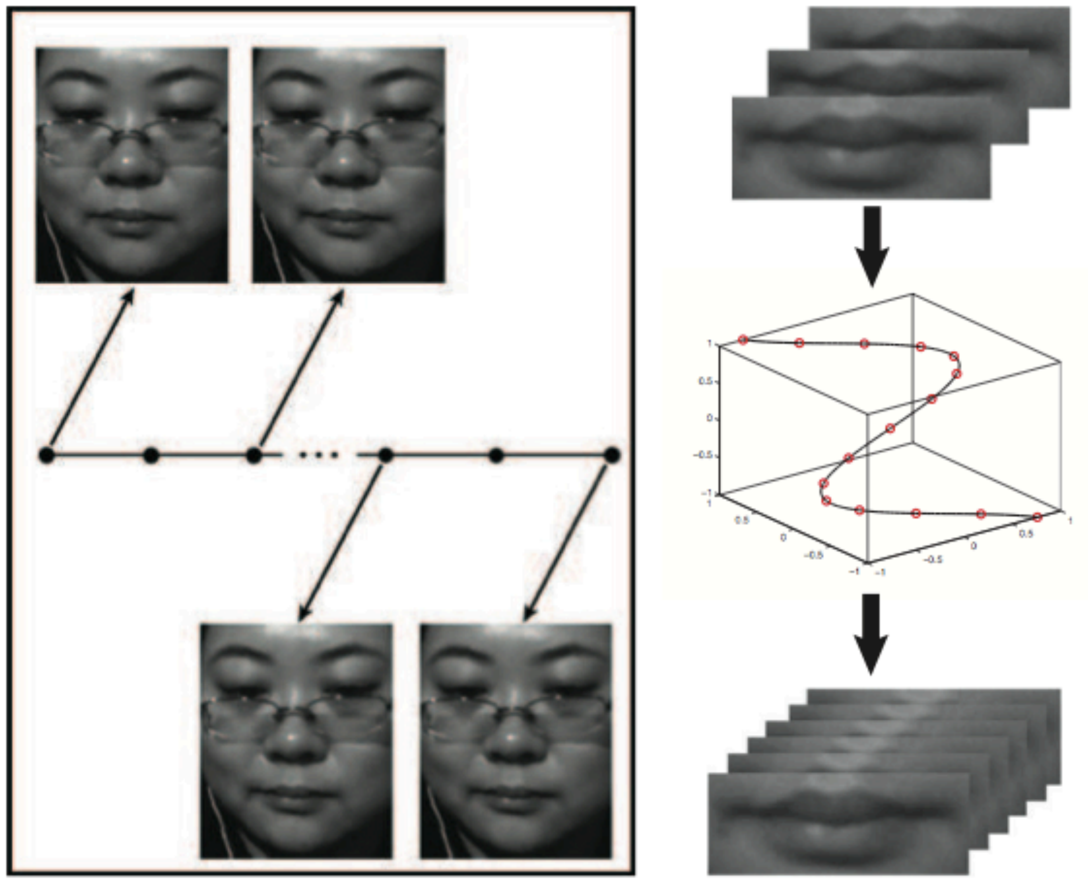
Classify: LIN | SVM | MKL | Random Forest | Fusion

Temporal Interpolation Method (TIM)

- Problem 1: micro-expressions are very short
  - E.g. 25fps: 1/25 sec …1/3 sec ~ 1…8 frames
  - At least 7 frames needed for LBP-TOP feature extraction
- Problem 2: Low # frames → histograms statistically unstable
- Solution: Interpolate between frames, then sample as wished
  - May lead to both a larger or smaller number of frames
  - Generic method to interpolate any kind of feature vectors

Temporal Interpolation Method (TIM)

- Basic idea:
  - Interpolate: Map feature vectors to continuous curve
  - Invert function
  - Create feature vectors from arbitrary position on curve (→ sampling)
- Values for # frames: 10, 20, 25 and 30 frames / video



(a) (b)

Overview of Method

Karlsruhe Institute of Technology

Collect videos
- Visual spectrum or near-infrared, onset to apex

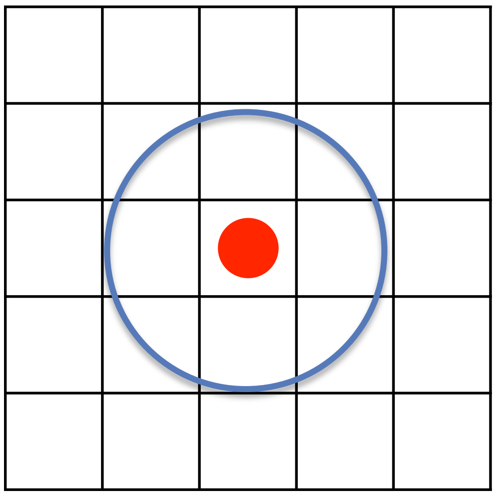
Normalize
- Spatial normalization: crop using eye-positions, ASM + LWM
- Temporal interpolation / normalization

Extract Features
- Spatio-temporal local texture descriptors: LBP-TOP / CLBP-TOP

Classify
- LIN | SVM | MKL | Random Forest | Fusion

Local Binary Pattern (LBP)

- Texture descriptor (2D)
- robust against changes in grey level (illumination), rotation, translation
- Describe “self-similarity” of a texture



Circle around center

100	142	143
131	130	125
119	118	155

Grey level values

-30	+12	+13
1		-5
-11	-12	+25

Differences

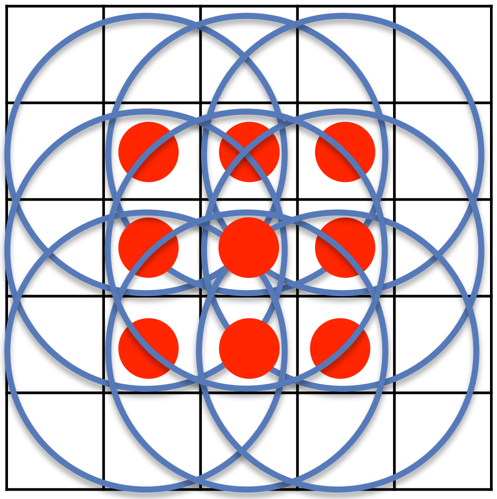
0	1	1
1		0
0	0	1

Threshold

01101001

Unique bit sequence

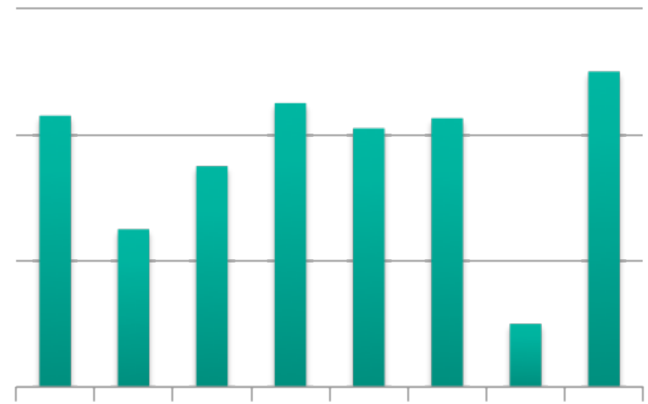
Local Binary Pattern (LBP)



Use all pixels..

