# Identifying Depression Without Lexical Content

presented by Helen Vernon and Anna Khlyzova

30.01.2018

# Contents

# Definitions

# Definitions

**Depression:**

- The exact causes of depression are not universally agreed upon
- Can be from interactions between a genetic predisposition and environmental factors, e.g. stress and emotional trauma (Nestler et al., 2002)
- Considered an illness when an individual has either a depressed mood or disinterested in their interests in combination with four or more symptoms for longer than a two-week period
- There are at least 1,497 individual profiles
- Hamilton rating scale can distinguish depressed and non depressed people

# Definitions

**Suicidality**:

An individual who is having a suicidal crisis is at imminent risk (minutes, hours, days) of attempting suicide.’ (Florentine and Crane, 2010)

- Small window - attempts are contemplated from 5 mins - 1 h.
  - idea -->non-fatal attempt-->fatal attempt
- Range of intense affective states such as desperation, extreme hopelessness, feelings of abandonment, self-hatred, rage, anxiety, loneliness and guilt (Hendin et al., 2007)
- It is not understood why people attempt suicide
- It is a behaviour, not an illness
- link with ‘non-lethal suicide behaviours’

# Difficulties

- **Depression is notoriously difficult to diagnose!**
- Is the person even diagnosed?
- Two people with no overlapping symptoms can receive the same diagnosis
- Absence of biological markers
- Depression is gradient
- Limited corpora
- Inconsistencies within and across corpora
- Medication
- **The search for an objective marker**

# Classifiers

# Classification and Prediction Systems

- 3 types of system:
  - Presence
    - Detection problem
    - 2 classes
    - Reported in terms of accuracy

# Classification and Prediction Systems

- 3 types of system:
  - Severity
    - 2 or more classes
    - Categorical assignment
    - Related to a scale of depression e.g HAMD scale
    - Reported in terms of accuracy

# Classification and Prediction Systems

- 3 types of system:
  - Level Prediction
    - Assigns a given item of speech input to a continuous scale related to mental state assessment.
    - Assessed in terms of the difference between prediction and observation
    - reported as **root square mean error** or **absolute mean error**

# Classification and Prediction Systems

- Models are typically Support Vector Machine or Gaussian Mixture Models
  - Good with sparse datasets
  - Lack of computational expense
- Limitations due to data sparsity
- Current models are not comparable
  - Different datasets
  - Variability within each dataset
  - Different experimental conditions within the datasets

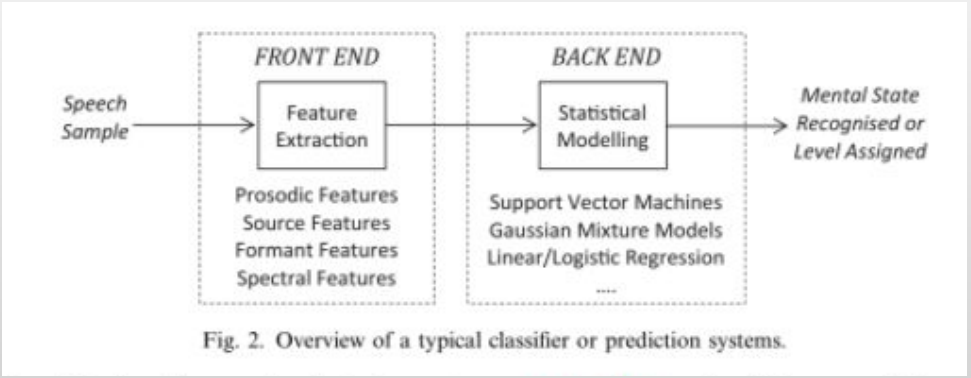


Fig. 2. Overview of a typical classifier or prediction systems.

