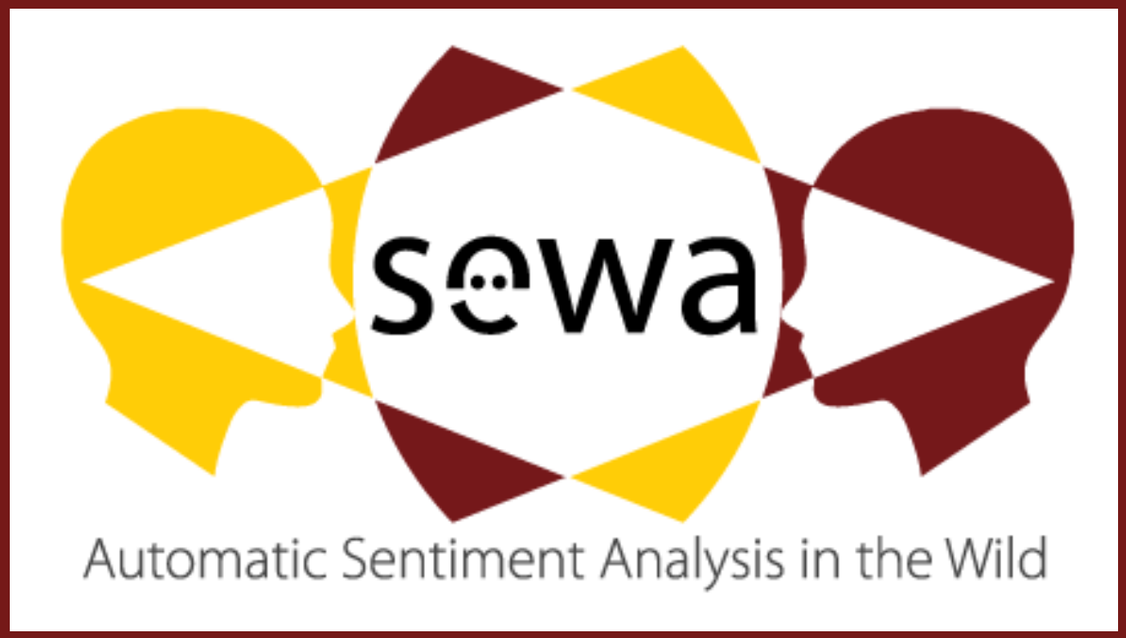
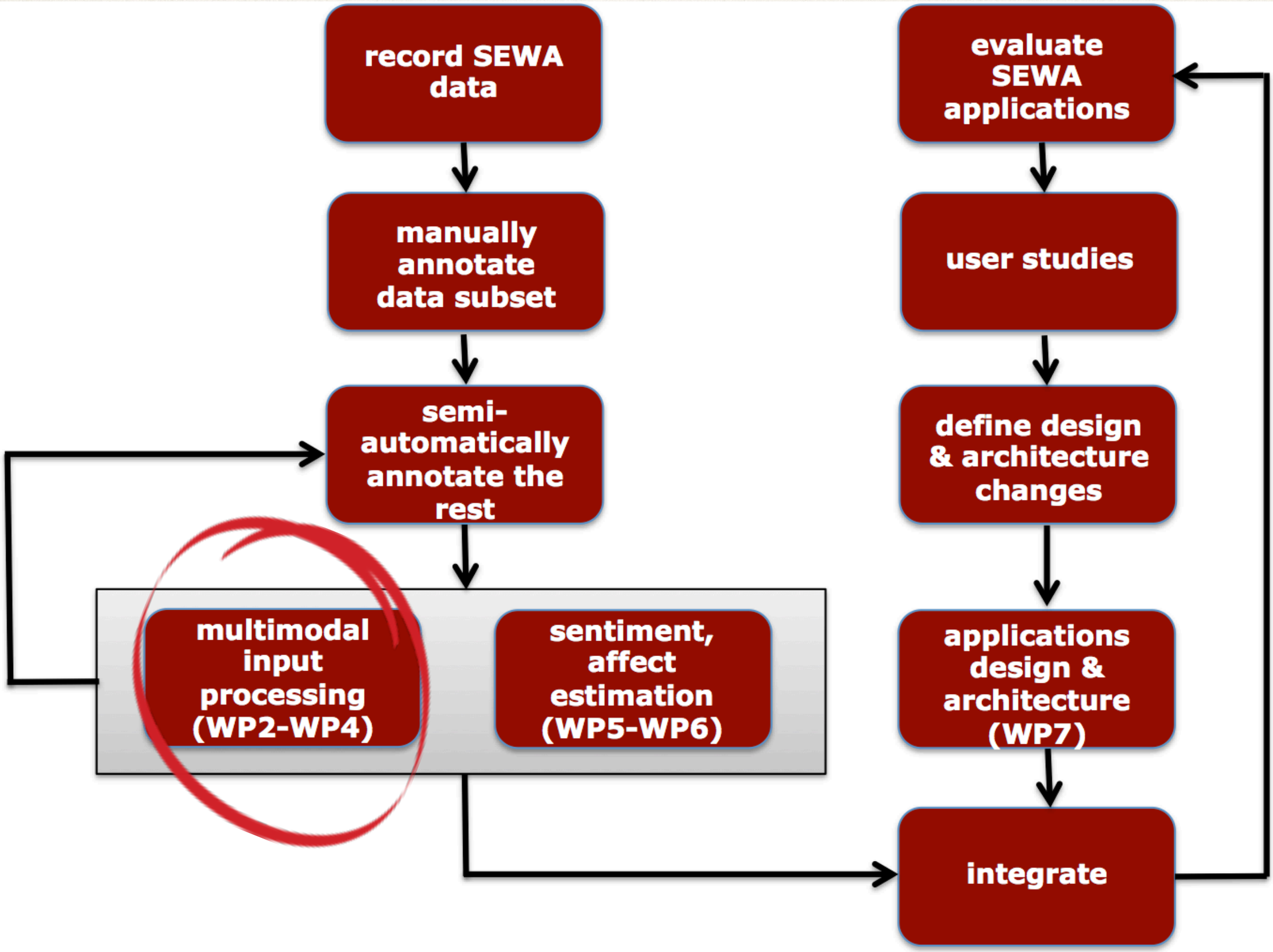
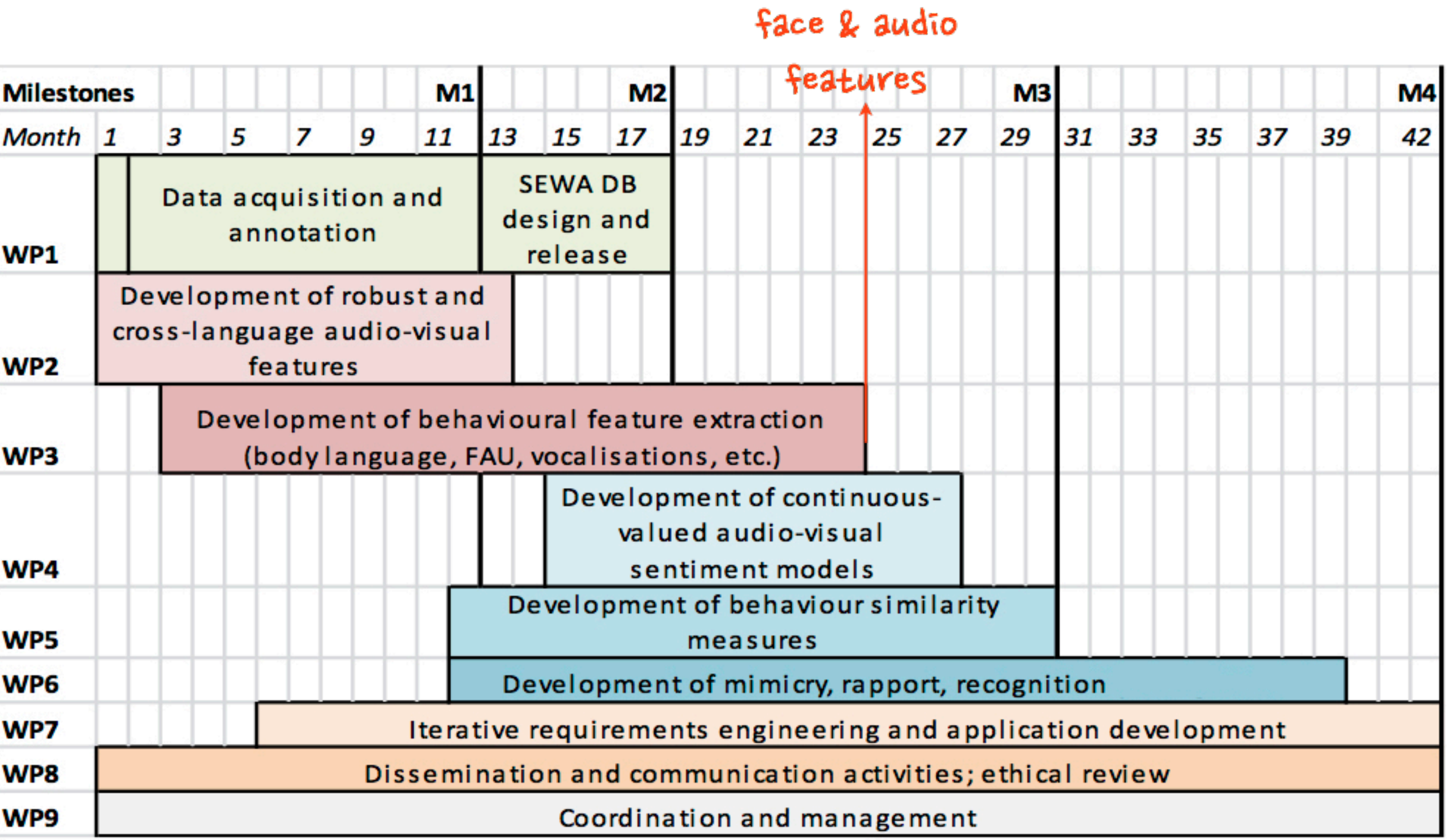
WP3:
Mid-level Feature Extraction

