SPIE 2011 on 24th Jan.

Appearance-based Human **Gesture Recognition** using Multimodal Features for Human Computer Interaction

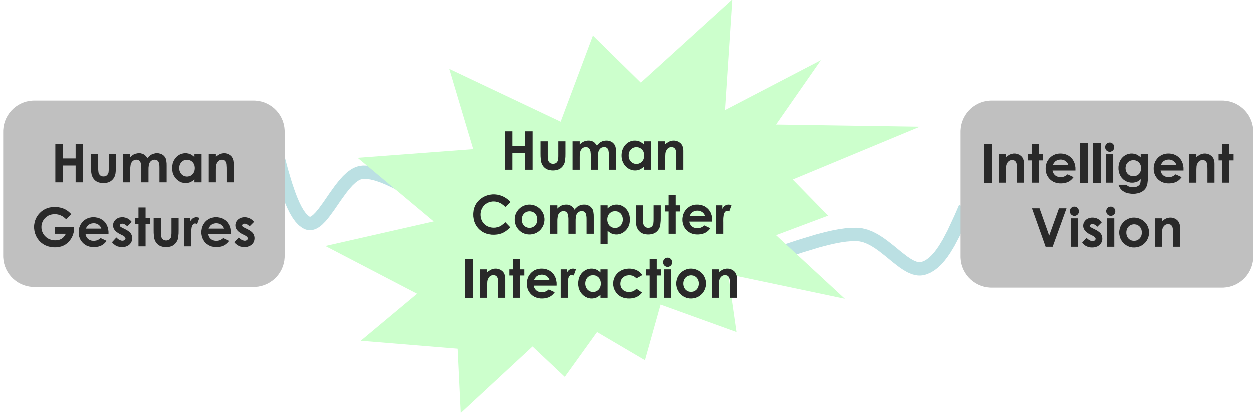
Luo Dan[a,b] Hua Gao[b] Hazim Kemal Ekenel[b] Jun Ohya[a]

GITS, Waseda University[a]
CVHCI, Karlsruhe Institute of Technology[b]

KIT Karlsruhe Institute of Technology

Waseda University

Introduction



- Control of consumer electronics
- Interaction with visualization systems
- Control of mechanical systems
- Computer games

Challenges

- Different components of human gestures

- Wide variety of signs (ambiguous)
- Variable appearance/clothing
- Unconstrained illumination
- Local-body Occlusions

Real Time Large Vocabulary Continuous Sign Language Recognition Based on OP/Viterbi Algorithm

Guilin Yao, ICPR2006

Two cyber gloves and **three pohelmus 3SPACE-position trackers** are used as input devices.