..to determine
bit sequences..

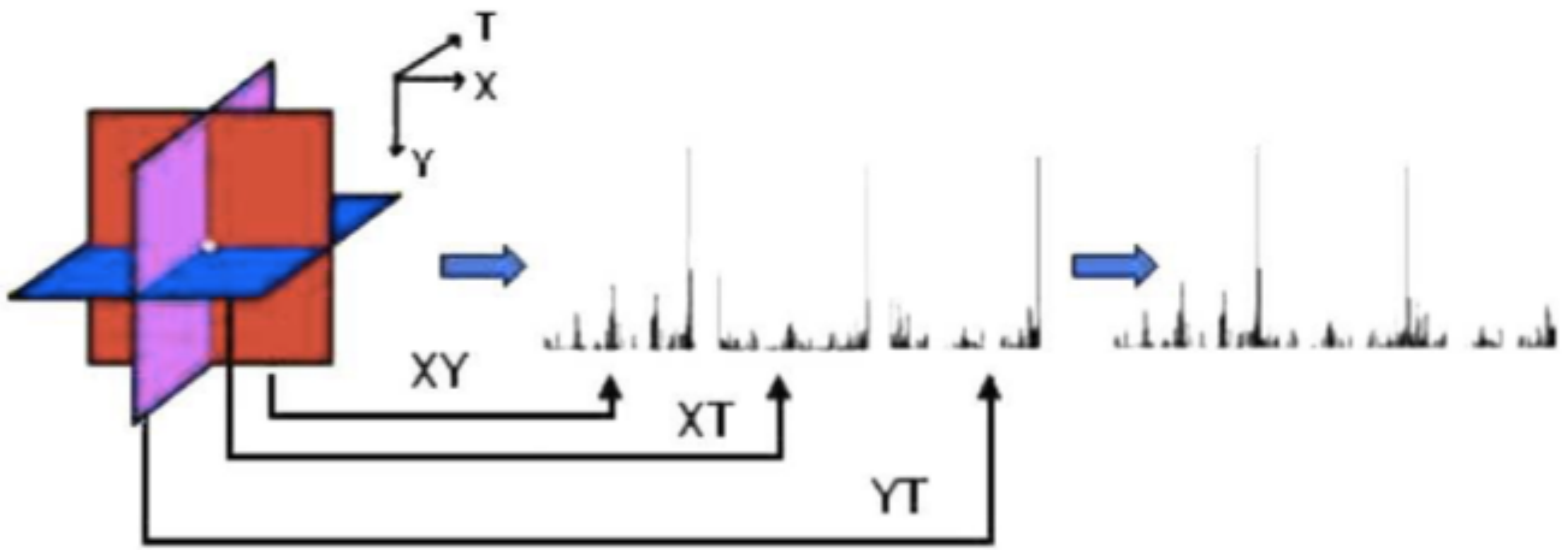
11101001
11010011
10111010
…
11101001



.. and calculate histogram

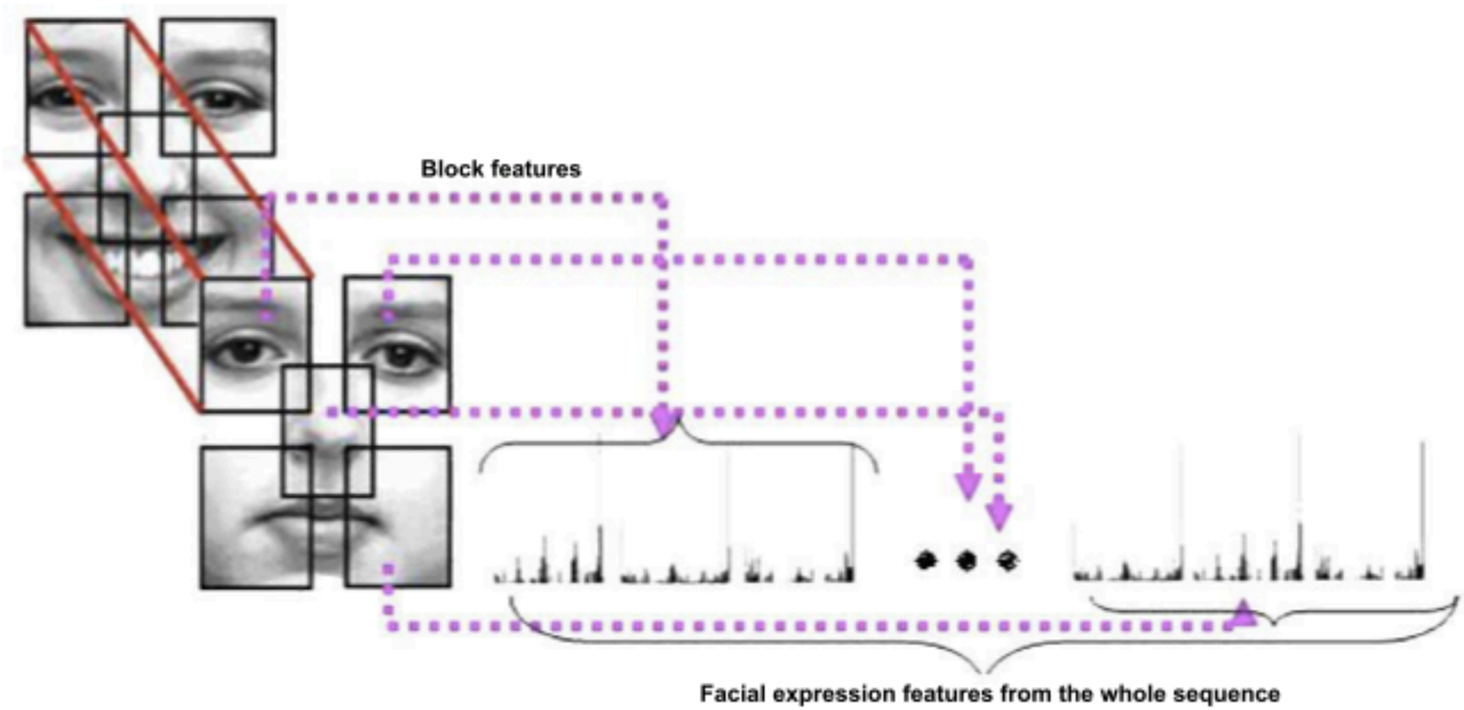
LBP on 3 Orthogonal Planes (LBP-TOP)

- Extend into temporal domain (i.e., make texture descriptor **dynamic**)
- View video as 3D space
- For each pixel, use circle on 3 planes (XY, XT, YT) in the same fashion
- Concatenate histograms



LBP on 3 Orthogonal Planes (LBP-TOP)

- To keep local and temporal context:
  - Divide into blocks (e.g. 8×8×1, 5×5×1, 8×8×2, 5×5×2, 8x8x3 etc.)
  - Use each block (=dynamic texture) to calculate LBP-TOP histograms
  - Concatenate histograms



Completed Local Binary Patterns (CLBP-TOP)

Karlsruhe Institute of Technology

100	142	143
131	130	125
119	118	155

-30	+12	+13
1		-5
-11	-12	+25

(1) Differences

Threshold
(> or < 0?)