Ognjen Rudovic



Automatic Sentiment Analysis in the Wild




face & audio features
Milestones
M1
M2
M3
M4
Month
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
42
WP1
Data acquisition and annotation
SEWA DB design and release
WP2
Development of robust and cross-language audio-visual features
WP3
Development of behavioural feature extraction (body language, FAU, vocalisations, etc.)
WP4
Development of continuous-valued audio-visual sentiment models
WP5
Development of behaviour similarity measures
WP6
Development of mimicry, rapport, recognition
WP7
Iterative requirements engineering and application development
WP8
Dissemination and communication activities; ethical review
WP9
Coordination and management

Objectives

- Automatic detection of head and hand gestures (D3.1)
- Facial Action Unit detection and intensity estimation (D3.1)
- Audiovisual detection of non-verbal vocalisations (D3.2)

Imperial College London

UNIVERSITÄT PASSAU

real eyes

PLAYGEN

Automatic Sentiment Analysis in the Wild

WP2: Detection of audio and visual features

WP3:
- head nods and shakes
- hand-touching-the-face gestures
- FAUs
- non-verbal vocalisations

WP4-WP6:
- sentiment
- affect
- intentions

Imperial College London

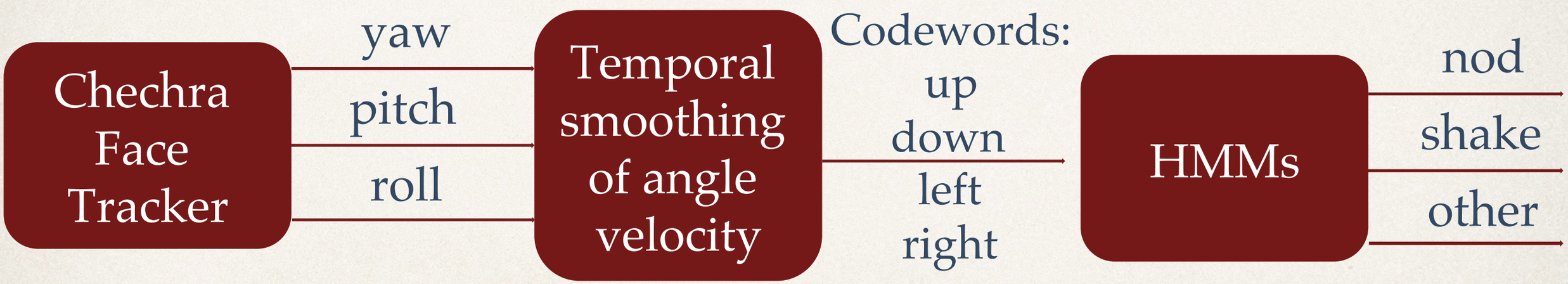
UNIVERSITÄT PASSAU

real eyes

PLAYGEN

Head Gestures

- Automatic detection of head nods and shakes using the state-of-the-art method for head node/shake detection based on HMMs