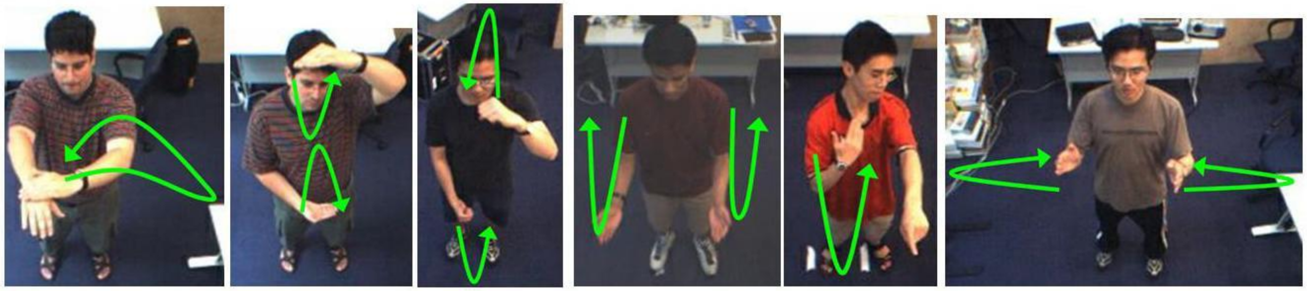
4942 isolated signs from two signers

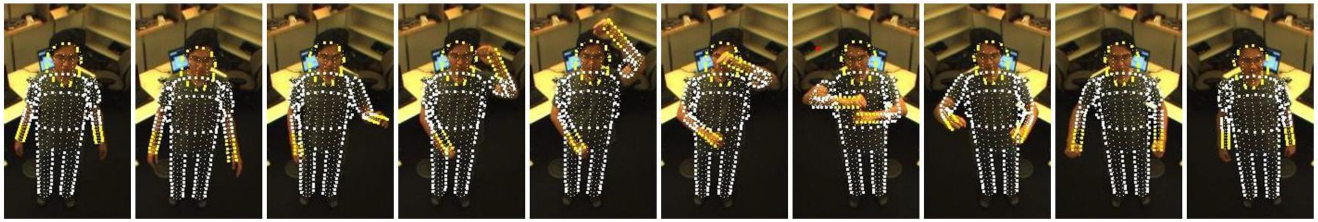
3312 in the test set

78.1%

Hidden Conditional Random Fields for Gesture Recognition

Sy Bor Wang, CVPR2006





A 3D cylindrical body model,

Stereo Camera

Head Gesture Dataset

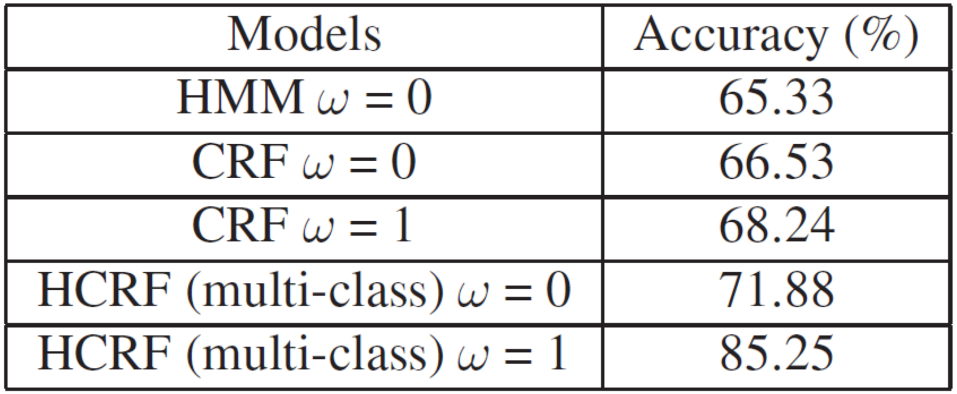
16 signer

A total of 152 head nods, 11 head shakes and 159 junk sequences

Arm Gesture Dataset

13 signer 6 classes

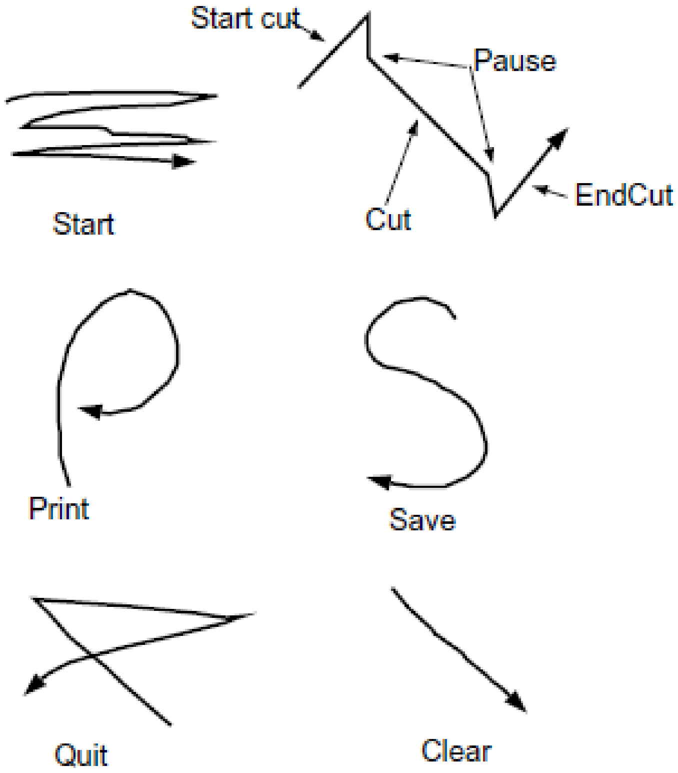
90 gestures for per class



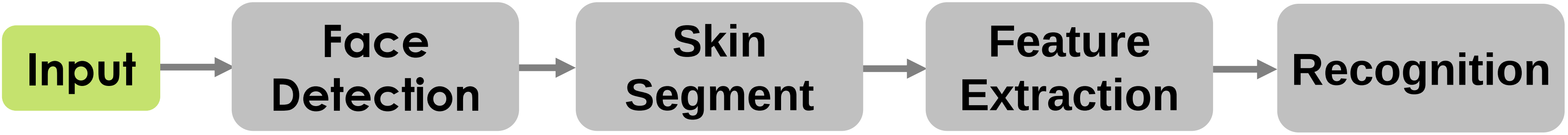
Models	Accuracy (%)
HMM $\omega = 0$	65.33
CRF $\omega = 0$	66.53
CRF $\omega = 1$	68.24
HCRF (multi-class) $\omega = 0$	71.88
HCRF (multi-class) $\omega = 1$	85.25