# Corpora and their Limitations

# Corpora

| 1st Published (Name) | Subjects | Clinical scores | Vocal exercises | Read speech | Free response or interview | Free speech | Additional notes | Other references |
|---|---|---|---|---|---|---|---|---|
| France et al. (2000) Vanderbilt II Study | 115: 59 DPRD (21M, 38F) 22 SCDL (all M) 34 NTRL (24M, 10F) | DSM-IV BDI (DPRS = BDI > 20) | | | √ | √ | *Recorded therapy sessions or suicide notes* Mean file length: *2 min 30 s* Age range: 25–65 Medications Present: *Imipramine-hydrochloride* | *Similar corpus used in:* Ozdas et al. (2004a,b, 2000) and Hashim et al. (2012) |
| Moore et al. (2004) | 33: 15 DPRD (6M, 9F) 18 NTRL (9M, 9F) | DSM-IV | | √ | | | Utterances per speaker: 65 Mean file length: *3 min* Age range: 19-57 | Moore et al. (2008) *Similar corpus used in:* Moore et al. (2003) |
| Yingthawornsuk et al. (2006) | 32(all M): 10 SCDL 13 DPRD 9 Remitted Patients | BDI (DPRD = BDI > 20) | | √ | √ | | Mean file length: *Free Response – 8 min Read Speech – 2 min* Age range: 25–65 | *Similar corpus used in:* Keskinpala et al. (2007), Landau et al. (2007), Yingthawornsuk et al. (2007), and Hashim et al. (2012) |
| Mundt et al. (2007) | 35: DPRD (15M, 20F) | HAMD Mean: 14.9 ± 6.3 Range: 3–27 QIDS Mean: 12.4 ± 6.1 Range: 0–26 | √ | √ | √ | | Mean age: 41.8 Medications: *Range present* | Sturim et al. (2011), Trevino et al. (2011), Quatieri and Malyska (2012), Cummins et al. (2013a,b), and Helfer et al. (2013) |
| Cohn et al. (2009) | 57: DPRD (24M, 34F) | DSM-IV HAMD (HAMD ≥ 5) | | | √ | | Min. vocalisation per speaker: *100 s* Mean age: 39.7 Age range: 19–65 Medications: *SSRI's present* | *Extended version:* Yang et al. (2012) |
| Low et al. (2009) | 139: 68 DPRD (49F, 19M) 71 NTRL (71M, 44F) | N/A | | | √ | √ | Recordings per subject: 3 Mean file length: *20 min* Age range: 12–19 | Memon et al. (2009), Low et al. (2011, 2010), and Ooi et al. (2013, 2012) |
| Alghowinem et al. (2012) | 80: 40 DPRD 40 NTRL | DSM-IV | | √ | √ | | Mean file length: *40 min* | Alghowinem et al. (2013a,b), and Cummins et al. (2013b) *Subset published in:* Cummins et al. (2011) |
| Mundt et al. (2012) | 165: All DPRD (61M, 104F) | DSM-IV HAMD QIDS | √ | √ | √ | | Age range: 21–75 Mean age: 37.8 Medications: *Sertraline* | None |
| Scherer et al. (2013a) | 60: 30 SCDL 30 NTRL | C-SSRS SIQ-Jr version | | | √ | | Age range: 13–17 | None |
| Scherer et al. (2013c) *Distress Assessment Interview Corpus* | 110: 29% DPRD 32% PSTD 62% Anxiety | PHQ-9 (DPRD = PHQ-9 > 10) | | | √ | | Data per participant: *30–60 min* Age range: 18–65 | Scherer et al. (2013b,d) |
| *Audio-Visual Depressive Language Corpus (AVID Corpus)* | files each containing a range of mix of vocal exercises, free and read speech tasks | *Training Set: 15.1 ± 12.3 Development Set: 14.8 ± 11.8* | | | | | Language Mean file length: *25 min* Age range: 18–63 Mean age: *31.5 ± 12.3* | et al. (2014a,b, 2013c), Kaya and Salah (2014), Kaya et al. (2014b), and Williamson et al. (2013) |
| | *AVEC 2014:* 150 files each containing a read speech passage (Die Sonne und der Wind) and an answer to a free response question. *Note* AVEC 2014 is a shortened (file length) version of AVEC 2013. 5 files were replaced in 2014 due to unsuitable data | *Mean AVEC 2014 Training Set: 15.0 ± 12.3 Development Set: 15.6 ± 12.0* | | | | | | *AVEC 2014 Papers:* Valstar et al. (2014), Gupta et al. (2014), Kaya et al. (2014a), Mitra et al. (2014), Pérez et al. (2014), Senoussaoui et al. (2014), Sidorov and Minker (2014), and Williamson et al. (2014). *Similar corpus used in:* (Hönig et al., 2014) – 1122 recordings taken from 219 speakers in AVDL Corpus |