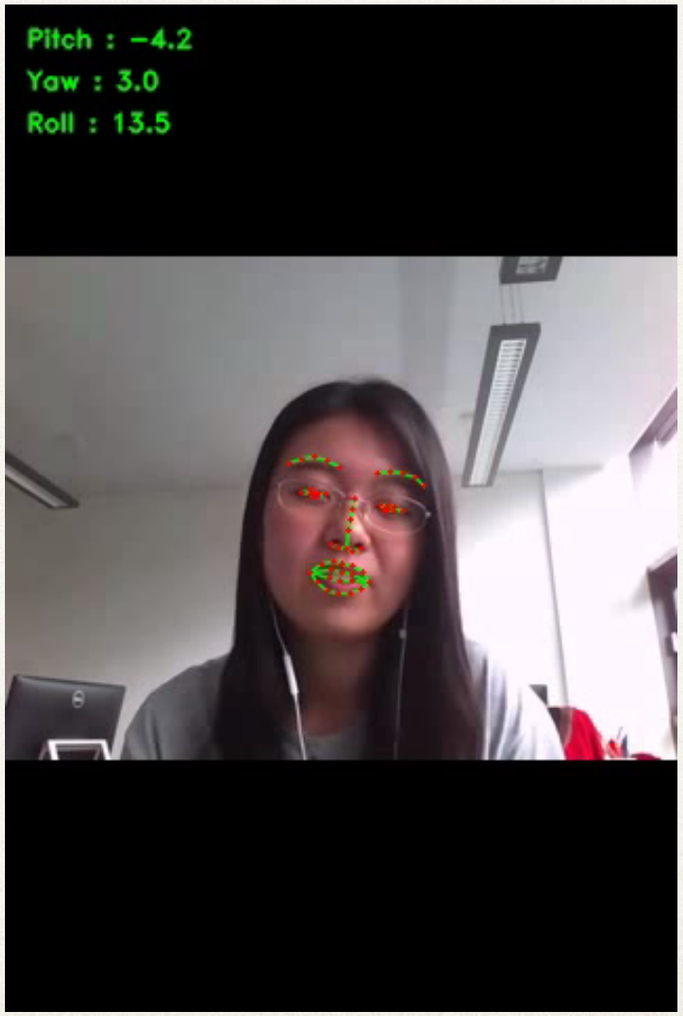


Imperial College London

Automatic Sentiment Analysis in the Wild

Head Gestures

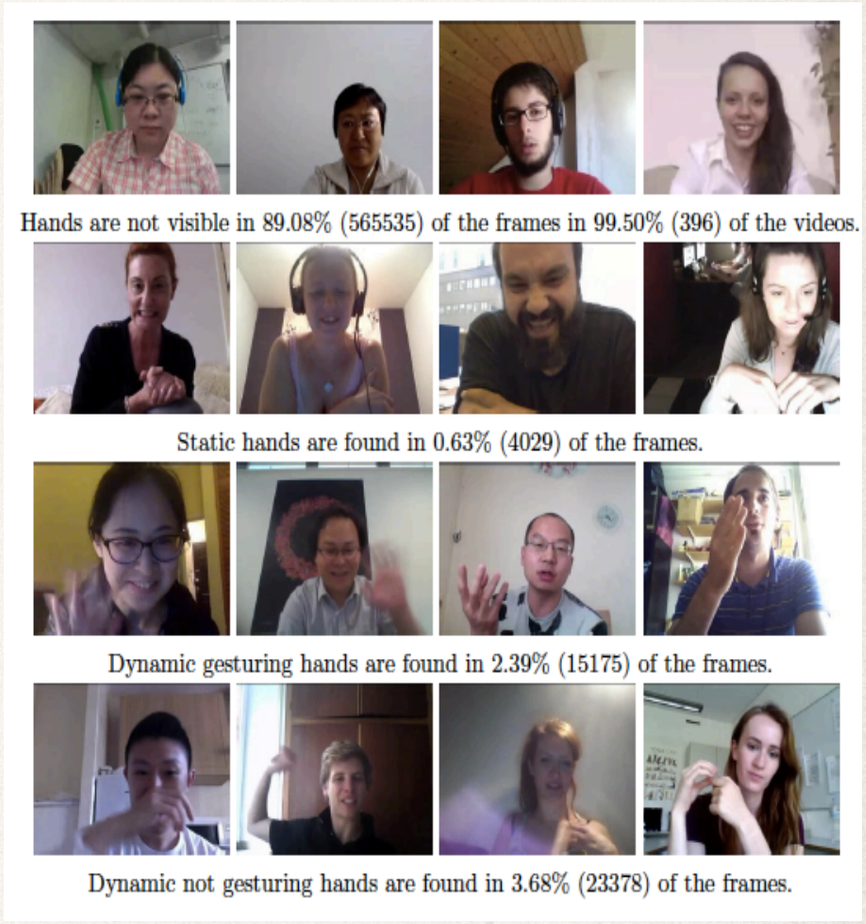
- Examples of automated detection of nods/shakes from SEWA videos



Imperial College London

Hand Gestures

- The automated extraction of touching-the-face gestures has been attempted in two ways: by analysis of dynamic hand movement and static face touching.
- The-face-touching and dynamic gestures cannot be seen in many videos, and most of the events are very short.
- The state-of-the-art hand trackers yield quite poor results, having many false positives and low true positives.
- For these reasons, we excluded the hand gestures from the set of mid-level features originally envisioned.



Facial Action Units

- The goal is to perform Action Unit (AU) intensity estimation and detection from SEWA videos.
- Target AUs:

AU1: (Inner Brown Raiser)

AU2: (Outer Brown Raiser)

AU4: (Brow Lowerer)

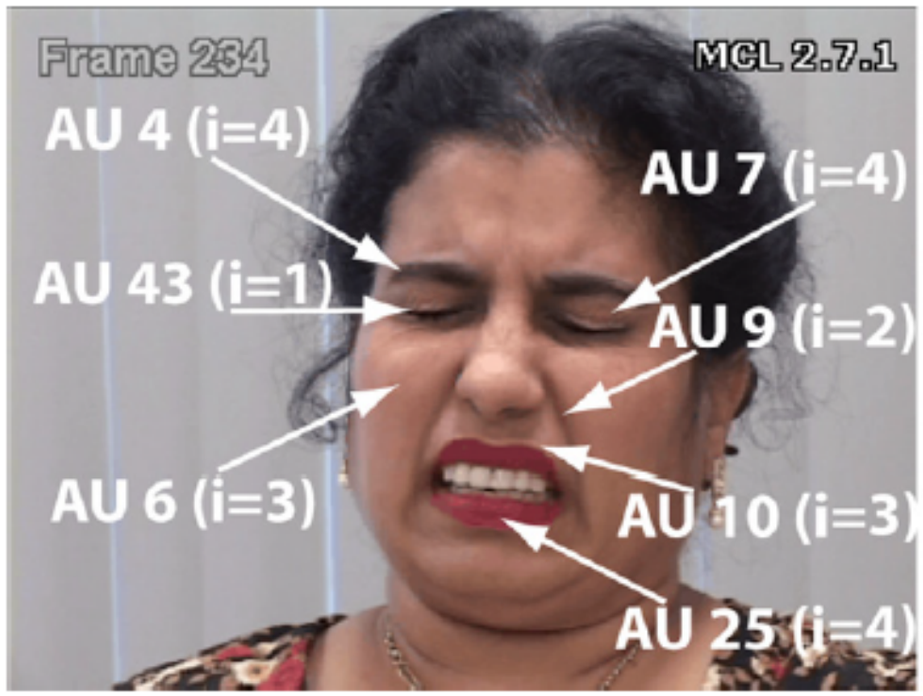
AU12: (Lip Corner Puller)

AU17: (Chin Raiser)

Imperial College London

Facial Action Units

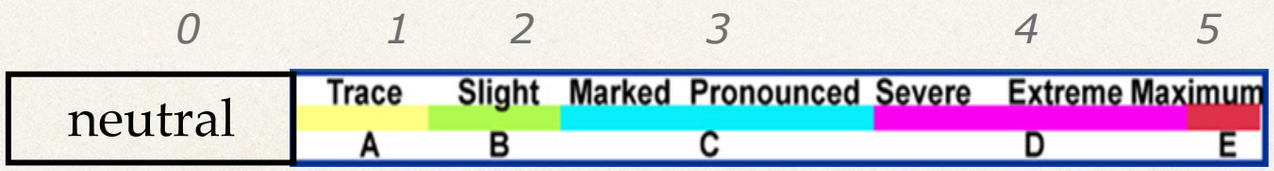
- Intensity coding: overview



Manual coding of AU intensity is extremely time-consuming and labor intensive!