Recognizing temporal trajectories using the condensation algorithm

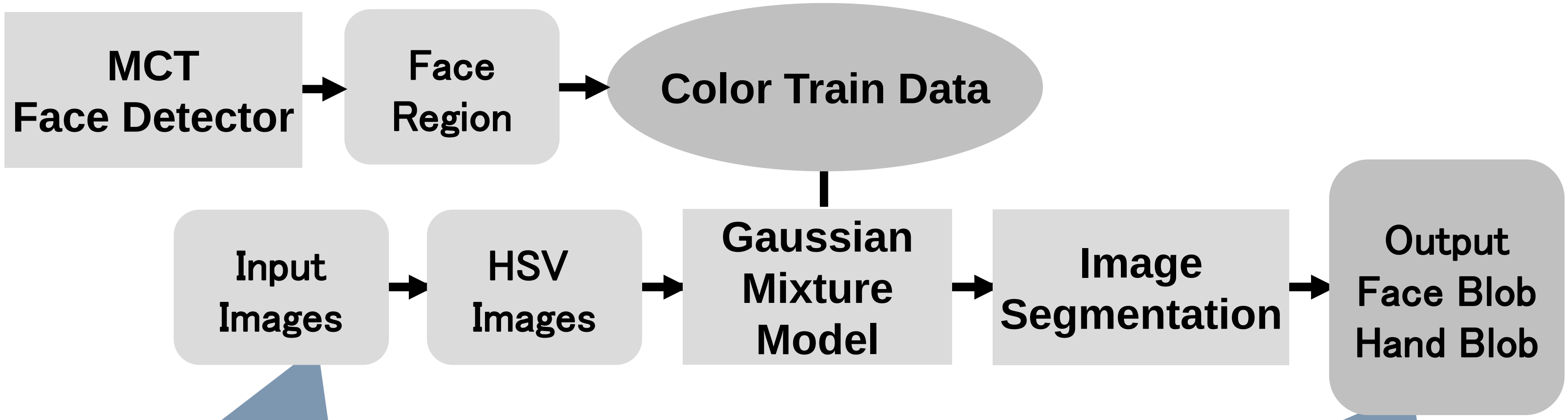
M. J. Black and Jepson, FG1998



System Overview

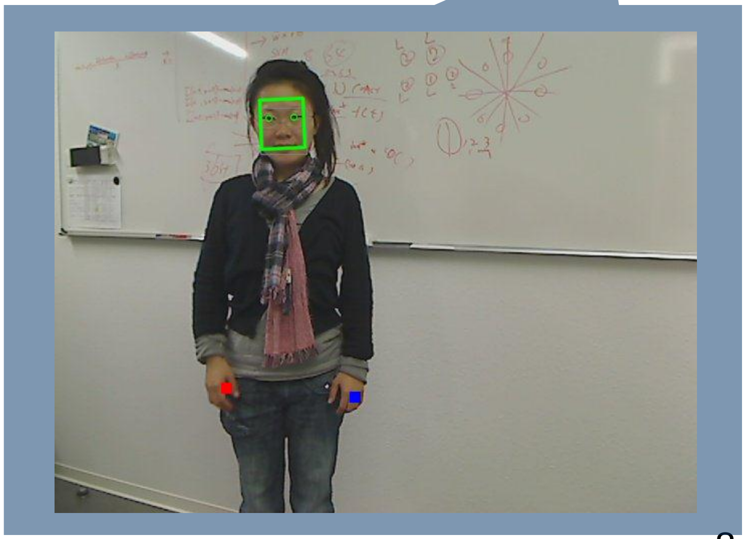


Multimodal Feature