30	12	13
1		5
11	12	25

Threshold
(> or < mean?)

(2) Magnitude of differences

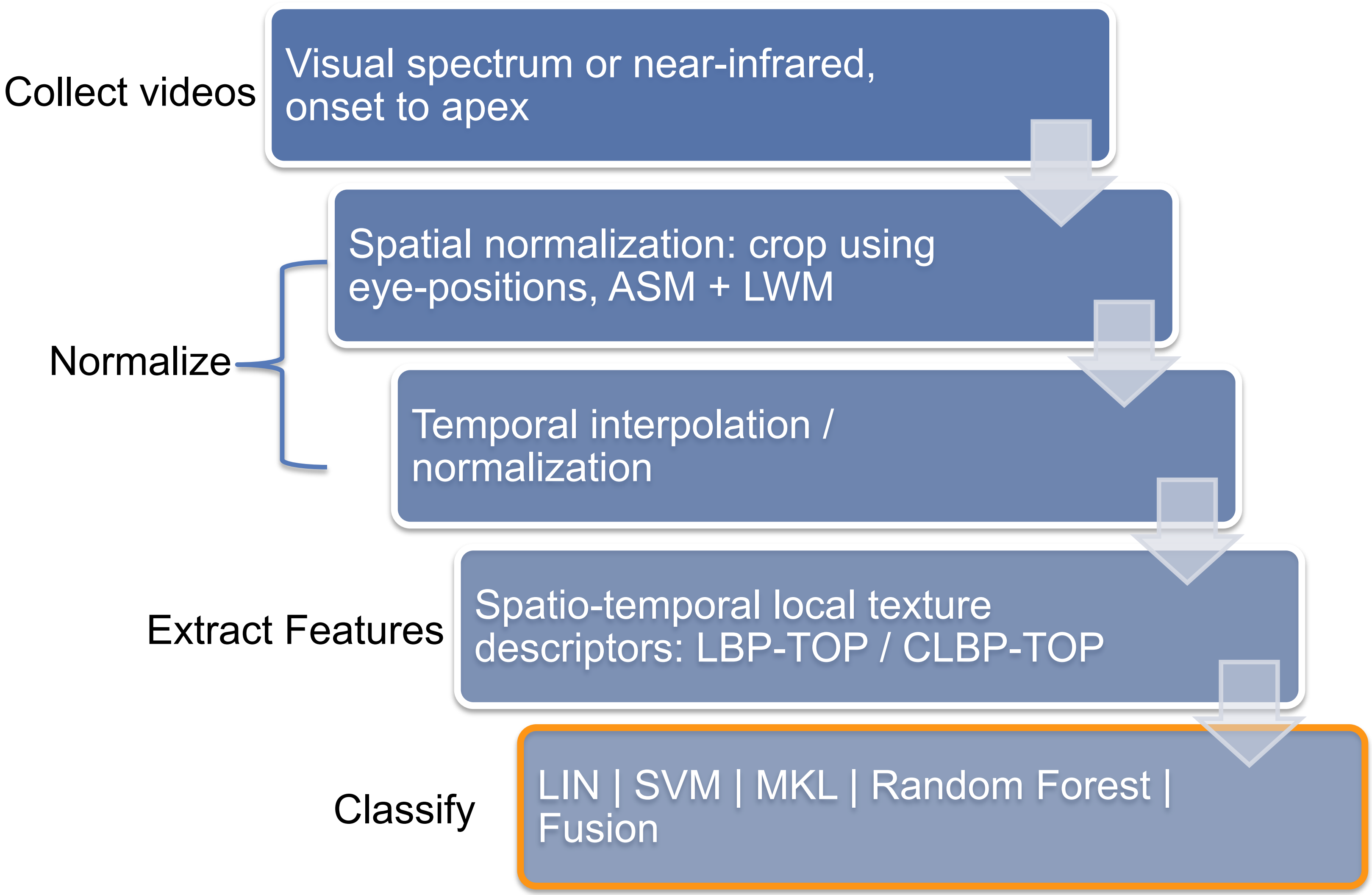
(3) Center grey level

Concatenate
bit sequences

01101001 |
00010001 |
14

- Pro: more discrimination power, better results
- Con: less robust to changes in illumination
(→works better on near-infrared)

Overview of Method



Classification

- Linear Support Vector Machine (SVM)
- SVM with polynomial kernel
- Multi-Kernel Learning (MKL)
  - Combine different kernels
- Random Forests (RF)
  - Combine randomized decision trees
- Fusion
  - Majority voting between linear, SVM, RF

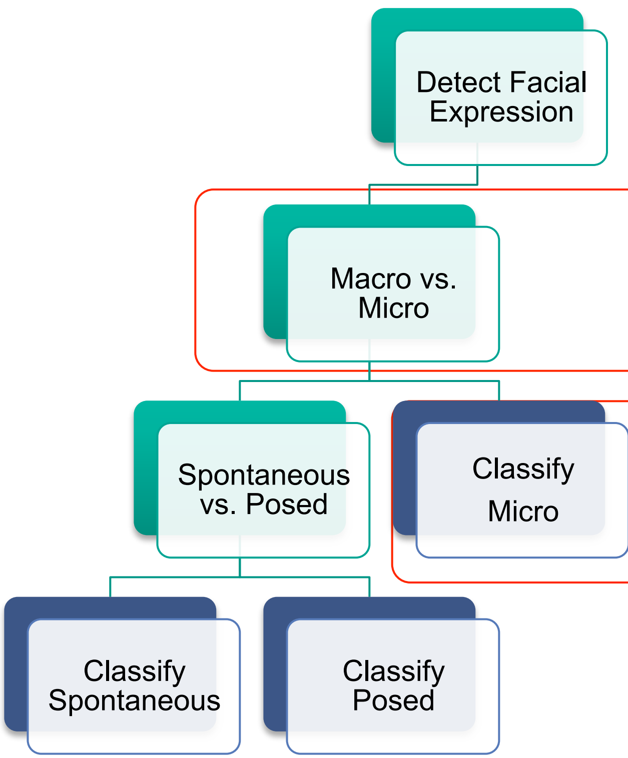
Evaluation

- Experiment Micro-Expressions:
  - Subjects watch videos that are supposed to induce 1 of basic 6 emotions
  - Carefully watch clips, but **suppress facial expressions**
  - Experimenters try to tell emotion from watching face
  - Threat of punishment if successful in telling
  - After experiment: subjects report true emotions