AU Intensity

[FACS (Ekman et al. '02)]



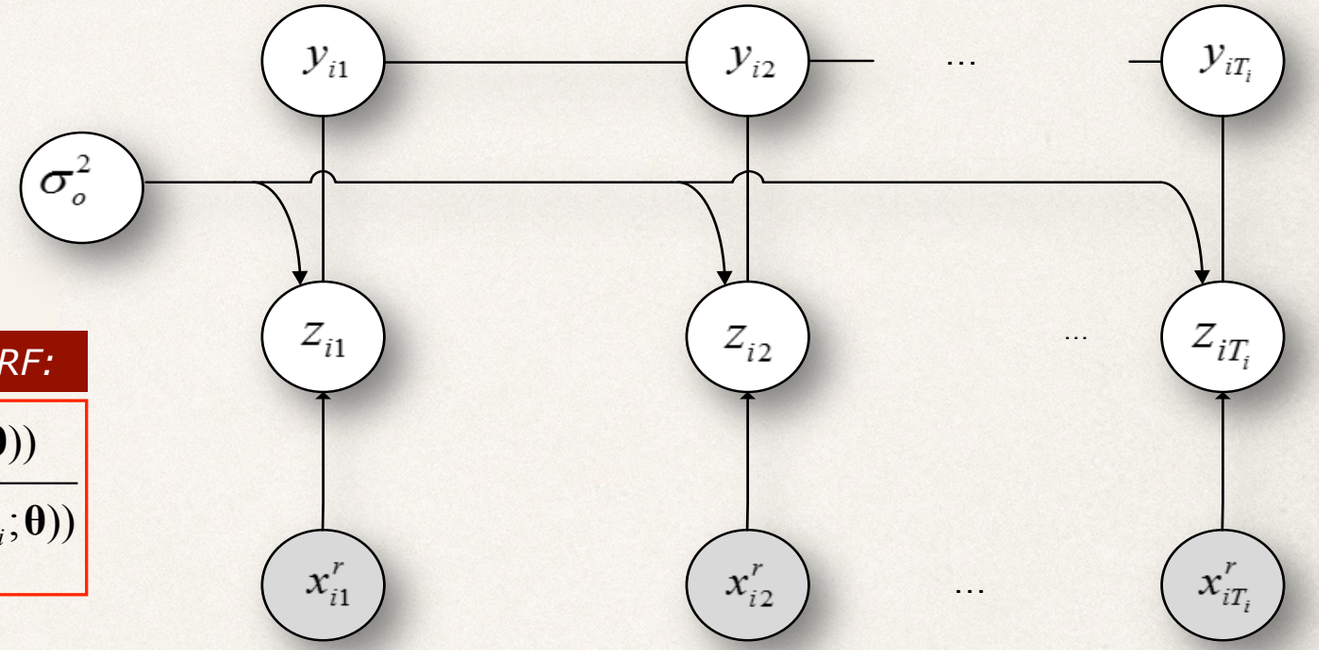
Imperial College London

Facial Action Units

❖ AU Intensity: Conditional Ordinal Random Fields (CORF)

$$\mathbf{x}_i = \{x_{i1}, \ldots, x_{iT_i}\}$$

$$\mathbf{y}_i = \{y_{i1}, \ldots, y_{iT_i}\}$$



Conditional likelihood of Linear-chain CRF:

$$P(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \frac{\exp\left(\sum_{j=1}^{T_i} \Psi(y_{i,j-1}, y_{ij}, x_i; \boldsymbol{\theta})\right)}{\sum_{\bar{\mathbf{y}} \in \Delta^{|T_i|}} \exp\left(\sum_{j=1}^{T_i} \Psi(\bar{y}_{i,j-1}, \bar{y}_{ij}, x_i; \boldsymbol{\theta})\right)}$$

$$\Psi_{ij}(y) = f_n(y_{ij}, x_i) + f_n(y_{i,j-1}, y_{ij})$$

Ordinal Node Potential

$$\sum_{k=1}^{K} I(y_{ij} = k) \log P(y_{ij} = k \mid z_{ij})$$

Edge Potential (1st order)

$$\sum_{m,k=1}^{K} I(y_{ij} = m \wedge y_{i,j-1} = k) u_{mk}$$

Learning cost: $\min_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) - \sum_{i=1}^{N} \log P(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta})$

Inference: $\mathbf{y}^* = \arg\max_{\mathbf{y} \in \Delta^{|T|}} P(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta})$

Imperial College London

Facial Action Units - Methodology

- AU detection: Variable-state Latent Variable Model (VSL-CRF)

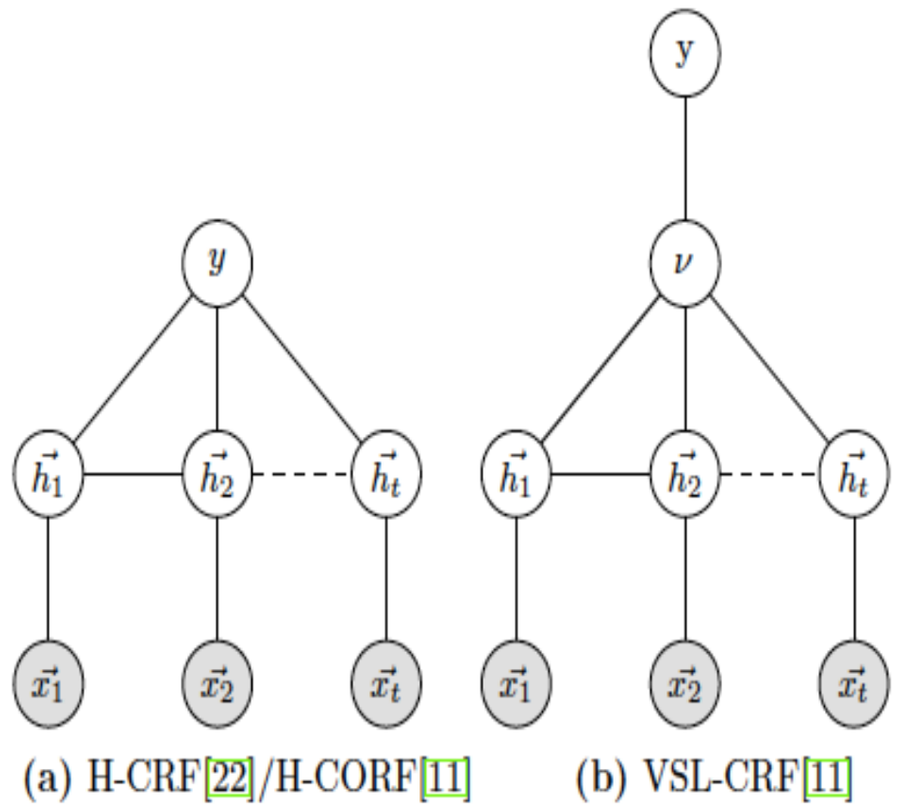
$$s(y, \mathbf{x}, \mathbf{h}, \nu; \Omega) = \begin{cases} \sum_{k=1}^{K} I(k = y) \cdot s(\mathbf{x}, \mathbf{h}; \theta_y^n), & \text{if } \nu_y = 0 \quad \text{(nominal)} \\ \sum_{k=1}^{K} I(k = y) \cdot s(\mathbf{x}, \mathbf{h}; \theta_y^o), & \text{if } \nu_y = 1 \quad \text{(ordinal)} \end{cases}$$

Marginal conditional probability of VSL-CRFs

$$P(y|\mathbf{x}, \Omega) = \frac{\max_{\nu}(\sum_{h} \exp(s(y, \mathbf{x}, \mathbf{h}, \nu, \Omega)))}{Z(\mathbf{x})}$$

$$Z(\mathbf{x}) = \sum_{k} Z_k(\mathbf{x}) = \sum_{k} \max_{\nu}(\sum_{h} \exp(s(k, \mathbf{x}, \mathbf{h}, \nu))) \text{ and } \Omega = \{\theta_k^n, \theta_k^o\}_{k=1}^{K}$$