# Corpora

| 1st Published | Subjects | Clinical Scores | Vocal Exercises | Read Speech | Free response/ Interview | Free Speech | Additional Notes |
|---|---|---|---|---|---|---|---|
| Moore et al 2004 | 33: 15 depr, 6M, 9F, 18 Neutr, 9M, 9F | DSM-IV | | X | | | Utterances per speaker: 65, Mean file length: 3 min, Age Range: 19-57 |
| Mundt et al 2007 | 35: depr, 15M, 20F, | HAMD Mean: 14.9 | X | X | X | | Mean age: 41.8 Medications: range present |
| Scherer et al 2013a | 60: 30 Suic, 30 Neutr | C-SSRS | | X | | | Age range: 13-17 |

# Speech Markers

# Feature Extraction

- Short-term time scale - overlapping frames of 10 - 40 ms in length
- Temporal information and long-term information - utterance level statistics/functionals or frame-based delta (D) and delta–delta (DD) coefficients, reflecting differences between the adjacent vectors' feature coefficients.
- Prosody - rhythm, stress, intonation
- Features - speaking rate, pitch, loudness, energy dynamics

# Speech Markers

- “patients speak in a low voice, slowly, hesitatingly, monotonously, sometimes stuttering, whispering, try several times before they bring out a word, become mute in the middle of a sentence” (Kraepelin, 1921).
- Cognitive and physiological changes:
  - Speech planning
  - Motor Coordination
  - Psychomotor Impairment
  - Correlated with severity and pause measurements
- Increase in muscle tension
  - Voice Quality
  - Prosody

# Speech Markers

- **Prosodic Features:**
  - Reduced pitch
  - Reduced pitch range
  - Slower speaking rate
  - Articulation errors
  - Lack of linguistic stress

# Speech Markers

- **Source Filters:**
- Results in this area are very inconsistent
  - **Jitter:** small cycle-cycle variations in pulse timing
  - **Shimmer:** cycle-cycle variations in pulse amplitude
  - **Harmonic - noise ratio:** ratio of harmonics to non harmonics in the speech signal
- Correlate strongly with depression and psychomotor impairment

# Speech Markers

- **Formant Features:**
- Decrease in $F^0$ correlates with levels of depression
- The phoneme /aɪ/ has the second formant location reduced
  - Low back → high front movement
  - Tongue moves more slowly
- Vocal tract becomes damp
  - Narrowing of bandwidth
  - 1st and 3rd formants show significant difference between depressed and control

# Speech Markers

- **Speech rate:**
  - **Best feature!!!**
  - Significantly correlated with Hamilton Rating Scale
  - Temporal features:
    - Total speech time
    - Total pause time
    - % pause time
    - Speech pause ratio
    - Speaking rate
    - Stronger relationship when these features are extracted at phoneme level than at global and sentence levels
    - Groups of phonemes correlated with specific sub symptoms

# An Example Study

# Trevino et al 2011

- Aims for the early diagnosis of major depressive disorder
- 35 speakers, free response questions
  - Data designed for depression severity
  - Clinic and phone call recordings
  - HAMD scores
- Hidden Markov Model for phoneme recognition