Institute for Anthropomatics, Computer Vision for Human-Computer Interaction Lab, FIPA Group

Evaluation

- Results:
  - Some results better than human recognition
  - Random Forest & MKL had best results (depending on task)
  - TIM10 yields large performance-boost, but more than 10 samples mostly did not lead to improvement (sometimes even scored below TIM10)



RF, 100fps: 74.3%
RF, 25fps: 78.9%

MKL, 100fps: 74.3%
MKL, 25fps: 64.9%

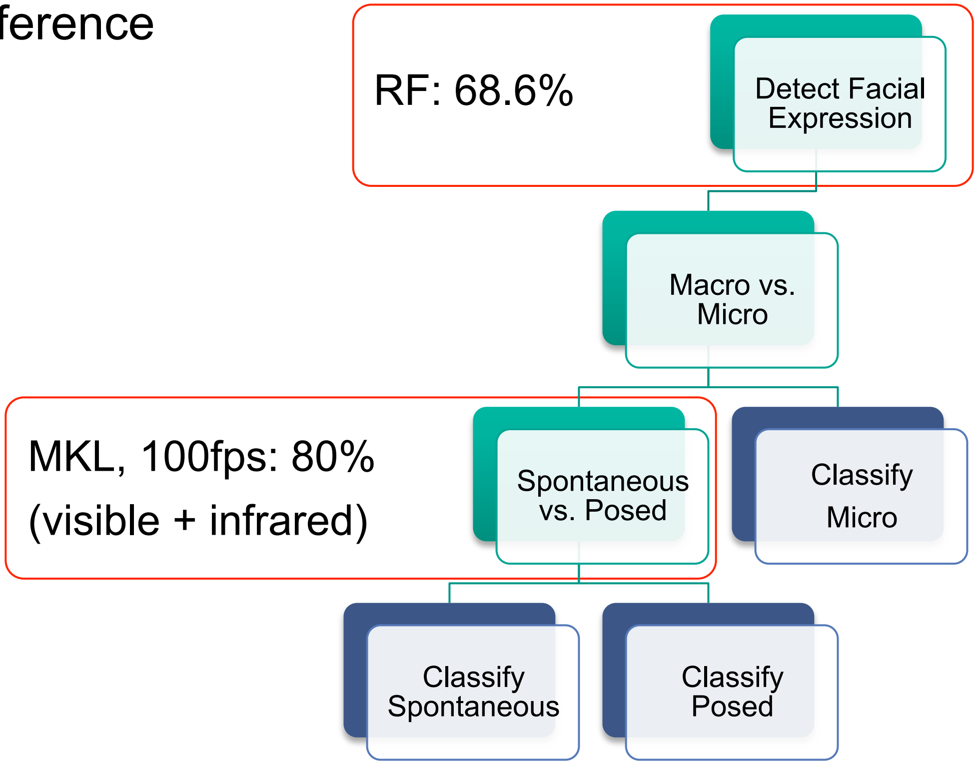
Institute for Anthropomatics, Computer Vision for Human-Computer Interaction Lab, FIPA Group

Evaluation

- Experiment Spontaneous vs Posed:
  - Subjects watched movie clips inducing the 6 basic emotions
  - This time: no suppression
  - Labeled according to subjects' reported emotions
  - Afterwards, subjects were asked to pose each emotion twice
  - Videos recorded with both visual-spectrum- and near-infrared-camera

Evaluation

- Results:
  - Near-infrared > visual-spectrum
  - CLBP-TOP > LBP-TOP for near-infrared data (up to 20% better); visual-spectrum data: difference much smaller



Summary

- Main contributions
  - Extend FE research to new tasks
  - Realistic but small corpora
  - FE recognition cascade
  - Method that can solve all subtasks in cascade

Discussion & Future Work

- Discussion
  - First experiments that use somewhat realistic data!
  - Used mostly existing methods, extended to new contexts
  - Dataset too small → results not very significant
- Future work
  - Make corpora larger & more realistic
  - Ekman: For lie detection, no single one good cue (micro-expressions etc.) exists
    → Several cues must be combined:
    - Classify micro-expressions (short but full involuntary expressions)
    - Classify "subtle expressions" (longer but only expression-fragments)
    - Body language (habits when nervous, …)
    - Voice characteristics (pitch, speed, ..)

Thank you for your attention

- Questions?!

Appendix: Local Weighted Means (LWM)

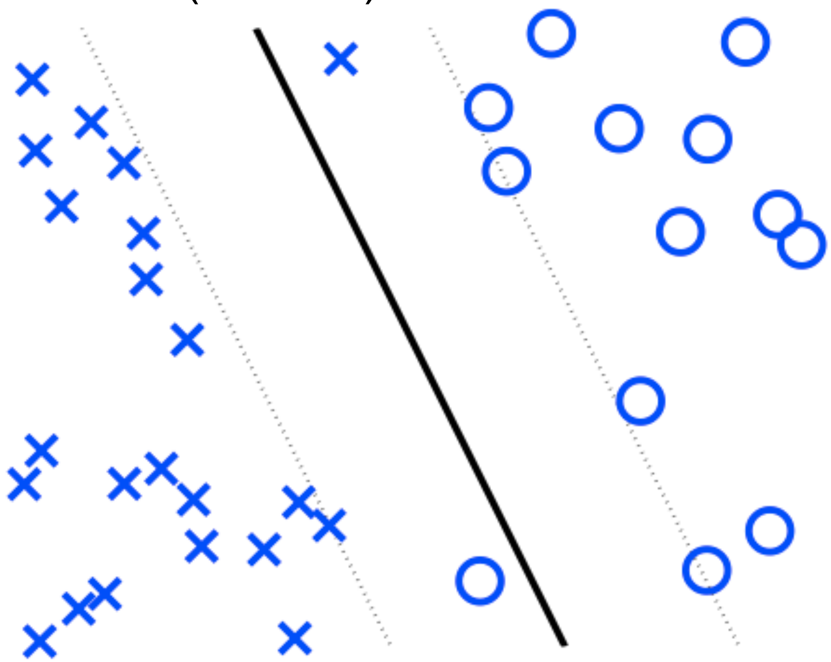
- Using ASM landmarks: compute transformation from first neutral frame to model face:

$$f : (x, y) \mapsto (x', y')$$

- Apply same transformation to all frames
  - This will normalize the expression spatially: certain facial features will always lie on the same spot
- Mathematically:
  - Let polynomials p_i pass over each control point & its *(n-1)* nearest neighbors
  - Compute weights w_i for each polynomial, according to distance of its control point to *(x,y)*. Set $w_i = 0$ for non-local control points
  - For given point *(x,y)*: Compute local weighted mean $\sum w_i p_i(x, y)$