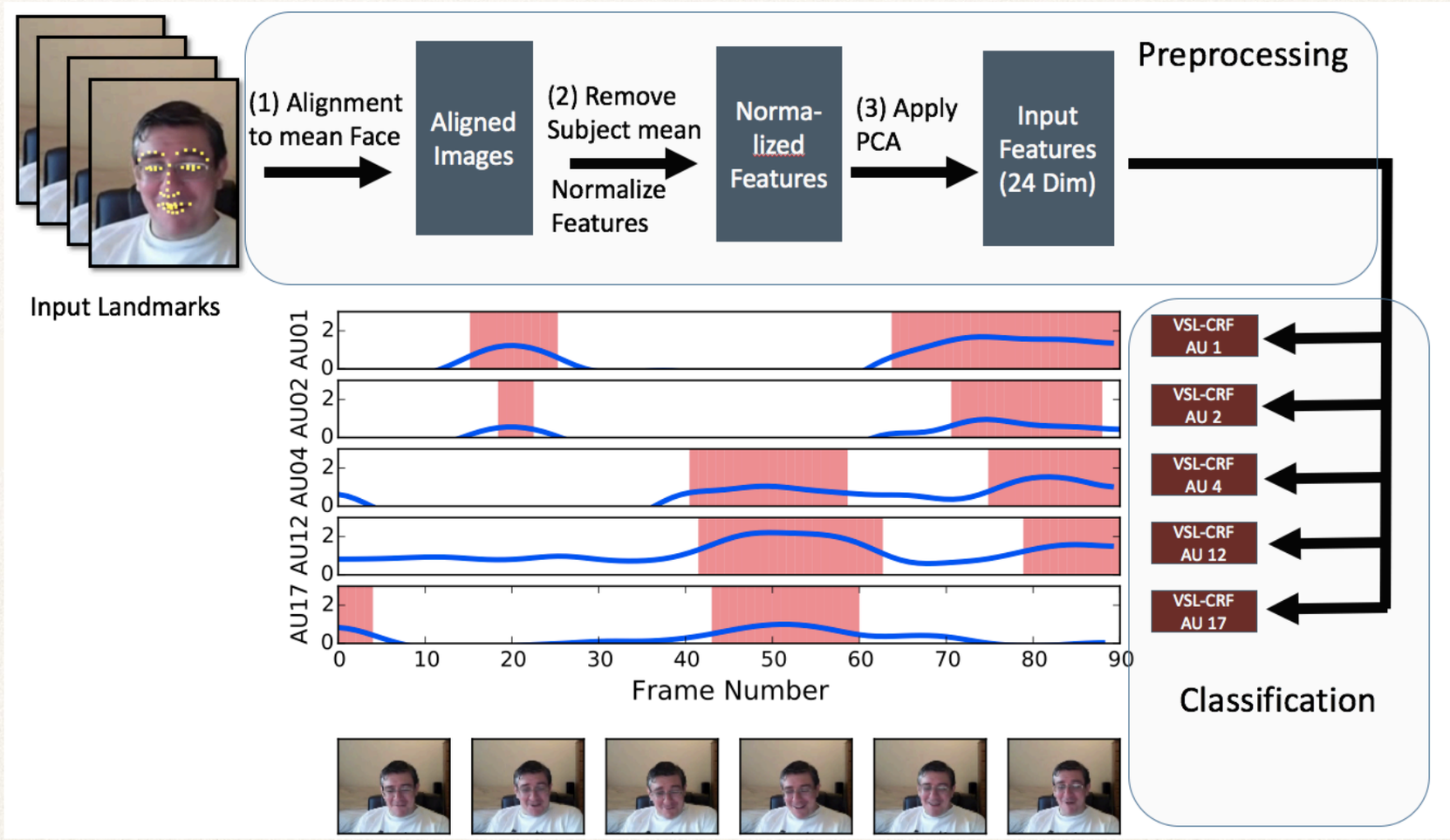
Prediction: $y^* = \operatorname{argmax}_y P(y|\mathbf{x}^*)$



(a) H-CRF[22]/H-CORF[11] (b) VSL-CRF[11]

Imperial College London

Facial Action Units– Software Impl.