# Trevino et al 2011

- Global level measurements
  - Speech units per second
  - Speaking rate (over the whole session, including pauses)
  - Phone rate (over speaking time only)
- Phoneme level measurements
  - Duration of individual phones
  - Phoneme specific relationships

# Trevino et al 2011

**Table 1 Score correlations with speaking and articulation rate**

| Rate Measure | Score Category | Spearman Correlation | p-value |
|---|---|---|---|
| Speaking-Phone Rate | HAMD Work and Activities | -0.20 | 0.01 < p < 0.05 |
| | *HAMD Psychomotor Retardation* | *-0.38* | *p = 3.6e-5* |
| | HAMD TOTAL | -0.22 | 0.01 < p < 0.05 |
| Articulation-Phone Rate | *HAMD Psychomotor Retardation* | *-0.46* | *p = 3.2e-7* |
| | HAMD Weight Loss | -0.19 | 0.01 < p < 0.05 |

Italic values indicate cases of high significance with $p < 0.01$.

**Table 3 Score correlations with signed aggregate phone length**

| *Phones used* | *Score Category* | *Spearman Correlation* | *p-value* |
|---|---|---|---|
| (sil, aa, g, jh, k, ng, s, t) | HAMD Mood | 0.43 | p = 2.7e-6 |
| (uh, b, jh, n, p, t, z) | HAMD Insomnia Middle of the Night | 0.37 | p = 6.8e-5 |
| (sil, aa, ih, ow, eh, s) | HAMD Work and Activities | 0.39 | p = 2.7e-5 |
| (sil, ae, iy, ay, ey, ao, ow, eh, aw, uh, er, g, k, ng, r, s, t, v, w, z) | HAMD Psychomotor Retardation | 0.58 | p = 1.7e-11 |
| (aw, jh, p, t) | HAMD Agitation | 0.34 | p = 2.0e-4 |
| (aa, uw, uh, b) | HAMD General Symptoms | 0.40 | p = 1.4e-5 |
| (aa, ao, s, w) | HAMD Genital Symptoms | 0.42 | p = 4.5e-6 |
| (sil, ao, g, n, ng, s) | HAMD Hypochondriasis | 0.39 | p = 2.0e-5 |
| (iy, ey, ih, eh, f, l, v) | HAMD Weight Loss | 0.39 | p = 2.6e-5 |
| (sil, s, k, ih, aa) | HAMD TOTAL | 0.35 | p = 1.8e-4 |

**Table 2 Score correlations with pause features**

| Measure | Score Category | Spearman Correlation | p-value |
|---|---|---|---|
| Pause Length | *HAMD Mood* | *0.28* | *p = 0.003* |
| | HAMD Guilt | 0.20 | 0.01 < p < 0.05 |
| | *HAMD Suicide* | *0.27* | *p = 0.004* |
| | *HAMD Work and Activities* | *0.28* | *p = 0.002* |
| | *HAMD Psychomotor Retardation* | *0.33* | *p = 0.0003* |
| | *HAMD Anxiety Psychic* | *0.24* | *p = 0.009* |
| | *HAMD Hypochondriasis* | *0.26* | *p = 0.005* |
| | *HAMD TOTAL* | *0.26* | *p = 0.005* |
| Ratio of Pause Time | HAMD Guilt | 0.21 | 0.01 < p < 0.05 |
| | HAMD Insomnia Early Morning | 0.20 | 0.01 < p < 0.05 |
| | HAMD Work and Activities | 0.19 | 0.01 < p < 0.05 |
| | HAMD Anxiety Psychic | 0.24 | 0.01 < p < 0.05 |
| | *HAMD TOTAL* | *0.25* | *p = 0.009* |

Pauses are identified by the phone recognizer; the average of all durations per session is used as the feature. Italic values indicate cases of high significance with $p < 0.01$.