Linear Classifier

- Basic support vector machine (SVM)
- Try to separate feature space linearly
  - Maximizing margin and..
  - ..minimizing penalty for off-liers (tradeoff)
- During training: normal vector *w* and bias *b* are learned

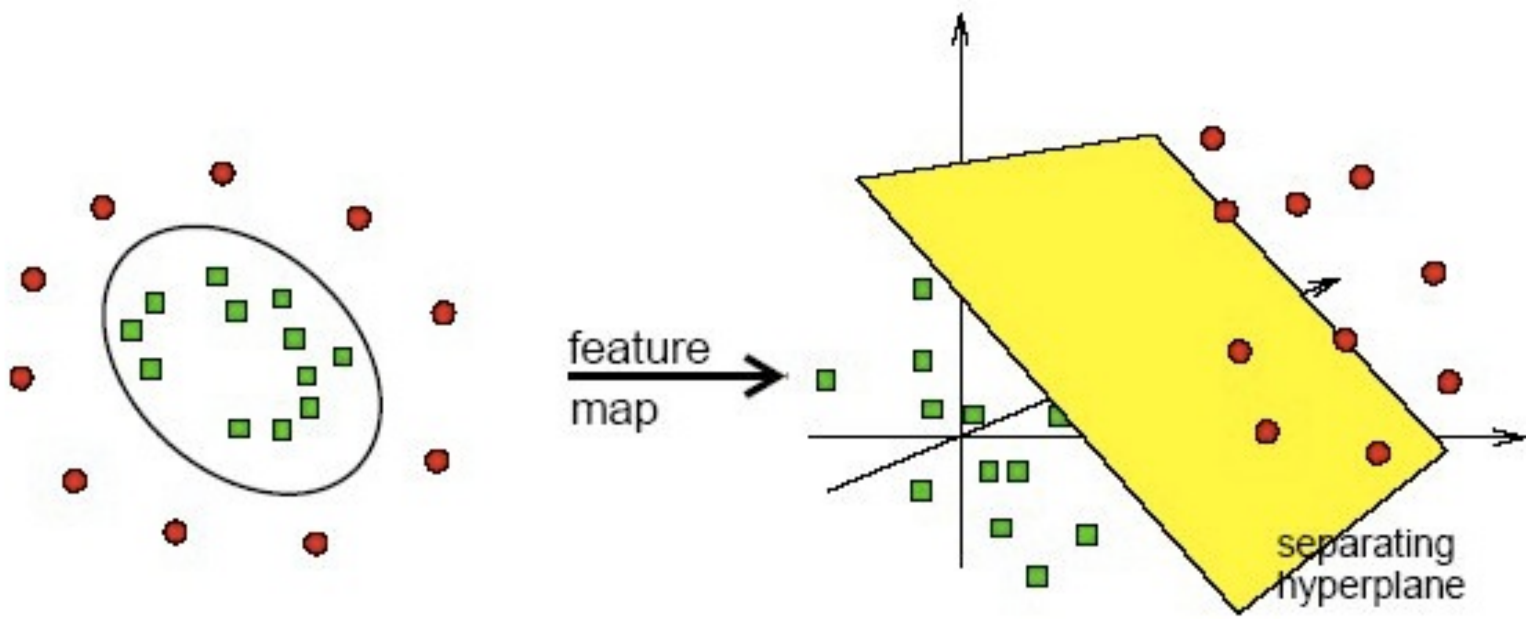


Soft margin SVM

Institute for Anthropomatics, Computer Vision for Human-Computer Interaction Lab, FIPA Group