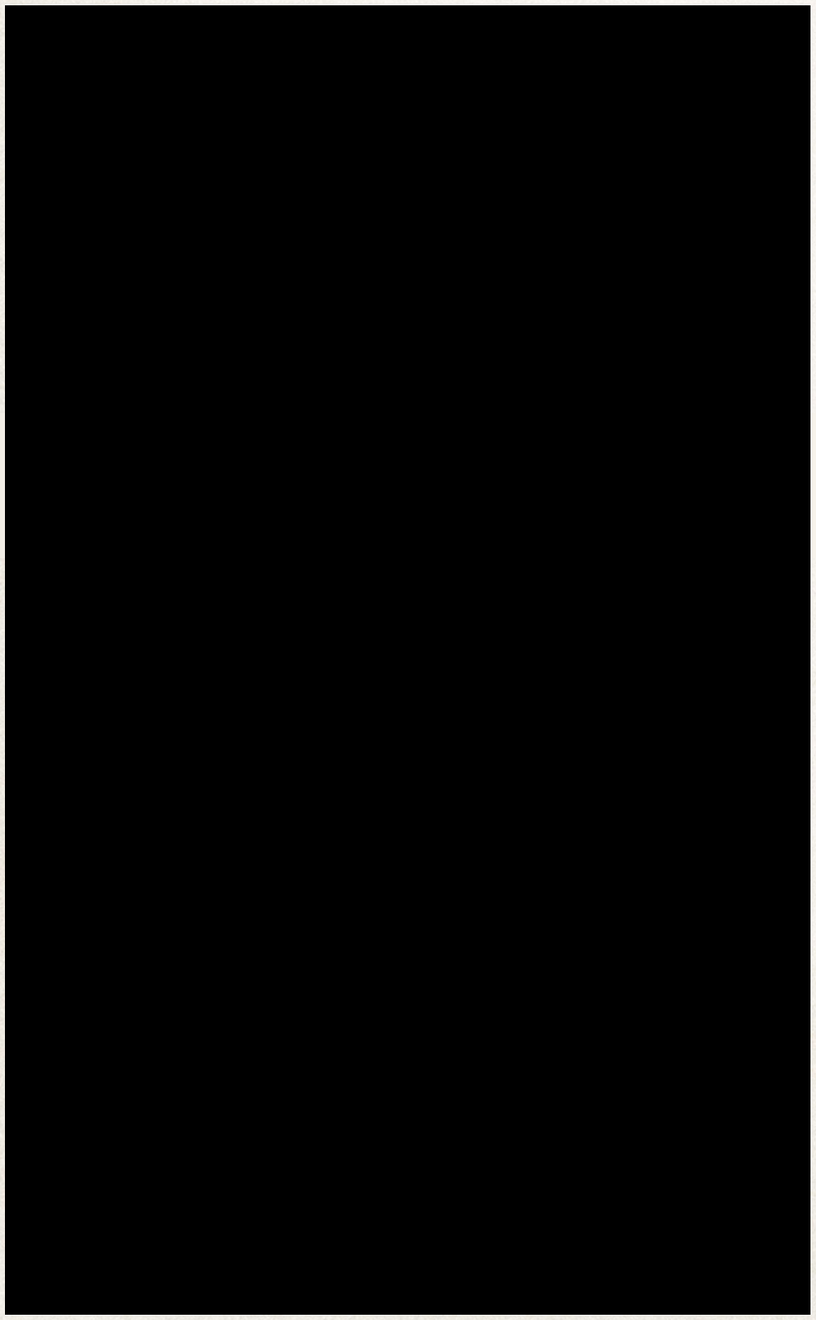
AU detection from SEWA videos

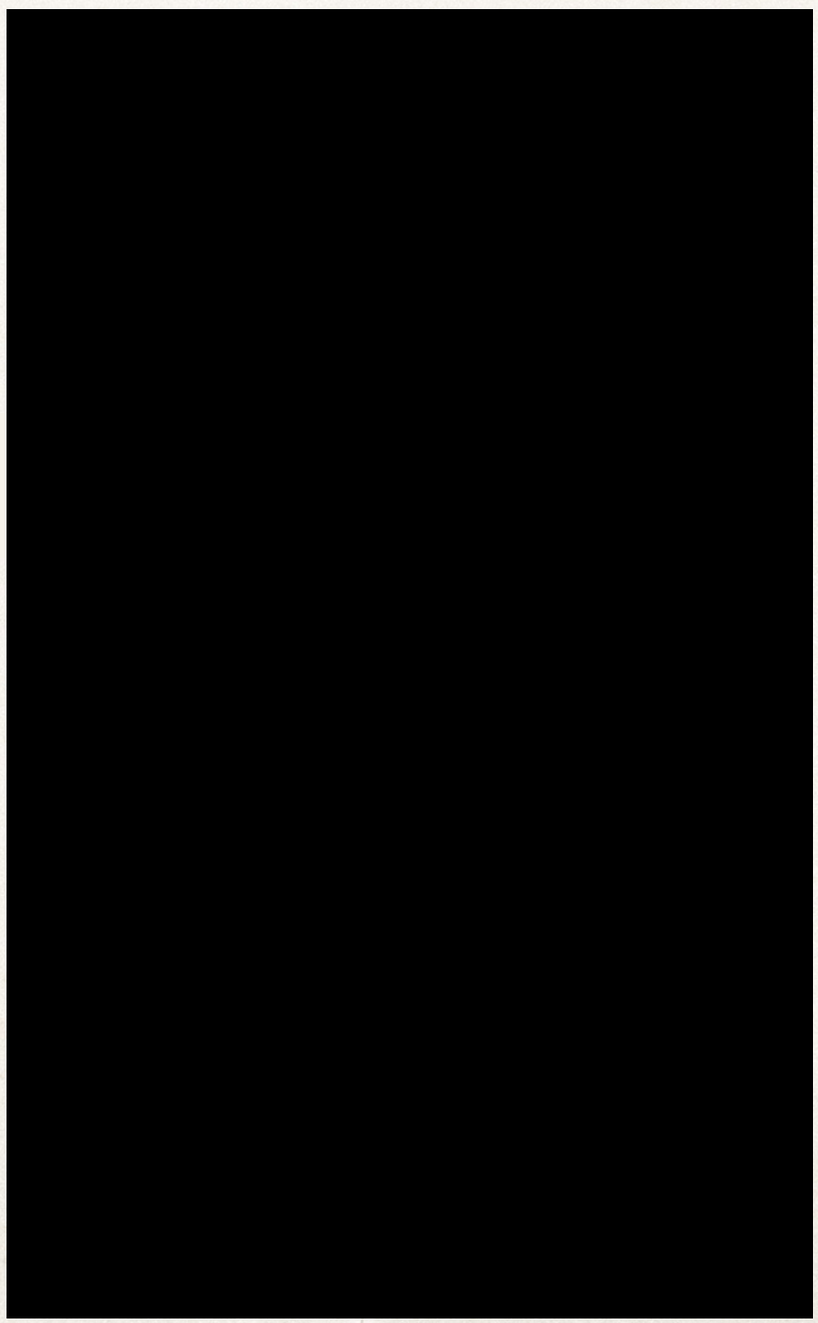
* AU detection pipeline. For each frame of the sequence, the facial points are (1) aligned to the mean face, (2) the median value of the subject is removed and (3) the dimensionality of the feature vector is reduced. The resulting sequential data is then classified using the VSL-CRF model.