MCT Face Detector → Face Region → Color Train Data

Input Images → HSV Images → Gaussian Mixture Model → Image Segmentation → Output Face Blob Hand Blob



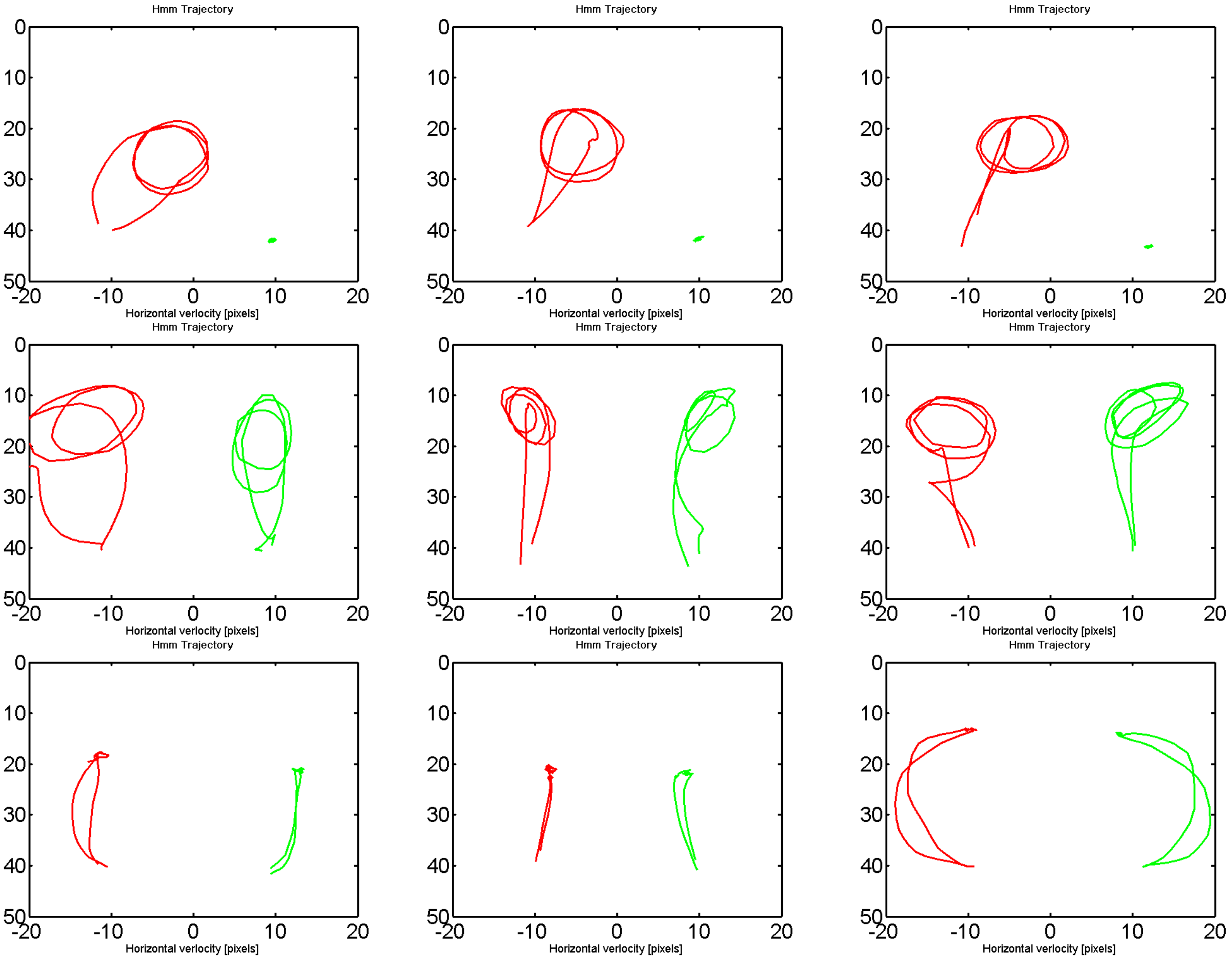
Input → Face Detection → Skin Segment → Feature Extraction → Recognition