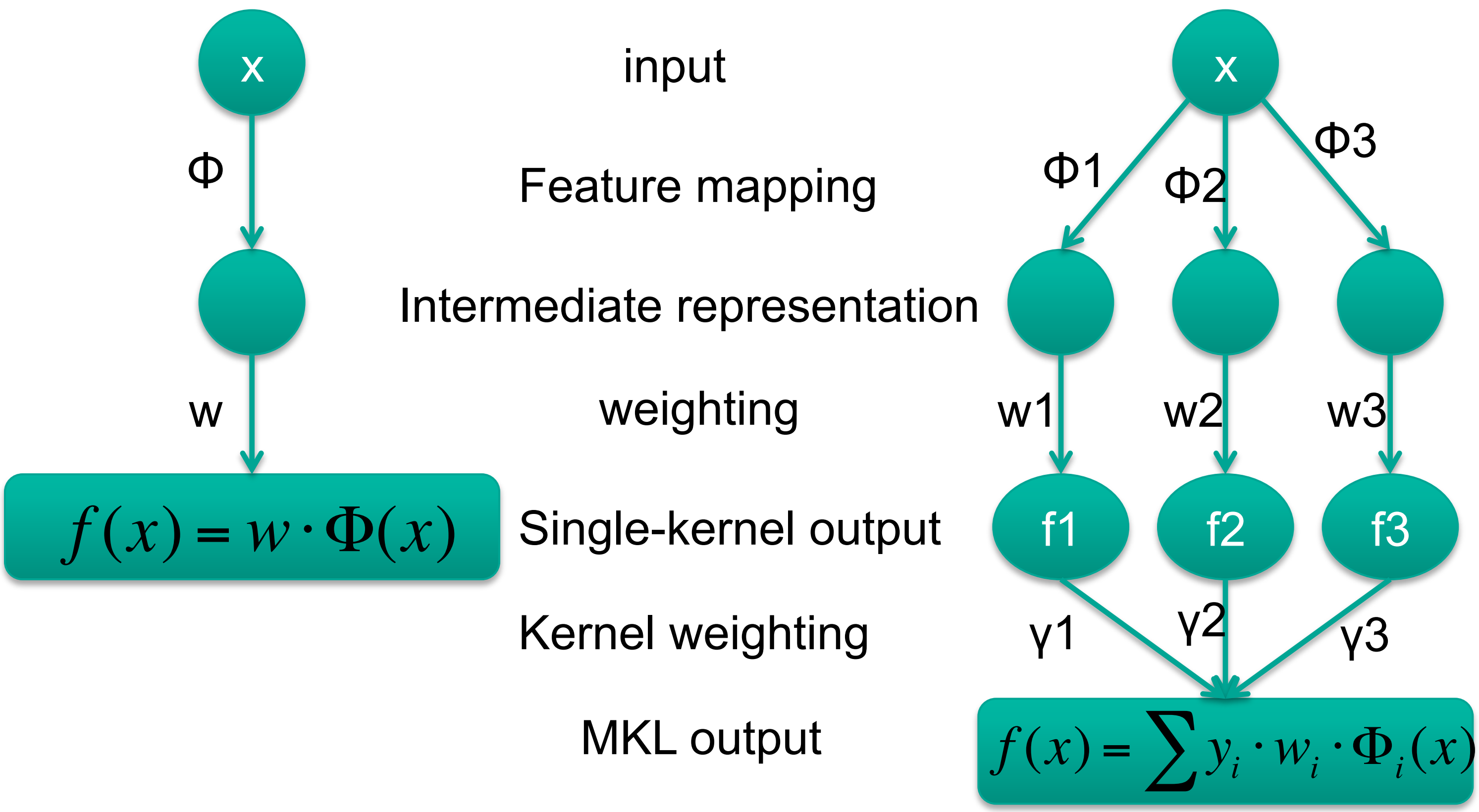
SVM with Kernel Trick

- Perform non-linear transformation into different feature space
- Certain types of non-linear separation are then possible using a standard SVM classifier
- Problem: how to know which Kernel to use?



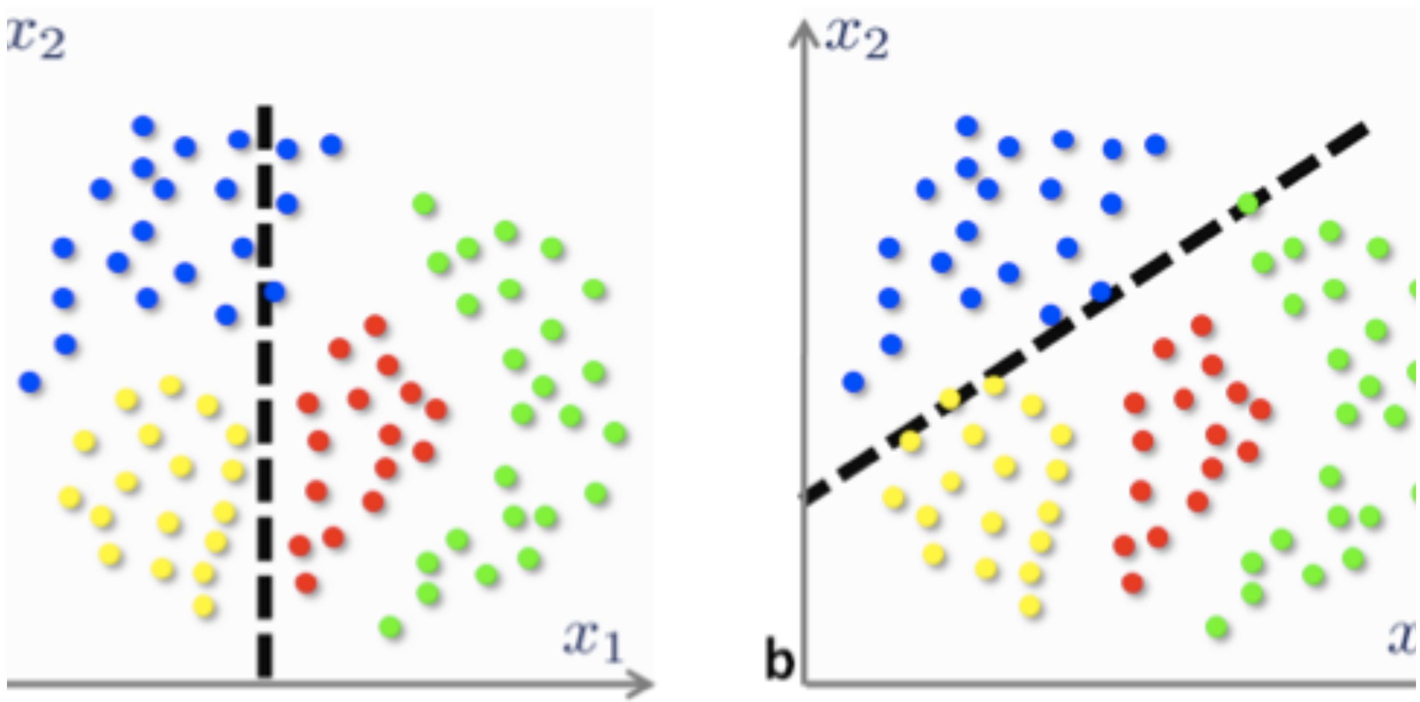
MKL

- Combine kernels (here: polynomial deg 2 & 6, histogram-intersection)
- Train SVM for each kernel
- Learn weights for different kernels and combine them



Random Forests

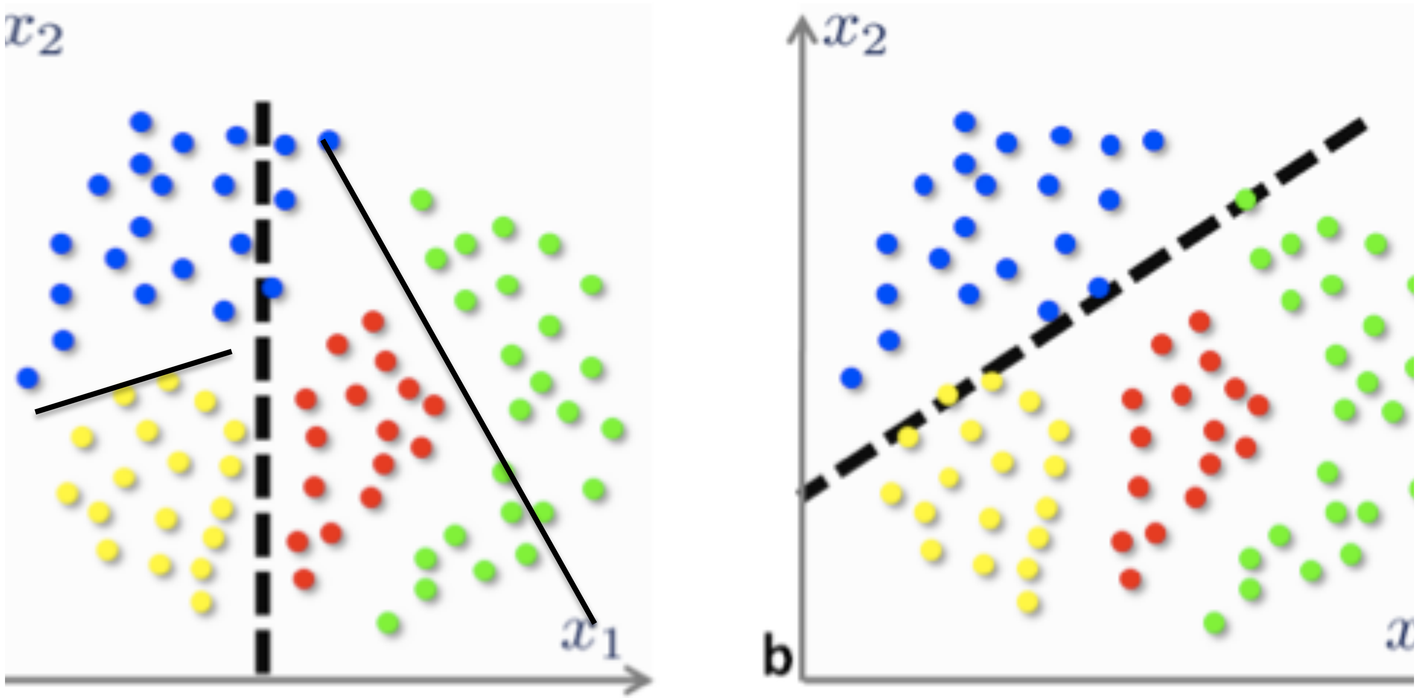
- Construct decision trees
  - Iteratively divide feature space in a way that separates the classes the best