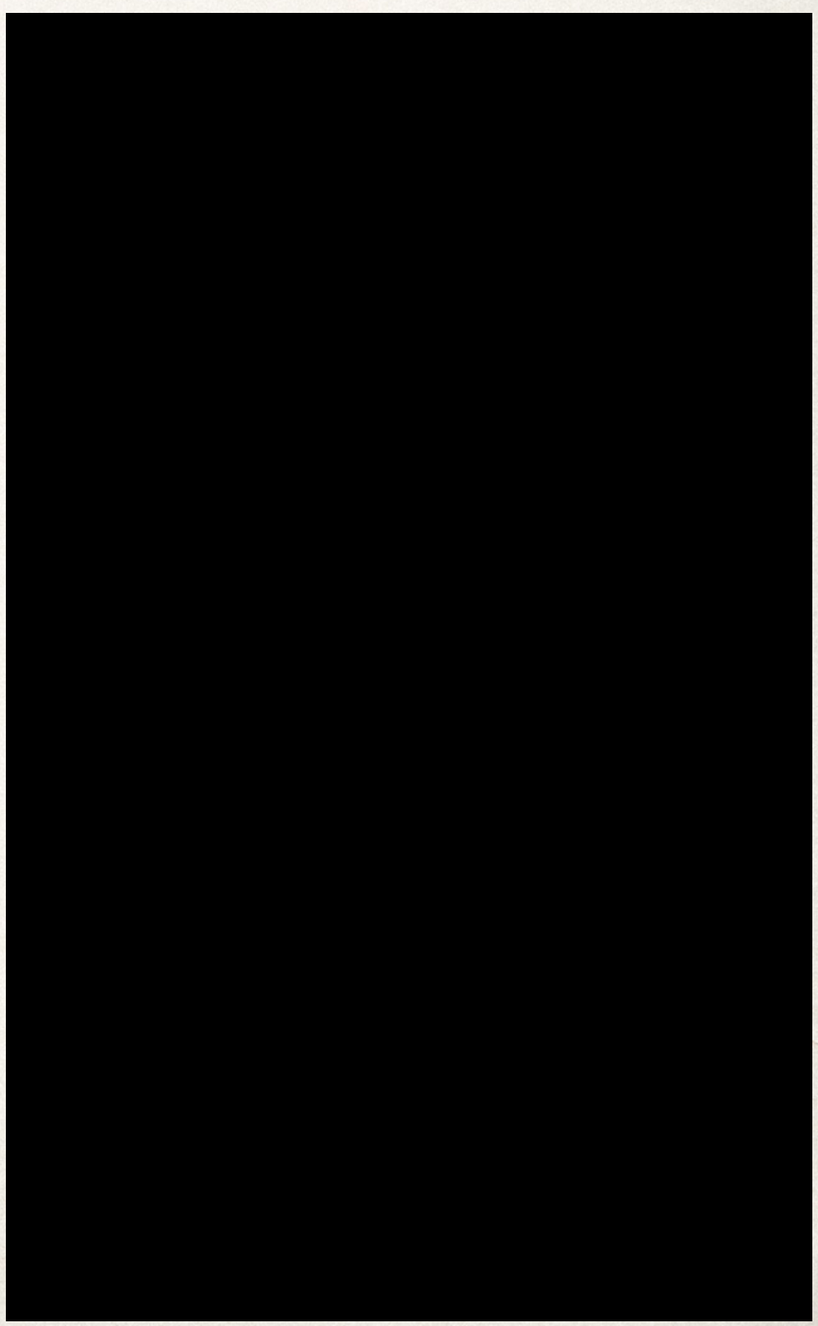
Imperial College London

Automatic Sentiment Analysis in the Wild

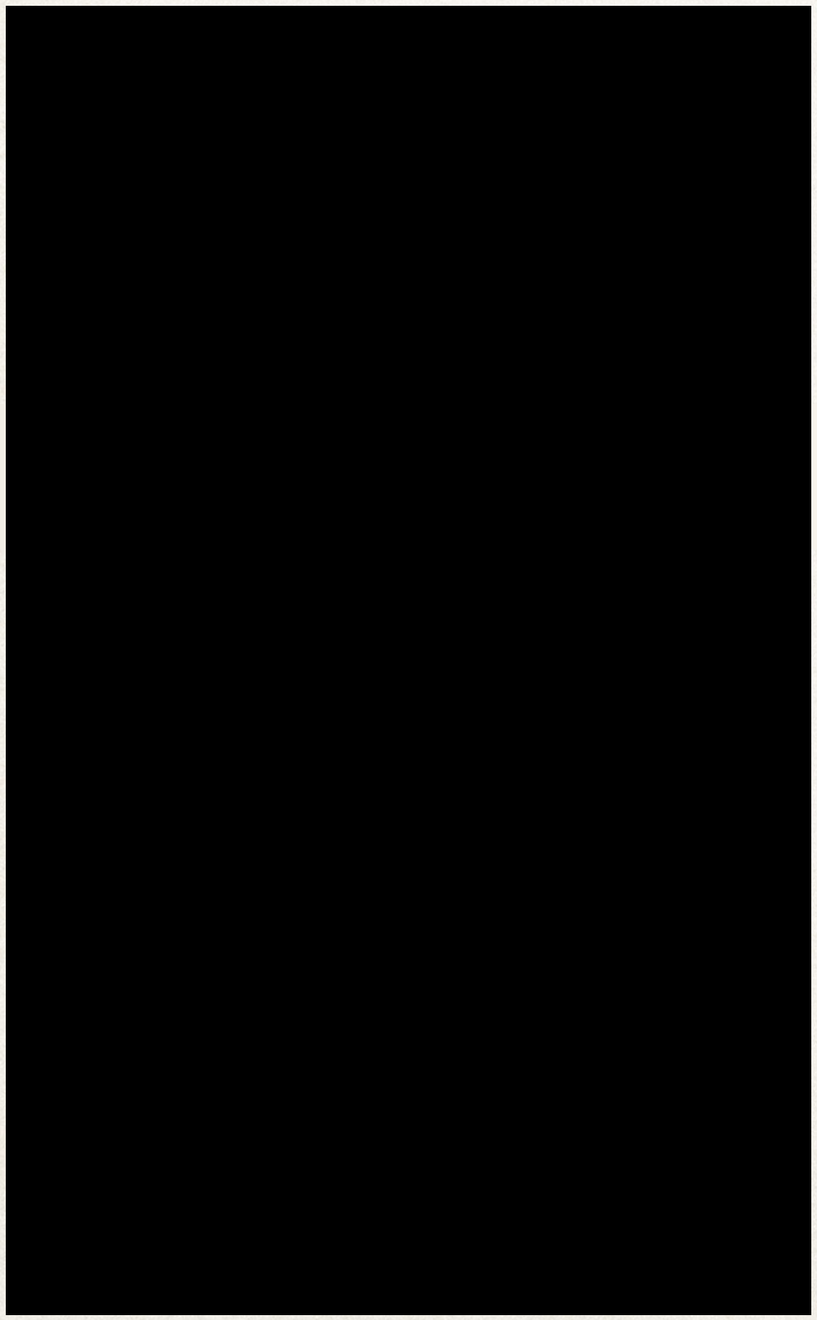
Facial Action Units - Experiments

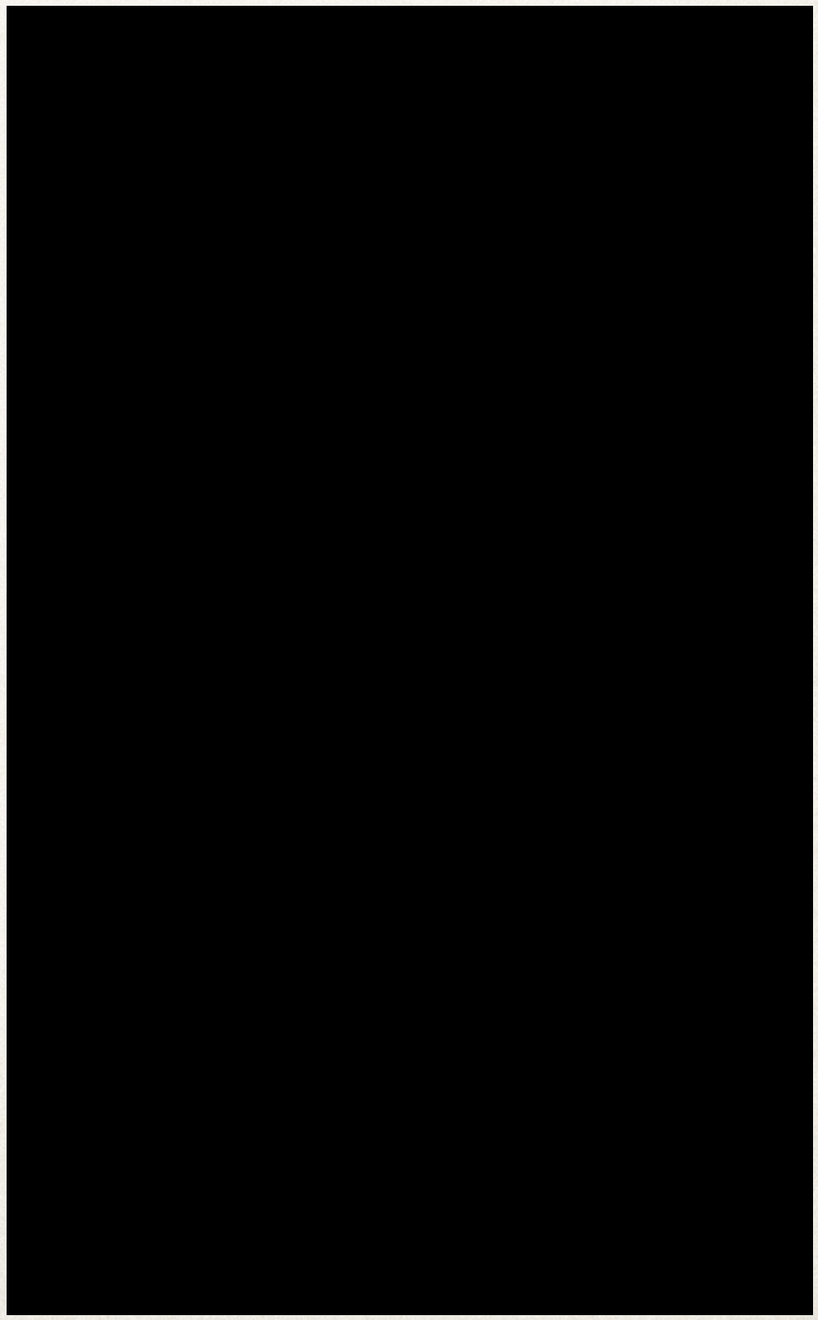






Facial Action Units - Experiments





Facial Action Units - Experiments

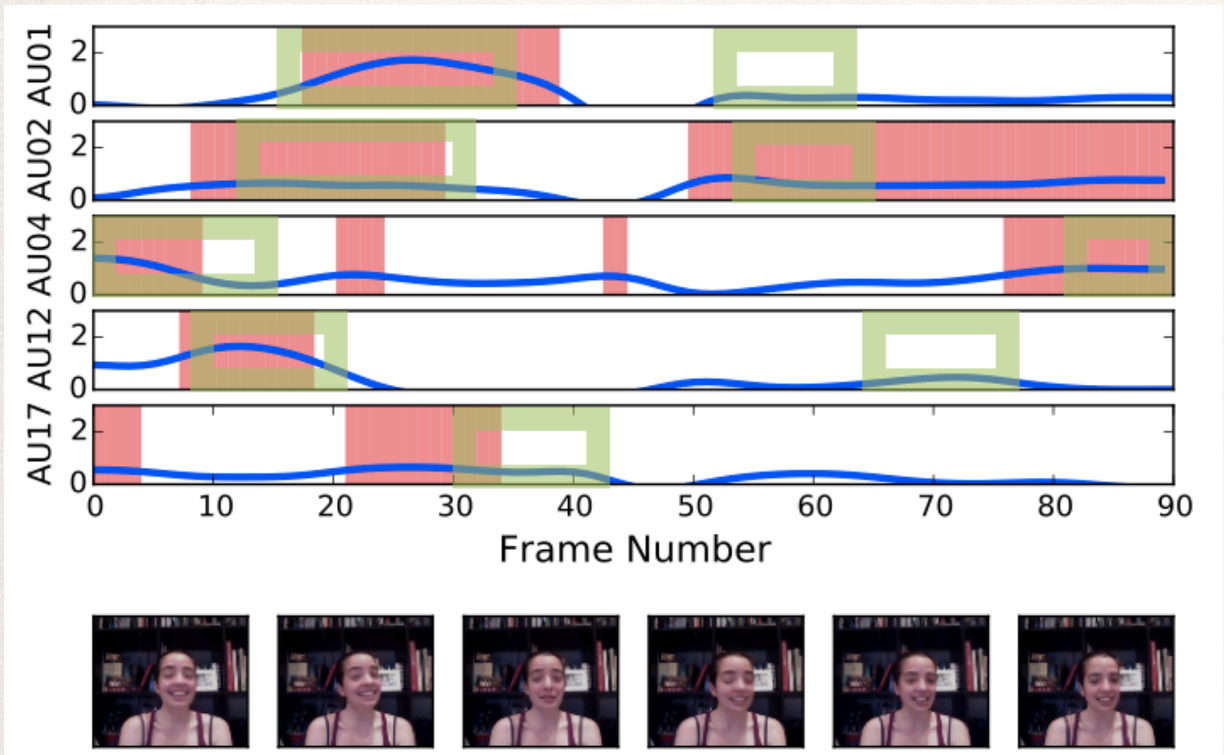
Mod.	AU1	AU2	AU4	AU12	AU17	av.
SVM	55.7	60.1	**53.5**	52.9	56.8	55.8
HCRF [22]	54.1	57.6	43.1	54.4	47.3	51.3
HCORF [11]	57.1	**65.7**	51.4	55.9	52.4	56.5
VSL-CRF	**61.4**	64.5	53.1	**56.2**	**56.3**	**58.3**

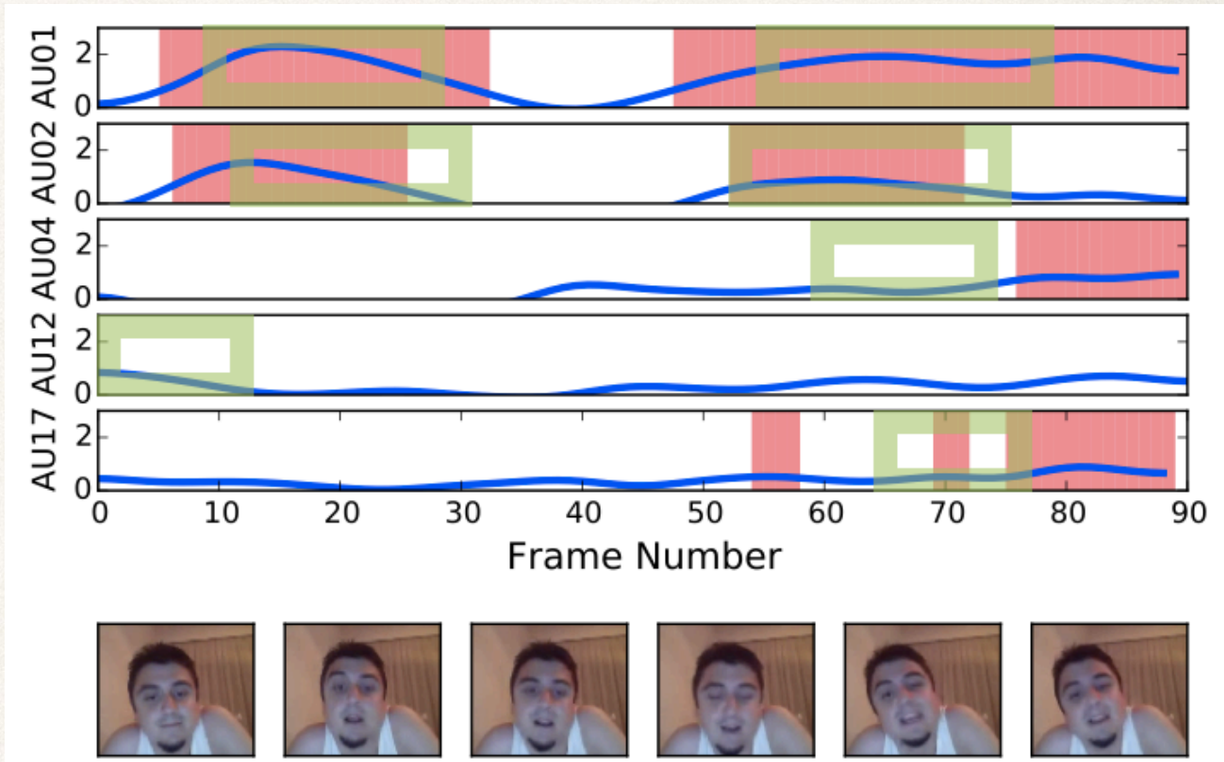
F-1 score for AU detection from SEWA videos

Imperial College London

Automatic Sentiment Analysis in the Wild

Facial Action Units - Experiments





AU detection from SEWA videos: Qualitative results

*The blue line depicts the (continuous) score by the VSL-CRF model for detection of the target AU, depicted in red. The ground truth for AU activations in target sequences is depicted in green.

Imperial College London

Objectives

- Automatic detection of head and hand gestures (D3.1)
- Facial Action Unit detection and intensity estimation (D3.1)
- Audiovisual detection of non-verbal vocalisations (D3.2)

Imperial College London

UNIVERSITÄT PASSAU

real eyes

PLAYGEN

WP3:
Mid-level Feature Extraction

Ognjen Rudovic