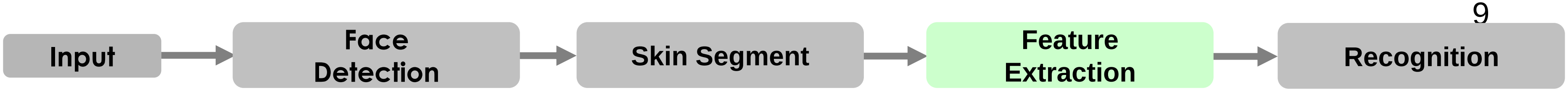
Hand Feature

Disgust

Excite

Nervous





Input

Face Detection

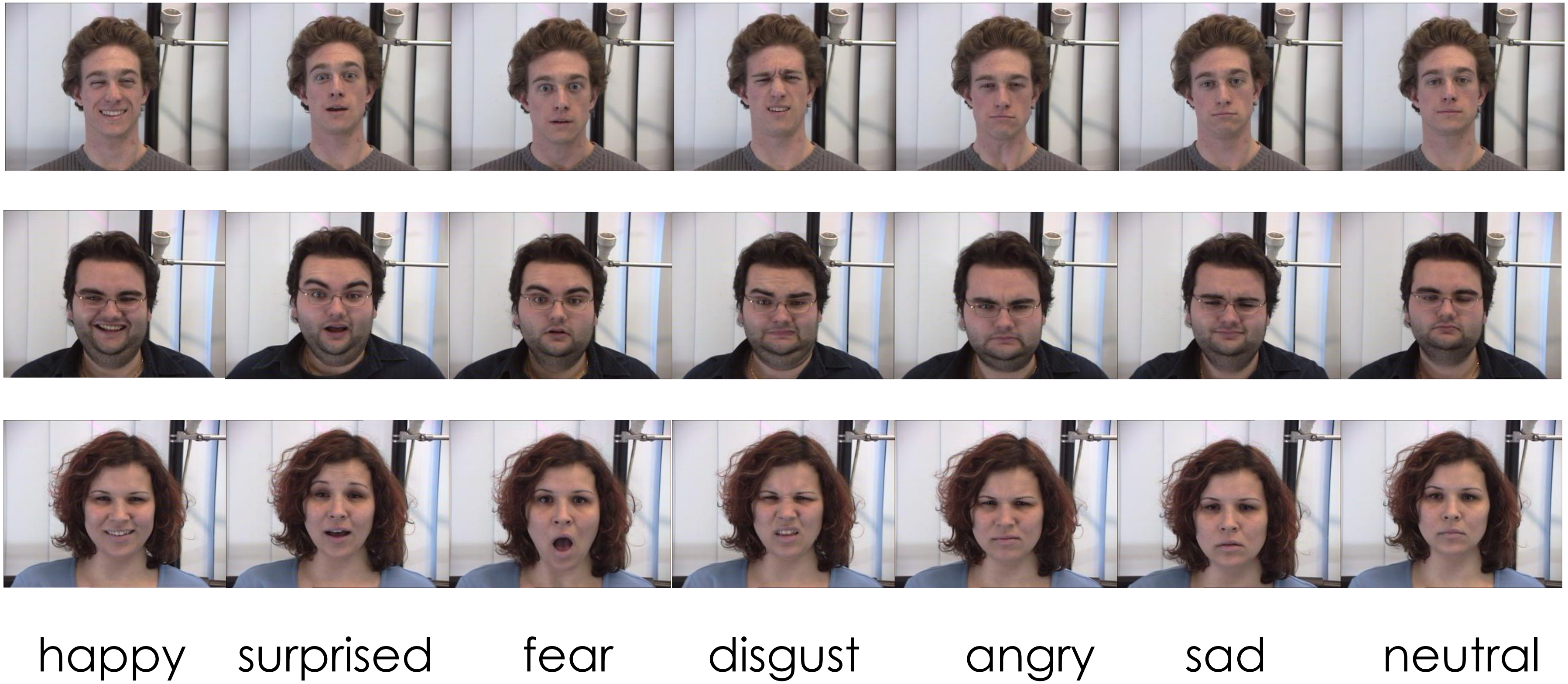
Skin Segment

Feature Extraction

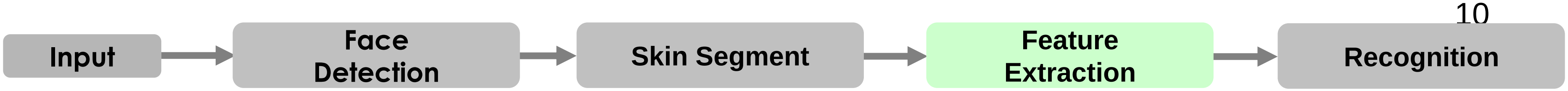
Recognition

Facial feature

7 Face expression subject [Training Dataset: "FEEDTUM"]



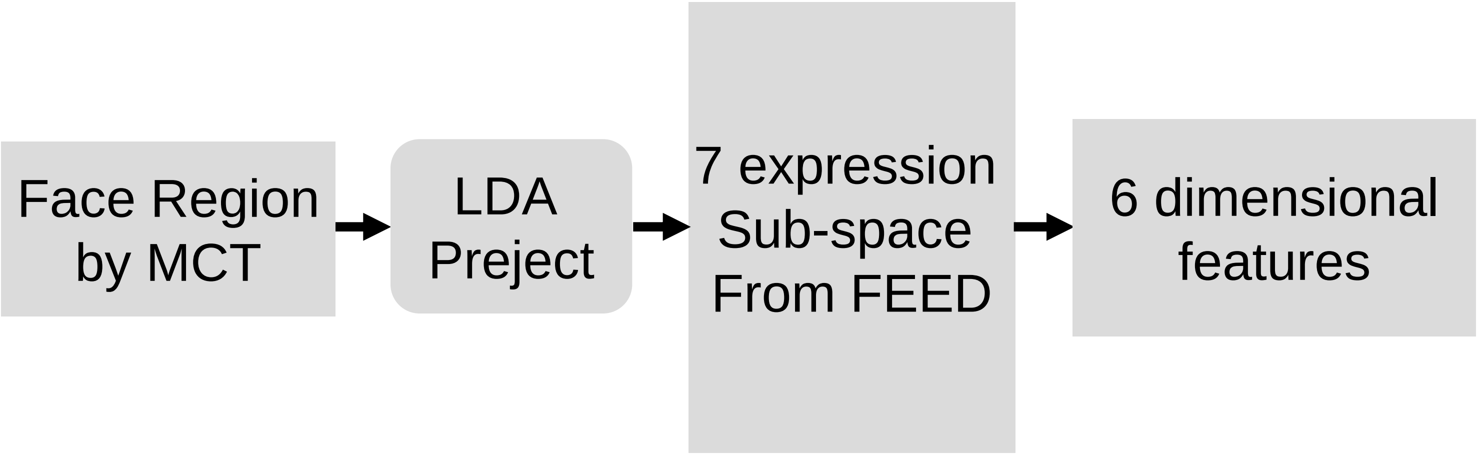
5 DCT coefficients from 64 blocks
Facial appearance feature vector(5 × 64=320 dimensional)