Institute for Anthropomatics, Computer Vision for Human-Computer Interaction Lab, FIPA Group

Random Forests

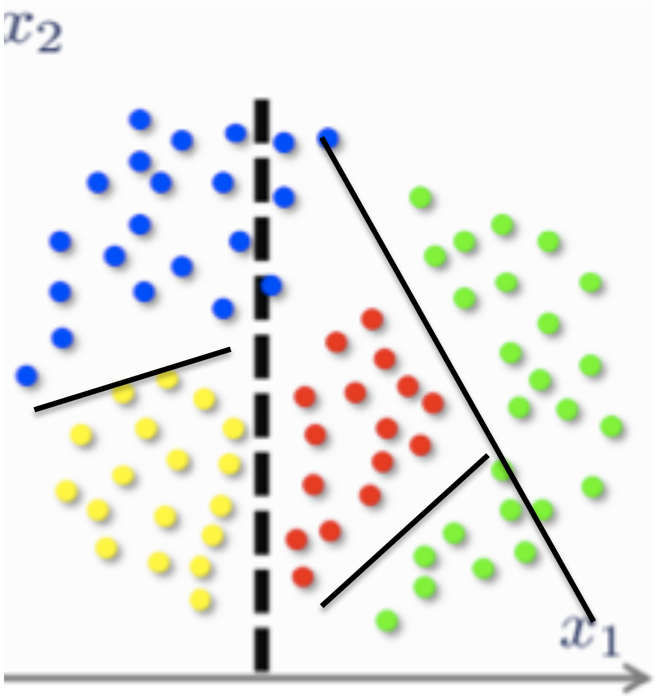
- Construct decision trees
  - Iteratively divide feature space in a way that separates the classes the best

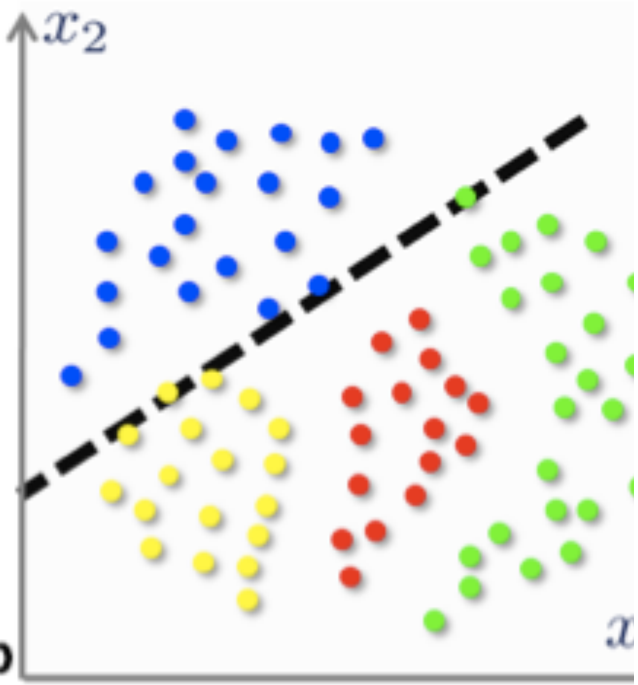


Institute for Anthropomatics, Computer Vision for Human-Computer Interaction Lab, FIPA Group

Random Forests

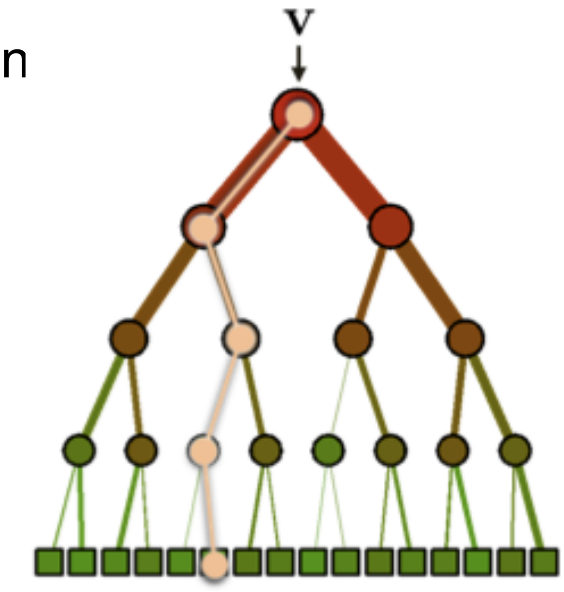
- Construct decision trees
  - Iteratively divide feature space in a way that separates the classes the best