# Limitations In This Area

- The classifiers used weren’t originally intended for this task
  - Front and back design too simplistic
  - Need to be multimodal including features from outside speech
- Confounding factors
  - Medication
  - Other illnesses/conditions with similar effects e.g. Parkinson's Disease

# Limitations In This Area

- Corpora
  - Too small
  - Features are ‘diluted’ by personal demographic information
    - Variability from age, gender, dialect, first language, emotion etc not considered
    - Sensitive corpora shared as a subset - features and measurements
    - Difficulties in data collection means that results are not generalisable
    - Some features depending upon content of speech
    - Lots of variability in recordings
    - $F^0$ in depressed people can be the same as for non-depressed people:
      - Underlying mood
      - Agitation and anxiety
      - Personality
    - Is the corpus even accurate? It is hard to diagnose after all!

# Limitations In This Area

- Muscle tension increase causes reduced pitch variation BUT should increase $F^0$
  - The average $F^0$ decreases.
- While there are some trends, many are not statistically significant
- No uniform indicator of depressed i.e. **no objective marker!**

**Maybe multimodality would help…….**

# Visual Indicators of Depression

# Intro

**Questions**:

1) How is nonverbal behavior related to the severity of depression? What parts should be indicative of the depressed state? (Girard et al., 2014)
2) Can body expressions contribute to automatic depression analysis? (Joshi et al., 2013)

**Motivation**: to support clinicians in the diagnosis of mental health disorders

# Two hypotheses

**The Affective Dysregulation hypothesis**

↓ positivity, ↑ negativity

**The Emotion Context Insensitivity hypothesis**

↓ positivity, ↓ negativity

Agree: in reduction of positive expressions
Disagree: excessive negativity or reduced?

# Limitations of Previous Works:

1) depressed participants vs. non-depressed controls ⇨ depression or stable personality traits?
2) viewing emo stimuli while alone ⇨ many behaviors are rare
3) limited range of nonverbal behavior: many expressions → 1 category or single expression → 1 category?

Joshi et al.:

patients against controls

# New hypothesis: Social Withdrawal hypothesis

- interprets depression in terms of affiliation
- ↓ affiliative behavior, ↑ non-affiliative

What's <u>new</u> in this study?

- ❑ control for stable personality traits
- ❑ head motion + facial movements
- ❑ manual and automatic coding measurement
- ❑ FACS (4 facial expressions)

<u>Joshi et al.:</u>

\+ upper body movements (incl. head and face movements)