Input → Face Detection → Skin Segment → Feature Extraction → Recognition

Facial feature

Expression Subspace-Expression Trajectories



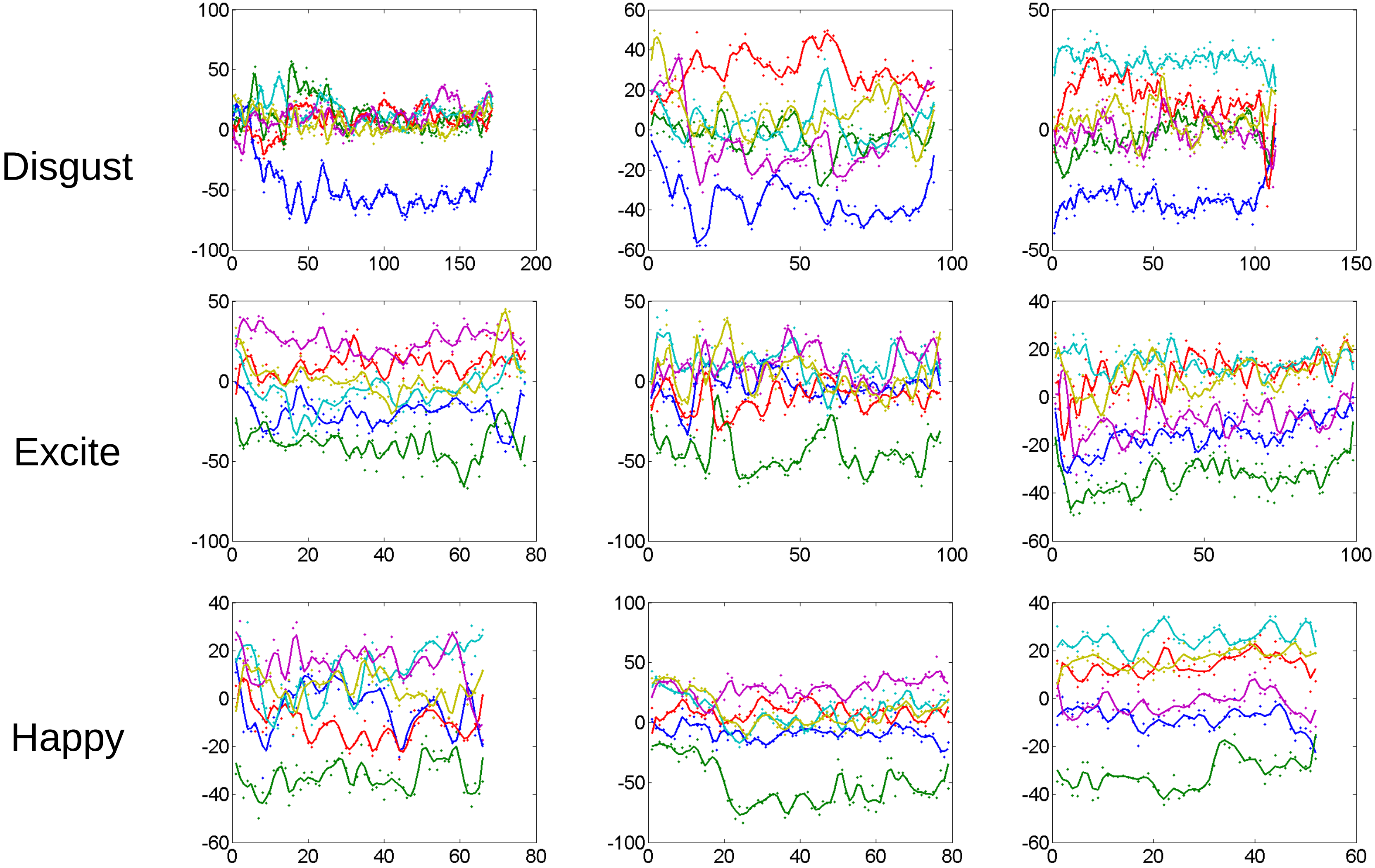
Face Region by MCT

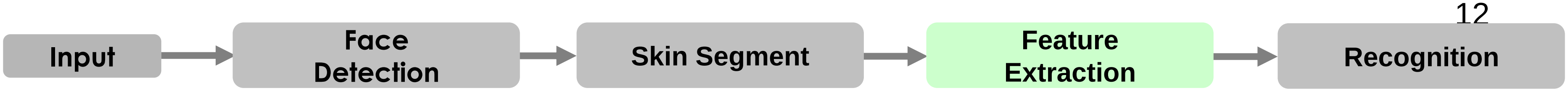
LDA Preject

7 expression Sub-space From FEED

6 dimensional features

Facial Feature




Input

Face Detection

Skin Segment

Feature Extraction

Recognition

Feature Combination

- Hand feature

Hand Location

- Facial feature

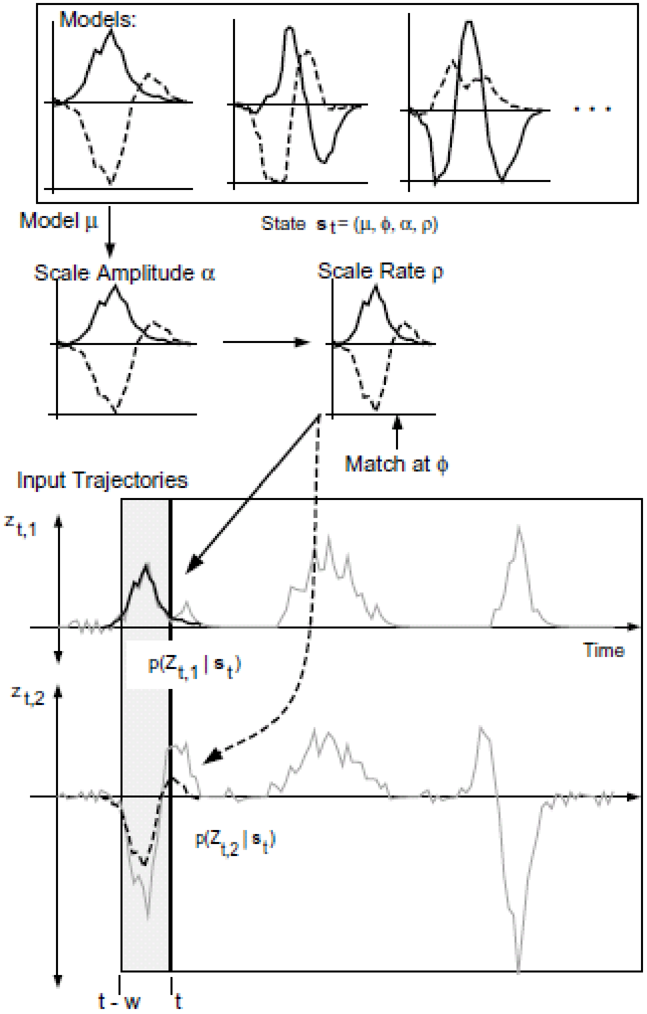
6 dimensional vectors

- Two different combination strategies

- The first one is at feature level by combining the feature vectors extracted from face and hands. A statistical method can be used afterwards to select the most discriminative features for classification.
- The second one is at decision level by combining the classification scores of each modality.

Input
Face Detection
Skin Segment
Feature Extraction
Recognition

Condensation



- **The sample set**

$$S_t = (\mu, \phi^l, \alpha^l, \rho^l, \phi^\gamma, \alpha^\gamma, \rho^\gamma)$$

- **Prediction**

$$\mu \in [1, \mu_{max}]$$

$$\varphi^i = \frac{1-\sqrt{y}}{\sqrt{y}}, \quad y \in [0,1]$$

$$\alpha^i \in [\alpha_{min}, \alpha_{max}]$$

$$\rho^i \in [\rho_{min}, \rho_{max}]$$

$$\mu_{t+1} = \mu_t$$

$$\varphi^j_{t+1} = \varphi^j_t + \varphi^j_t + N(\sigma_\varphi)$$

$$\alpha^j_{t+1} = \alpha^j_t + N(\sigma_\alpha)$$

$$\rho^j_{t+1} = \rho^j_t + N(\sigma_\rho)$$

- **Updating**

$$p(Z_{t,i} \mid s_t) = \frac{1}{\sqrt{2\pi}} \exp \frac{-\sum_{j=0}^{\omega-1} (x_{(t-j),i} - \alpha^* m^{(\mu\mu}_{(\varphi^* - \rho^* j),i})^2}{2(\omega - 1)}$$

Condensation