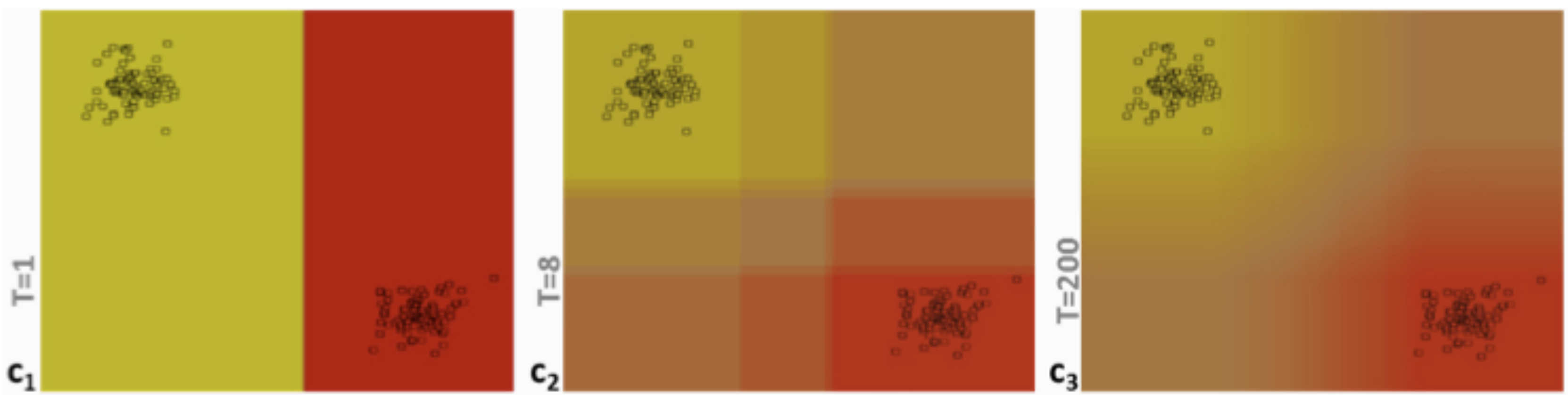


b


Random Forests

- Introduce random component to construction of trees
  - Subsets of samples
  - Randomly disturb separating lines
- Combine trees to make output smoother





Appendix: Tables (Classify Micro)

Phase	Classes	Method	Accuracy (%)
1	detection	RF+TIM15	67.7
1	detection	SVM	70.3
1	detection	RF+TIM20	70.3
1	detection	MKL	71.4
1	detection	RF+TIM10	**74.3**
2	neg/pos	SVM	54.2
2	neg/pos	SVM+TIM15	59.8
2	neg/pos	MKL	60.2
2	neg/pos	MKL+TIM10	**71.4**

Table 2. Leave-one-subject-out results on the SMIC corpus. MKL denotes Multiple Kernel Learning; TIMn denotes temporal interpolation to n frames; RF denotes the Random Forest decision tree classifier; neg/pos denotes classifying negative vs. positive micro-expressions.

Appendix: Tables

Phase	Classes	Method	Accuracy (%)
1	detection	RF+TIM10	58.5
1	detection	SVM+TIM10	65.0
1	detection	MKL+TIM10	70.3
1	detection	RF+TIM15	76.3
1	detection	RF+TIM20	**78.9**
2	neg/pos	SVM+TIM10	51.4
2	neg/pos	MKL+TIM10	60.0
2	neg/pos	MKL+TIM10	60.0
2	neg/pos	SVM+TIM15	62.8
2	neg/pos	MKL+TIM15	**64.9**

Table 3. Leave-one-subject-out results on the SMIC corpus down-sampled to 25fps. MKL denotes Multiple Kernel Learning; TIMn denotes temporal interpolation to n frames; RF denotes the Random Forest decision tree classifier; neg/pos denotes classifying negative vs. positive micro-expressions.

Appendix: Tables (SVP)

Channel	Method	Accuracy LBP (%)	Accuracy CLBP (%)
NIR	SVM	49.3	66.6
NIR	FUS+TIM10	55.7	73.0
NIR	LIN+TIM25	58.0	**78.2**
NIR	LIN+TIM30	62.8	76.9
VIS	SVM	65.3	70.3
VIS	FUS+TIM20	66.0	**72.0**
VIS	SVM+TIM25	66.6	70.0
VIS	SVM+TIM30	66.6	70.0
NIR+VIS	MKL+TIM25	66.8	**80.0**

Table 1. Leave-one-subject-out results on the SPOS corpus with CLBP-TOP and LBP-TOP. NIR denotes the near-infrared channel; VIS denotes the visual channel; SVM denotes support vector machines; MKL denotes Multiple Kernel Learning; TIMn denotes temporal interpolation to n frames; LIN denotes the LINEAR classifier; FUS denotes fusion of SVM, LINEAR and Random Forest through majority voting.

Appendix: Tables (CLBP-TOP)

Components	Method	Accuracy (%)
S+M+C	FUS+TIM10	**73.0**
S+M	FUS+TIM10	72.7
M+C	FUS+TIM10	71.7
S+C	FUS+TIM10	56.4
S	FUS+TIM10	55.7
S+M+C	LIN+TIM25	**78.2**
S+M	LIN+TIM25	76.2
M+C	LIN+TIM25	73.0
S+C	LIN+TIM25	62.1
S	LIN+TIM25	58.0

Table 2. Leave-one-subject-out results on the SPOS corpus comparing different CLBP-TOP components. NIR data were used for this experiment. C is the centre grey level; S is the sign and M is the magnitude of the local difference d_p. TIMn denotes temporal interpolation to n frames; LIN denotes the LINEAR classifier; FUS denotes fusion of SVM, LINEAR and Random Forest through majority voting.

Appendix: Tables (FED)

Method	Accuracy (%)
CLBP+SVM	58.8
CLBP+MKL	64.7
CLBP+RF	68.6

Table 3. Leave-one-subject-out results for LAYER1-FED with visual data. SVM denotes support vector machines; MKL denotes multiple kernel learning; RF denotes the Random Forest decision tree classifier.

References

- Main papers:
  - [1] Pfister et al.: Recognising spontaneous facial micro-expressions. (2011)
  - [2] Pfister et al..: Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. (2011)
- Referenced literature:
  - [3] Ekman, P.: Lie Catching and Microexpressions. (2009)
  - [4] Ekman et al.: Detecting deception from the body or face. (1974)
  - [5] Polikovsky et al.: Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. (2009)
  - [6] Shreve, M. et al.: Macro- and micro-expression spotting in long videos using spatio-temporal strain. (2011)
  - [7] Valstar et al.: How to distinguish posed from spontaneous smiles using geometric features.(2007)

- Referenced literature (continued):

  - [8] Bartlett et al.: Verbal and nonverbal features of human-human and human-machine interaction. (2008)
  - [9] Warren et al.: Detecting deception from emotional and unemotional cues.
  - [10] Mihalcea et al.: The lie detector: explorations in the automatic recognition of deceptive language. (2009)
  - [11] Michael et al.: Motion profiles for deception detection using visual cues. (2010)
  - [12] Cootes et al.: Active shape models - their training and application. (1995)
  - [13] Milborrow et al.: Locating facial features with an extended active shape model. (2008)
  - [14] Papageorgiou et al.: A general framework for object detection. (1998)

- Referenced literature (continued):

  - [15] Yan et al.: Graph embedding and extension: A general framework for dimensionality reduction (2007)
  - [16] Zhou et al.: Towards a practical lipreading system. (2011)
  - [17] Ojala et al.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. (2002)
  - [18] Guo et al.: A completed modeling of local binary pattern operator for texture classification. (2010)
  - [19] Valstar et al.: Spontaneous vs. posed facial behavior: Automatic analysisof brow actions. (2006)