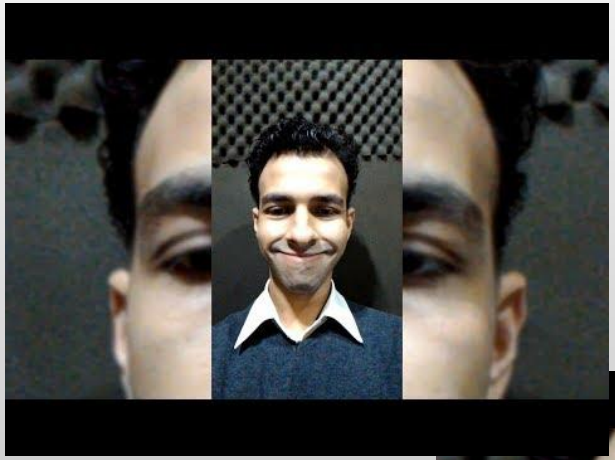
<u>Joshi et al.:</u>

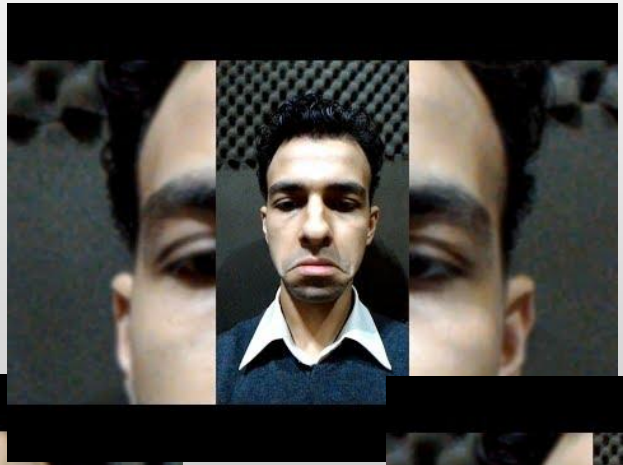
only automatic

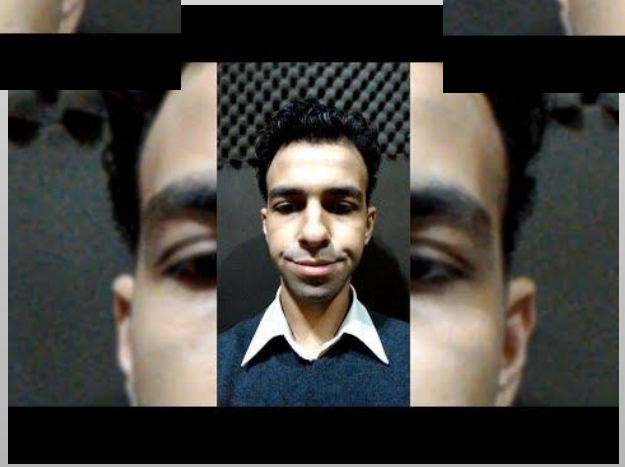
<u>Joshi et al.:</u>

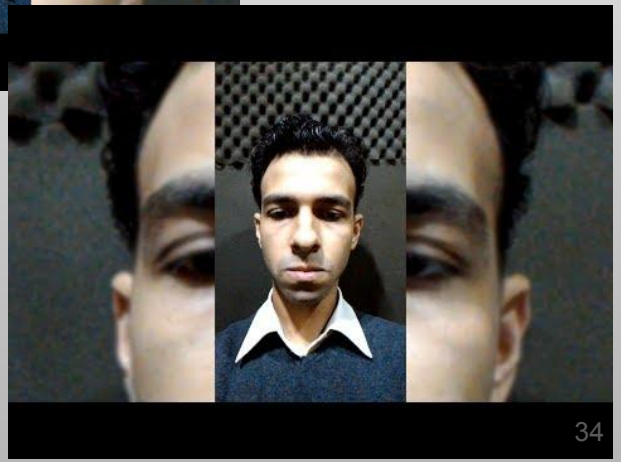
Viola-Jones object detector (common/not common)

# FACS expressions









# In case videos don’t work:

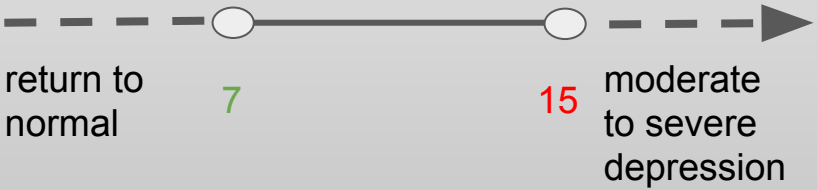
# 3 hypotheses compared

| | **The Affective Dysregulation hypothesis** | **The Emotion Context Insensitivity hypothesis** | **Social Withdrawal hypothesis** |
|---|---|---|---|
| AU12 |  |  |  |
| AU14 |  |  |  |
| AU15 |  |  |  |
| AU24 |  |  |  |

# Methods

# Participants

- 33 adults from a clinical trial
- evaluation of symptoms severity at 1,7,13,and 21 weeks
- Hamilton rating scale




- 4 cameras
- analysis of only first 3 questions (segments of 28-242 sec)

- FACS trained on data from all 33 participants ⇨ but analysis is only on improved ones
- 38 interviews from 19 participants

Participant groups and demographic data

| | Database | Responders |
|---|---|---|
| Number of Subjects | 33 | 19 |
| Number of Sessions | 69 | 38 |
| Average Age (years) | 42.3 | 42.1 |
| Gender (% female) | 66.7% | 63.2% |
| Ethnicity (% white) | 87.9% | 89.5% |
| Medicated (% SSRI) | 45.5% | 42.1% |