- **The sample set in each state**

Hand feature trajectories

$$S_t = (\mu, \phi^l, \alpha^l, \rho^l, \phi^\gamma, \alpha^\gamma, \rho^\gamma)$$

Facial feature trajectories

$$\mathrm{S}_\mathrm{t} = (\mu, \phi^f, \alpha^f, \rho^f)$$

Hand-face feature trajectories

$$\mathrm{S}_\mathrm{t} = (\mu, \phi^\mathrm{l}, \alpha^\mathrm{l}, \rho^\mathrm{l}, \phi^\gamma, \alpha^\gamma, \rho^\gamma, \phi^f, \alpha^f, \rho^f)$$

Input

Face Detection

Skin Segment

Feature Extraction

Recognition

Experiment – dataset

•180 video clips of 12 human gestures with facial expression
(1)anger, (2)apologize,(3)appreciate, (4)desire, (5)disgust, (6)excite,
(7)fear, (8)happy, (9)nervous, (10)sad, (11)so-so and (12)surprise,
Selected from ASL.

•Each sign includes three phases of a gesture: prestroke, stroke and poststroke.

•3 people perform 3 to 7 times for each gesture. 1 as test data and the other 2 as train data.

•A training set and a testing data-set for evaluation. The training set contains one recording session per person,
i.e. 12 × 3 = 36 video clips. The rest of the clips are used for test.

•Each video clip has a spatial resolution of 640 × 480 pixels with a frame-rate of 25fps and it is captured by a Logitech Webcam Pro 9000 from frontal view.

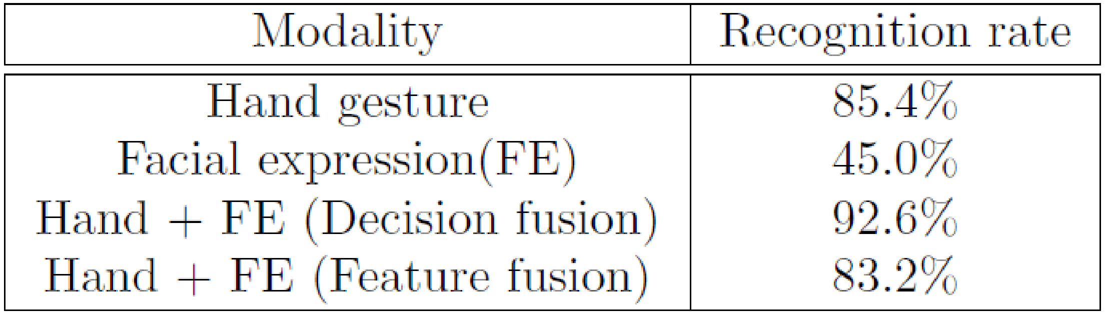
Experiments

Two different combination strategies
- Feature Level

by combining the feature vectors extracted from LDA face projection feature and hand trajectories. A statistical method (condensation) can be used afterwards to select the most discriminative features for classification.
Feature [Face, Hand], Recognition result: 83.2%