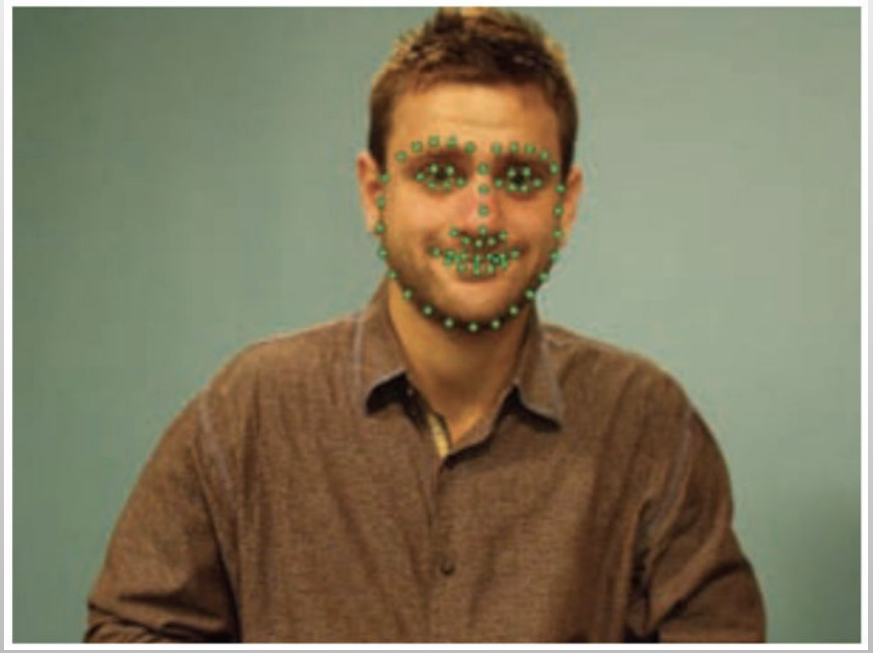
# Manual FACS coding

- gold standard
- decomposition of facial expressions into AUs
- coding by certified and experienced coders
- Cohen’s Kappa = 0.75 ⇨ good

# Automatic FACS coding

Face registration

- landmark points
- used Active Appearance Models to track 66 points
- to train AMMs, ≅3% of video frames manually annotated



# Automatic FACS coding

Feature extraction

- facial expressions: changes in shape and appearance
- Localized Gabor features: 40 filters applied around 66 l.p.
  = 2640 per frame

Dimensionality reduction:

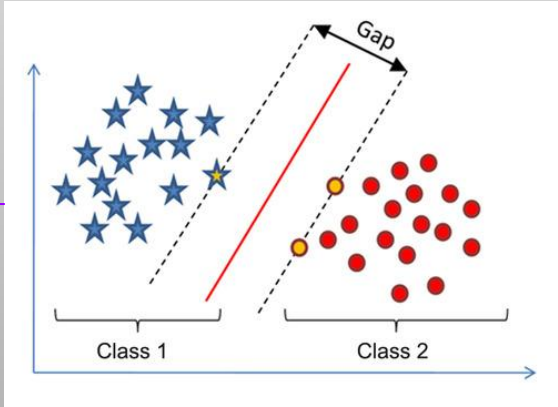
- too many features ⇨ hard to classify
- manifold learning technique to reduce them
- result: 29 features per frame

used for recognition of shapes of objects

# Automatic FACS coding

Classifier training

- SVM
- to build a training set: randomly sampled pos and neg frames
- to train and test classifiers: LIBSVM library
- to find the best classifier: “grid-search” during leave-one-out cross-validation