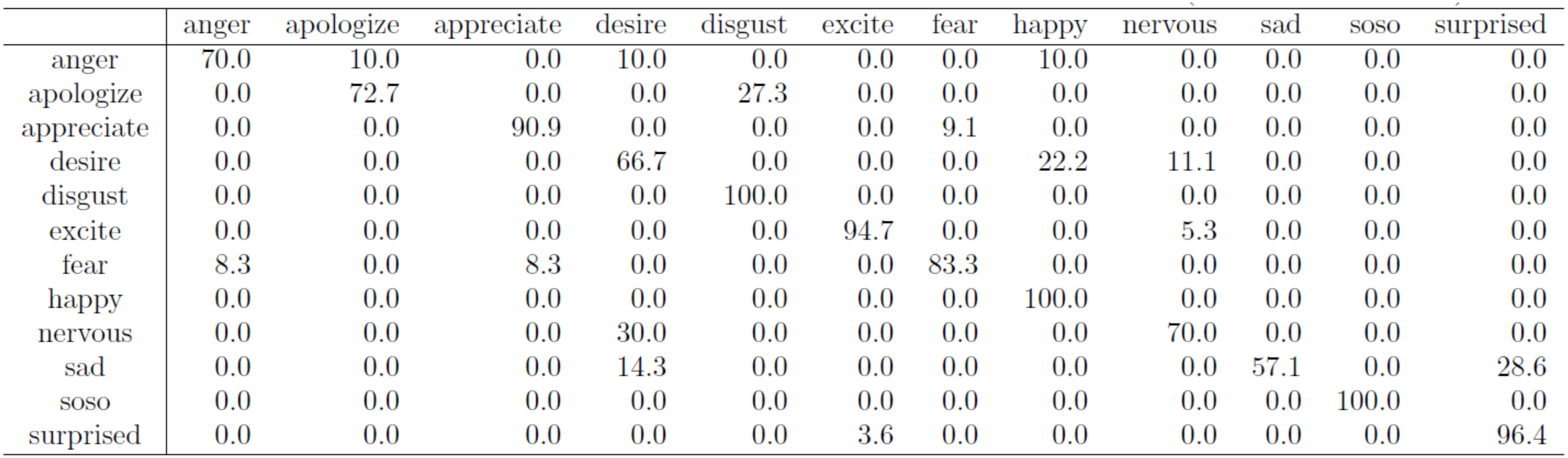
- Decision level

by combining the classification scores of each modality.
Feature [Hand] | Feature [face], Recognition result: 92.6%



Modality	Recognition rate
Hand gesture	85.4%
Facial expression(FE)	45.0%
Hand + FE (Decision fusion)	92.6%
Hand + FE (Feature fusion)	83.2%

Experiments



	anger	apologize	appreciate	desire	disgust	excite	fear	happy	nervous	sad	soso	surprised
anger	70.0	10.0	0.0	10.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0
apologize	0.0	72.7	0.0	0.0	27.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
appreciate	0.0	0.0	90.9	0.0	0.0	0.0	9.1	0.0	0.0	0.0	0.0	0.0
desire	0.0	0.0	0.0	66.7	0.0	0.0	0.0	22.2	11.1	0.0	0.0	0.0
disgust	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
excite	0.0	0.0	0.0	0.0	0.0	94.7	0.0	0.0	5.3	0.0	0.0	0.0
fear	8.3	0.0	8.3	0.0	0.0	0.0	83.3	0.0	0.0	0.0	0.0	0.0
happy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
nervous	0.0	0.0	0.0	30.0	0.0	0.0	0.0	0.0	70.0	0.0	0.0	0.0
sad	0.0	0.0	0.0	14.3	0.0	0.0	0.0	0.0	0.0	57.1	0.0	28.6
soso	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
surprised	0.0	0.0	0.0	0.0	0.0	3.6	0.0	0.0	0.0	0.0	0.0	96.4

Confusion matrix for the condensation-based classification on database (hand motion result)

Conclusion

•an appearance-based multi-modal gesture recognition framework, which combines facial expression and hand motions.

•12 classes of human gestures with facial expression from ASL.

•Two fusion strategies: the decision fusion and feature fusion.

•Experimental results showed that the analysis of facial expression helps distinguishing ambiguous hand gestures and facial analysis improves hand gesture recognition.

•In particular, decision level fusion improves the recognition rate from 85:4% to 92:6%.

Acknowledgment

- InterACT program
  Waseda University and Karlsruhe Institute of Technology

- German Excellence Initiative
  “Concept for the Future”

Thank you !

Q?