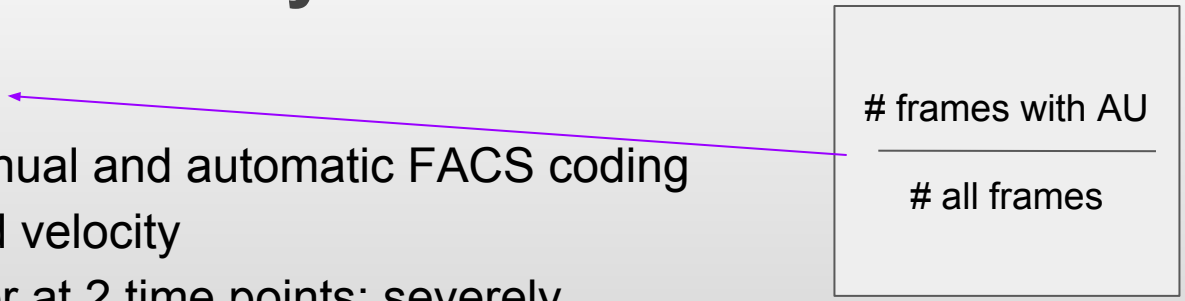


# Automatic Head Pose Analysis

- cylinder-based 3D head tracker
- output: 6 degrees of freedom of head motion or error message ⇨ 4.61% cannot be tracked
- poor connection in 11.37% of frames (excluded)
- head angles ⇨ angular amplitude and angular velocity

# Measures and Data Analysis

- base rate for each AU
- measured for both manual and automatic FACS coding
- average amplitude and velocity
- comparison of behavior at 2 time points: severely depressed and recovered
- comparison of manual and automatic FACS coding
  1. at frame level
  2. at session level



# Results

Facial expression during high and low severity interviews, Average base rate

| | Manual Analysis | | Automatic Analysis | |
|---|---|---|---|---|
| | High Severity | Low Severity | High Severity | Low Severity |
| AU 12 | 19.1% | 39.2% * | 22.3% | 31.2% |
| AU 14 | 24.7% | 13.9% * | 27.8% | 17.0% * |
| AU 15 | 05.9% | 11.9% * | 08.5% | 16.5% * |
| AU 24 | 12.3% | 14.2% | 18.4% | 16.9% |

same

* $p<0.05$ vs. high severity interview by Wilcoxon Signed Rank test

Head motion during high and low severity interviews

| Head Motion | High Severity | Low Severity |
|---|---|---|
| Vertical Amplitude | 0.0013 | 0.0029 * |
| Vertical Velocity | 0.0001 | 0.0005 ** |
| Horizontal Amplitude | 0.0014 | 0.0034 ** |
| Horizontal Velocity | 0.0002 | 0.0005 ** |

Reliability between manual and automatic FACS coding

| Facial Expression | Frame-level | Session-level |
|---|---|---|
| AU 12 | AUC = 0.82 | ICC = 0.93 |
| AU 14 | AUC = 0.88 | ICC = 0.88 |
| AU 15 | AUC = 0.78 | ICC = 0.90 |
| AU 24 | AUC = 0.95 | ICC = 0.94 |

# Discussion

Results: ↓ AU12, ↑ AU14, ↓ AU15, AU24 - same

| | The Affective Dysregulation hypothesis | The Emotion Context Insensitivity hypothesis | Social Withdrawal hypothesis |
|---|---|---|---|
| AU12 | ↓ ✓ | ↓ ✓ | ↓ ✓ |
| AU14 | ↑ ✓ | ↓ ✗ | ↑ ✓ |
| AU15 | ↑ ✗ | ↓ ✓ | ↓ ✓ |
| AU24 | ↑ ✗ | ↓ ✗ | ↑ ✗ |

It looks like the best models will be multimodal…….so…….

YOUR TURN!

What would you include?

# References


Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10-49.

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., & Rosenwald, D. P. (2014). Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, *32*(10), 641-647.

Joshi, J., Goecke, R., Parker, G., & Breakspear, M. (2013, April). Can body expressions contribute to automatic depression analysis?. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on* (pp. 1-7). IEEE.

Trevino, A. C., Quatieri, T. F., & Malyska, N. (2011). Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, *2011*(1), 42.