**Existing head gesture detectors [see table comparison at the end]**

*Most previous works used HMM-based models …*

*The problem in this domain is the lack of annotated corpora.*

NOD / SHAKE

1. A real-time head nod and shake detector using HMMs
   1. <https://www.sciencedirect.com/science/article/pii/S0957417403000885>
   2. HMM based
   3. 82-89% acc, 85% overall
   4. Real-time
   5. changes of the eye's x-position and y-position indicate the direction of head movement
   6. Collected their own dataset: 80 samples train, 110 samples test
2. Hatice, with emotion predictions
   1. <https://ibug.doc.ic.ac.uk/media/uploads/documents/IVA-2010-GunesPantic-CAMERA.pdf>
   2. <https://ibug.doc.ic.ac.uk/resources/nod-shake-detector-and-5-dimensional-emotion-predi/>
   3. Hmm-based
   4. SEMAINE data, annotations per seq :(
   5. %
   6. Can run EXE and compare performance!!!
3. Real-time head nod and shake detection for continuous human affect recognition
   1. Realtime
   2. Nod and shake
   3. Hmm-based
   4. Kinect-based, collected own database, 150 samples, 50 nods, 50 shakes
   5. 86%
4. **[RECENT, 2018] Recognizing Visual Signatures of Spontaneous Head Gestures**
   1. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8354154>
   2. FIPCO and CCdb datasets
   3. CNN + LSTM
   4. Input: 6DOF head pose - position+velocity profiles, 16 landmarks
   5. NOT realtime - frame at time t we extract features from multiple temporal windows that extend from [t−k,t+k]
      1. K = 16, 32, 64
   6. Per-frame F1 score (detect gesture at each frame), weighted F1
   7. Data augmentations
   8. Coarser gesture categories - 5 or 7 instead of 11
      1. As THEY found
         1. Nod+tick+jerk
         2. Up+down
         3. Shake+turn
         4. Fwd+bwd
   9. [ASKED FOR CODE, GOT]
      1. <https://github.com/mohitsharma0690/multi_scale_head_gesture>
5. [**RECENT**] Real-time on-device nod and shake recognition
   1. <https://arxiv.org/pdf/1806.04776.pdf>
   2. Use depth sensor, IPhone
   3. Real-time
   4. NOT per-frame
   5. Data augmentations
      1. shrinking and stretching on a full sequence.
      2. shrinking and stretching heads and tails of a sequence
   6. LSTM / GRU
6. Online nod detection in human-robot interaction
   1. <https://ieeexplore.ieee.org/abstract/document/8172396>
   2. SVMs
   3. Realtime
   4. KOMPASS WOZ1 dataset
   5. Kinect based - KTH-IDIAP
7. Head Nod Detection from a Full 3D Model
   1. SVMs
   2. Kinect, full 3D
   3. <https://publications.idiap.ch/downloads/papers/2015/Chen_ICCV2015WORKSHOP_2015.pdf>
8. Robust Head Gestures Recognition for Assistive Technology
   1. nodding, shaking, turning right, turning left, looking up, and looking down
   2. HMM-based
   3. <https://link.springer.com/content/pdf/10.1007/978-3-319-07491-7_16.pdf>
   4. Data collected:
   5. [ASKED FOR ACCESS]
9. Using Self-Context for Multimodal Detection of Head Nods in Face-to-Face Interactions
   1. Dataset - collected 8 interactions, total 1h40min
   2. 2 classes of head nods: obvious and subtle
   3. Use bounding box and motion estimation, FT
   4. SVM, binary, not use subtle nods
   5. 5000 training examples
   6. Audio + visual
   7. Fvisual1= 0.559,Fmulti1= 0.6283
   8. [ASKED FOR ACCESS]
10. An ICA based head movement classification system using videosignals
    1. Not of much use, but refs to datasets! [datasets not useful]
    2. <https://link.springer.com/content/pdf/10.1007%2Fs12193-017-0244-0.pdf>
    3. Independent component analysis with NN
    4. Left, right, up, down movements
11. A Real Time Decision Support System using Head Nod and Shake
    1. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7530336&tag=1>
    2. Realtime
    3. Vision based
    4. 91.1% nod and shake
    5. SVM
    6. Own dataset
    7. Not per-frame annotations
    8. [ASKED FOR ACCESS]
12. Recognition of Simple Head Gestures Based on Head Pose Estimation Analysis
    1. up/down/left/right
    2. Depth camera
    3. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.676.6128&rep=rep1&type=pdf>
13. Classifying head movements in video-recorded conversations based on movement velocity, acceleration and jerk
    1. <http://www.ep.liu.se/ecp/141/003/ecp17141003.pdf>
    2. SVM
    3. Use: velocity, acceleration, jerk
    4. NOMCO dataset
    5. **show that using jerk improves accuracy**
       1. We expect that a sequence of frames for which jerk has a high value in the horizontalor vertical direction will correspond to the most effortful part of the head movement (often calledstroke(Kendon, 2004), orapex(Loehr, 2007)).
    6. With audio features:
       1. + Pitch and intensity of cooccuring speech
       2. Detecting head movements invideo-recorded dyadic conversations
       3. MLP, 75%
       4. adding additional pitch and intensity features increases the accuracy
14. Evaluating Models of Speaker Head Nods for Virtual Agents
    1. [https://ict.usc.edu/pubs/Evaluating%20Models%20of%20Speaker%20Head%20Nods%20for%20Virtual%20Agents.pdf](https://ict.usc.edu/pubs/Evaluating Models of Speaker Head Nods for Virtual Agents.pdf)
    2. Nods of speaker
    3. In introduction, they say that “head movements are also influenced by our emotions” => using action units to predict head gestures is reasonable and might be beneficial
    4. Use linguistic features
    5. Compare rule-based and data-driven approach
    6. Models predict only timing of the nods, not their dynamics (magnitude, velocity and length of the nods were unified)
    7. Conclusion: data-driven approach had significantly less nods displayed at inappropriate times than the rule-based approach

**Novelties / important features**

* Combination of:
  + Data-driven & Deep learning approach
  + Trained on spontaneous gestures / conversations
  + Offline / Real-time
  + Robust to speaking
    - Trained on both kinds of data (speaker, listener)
  + Multimodal? [head pose + eye gaze? + land marks?]
    - Can experiment with speaker **audio**
      * This would be extra work, not usable for DVRA
  + Per-frame prediction
  + Based on single-view RGB input (no depth info / Kinect)
  + Trained on multiple datasets?
  + Nod, shake and tilt

Comparison of previous works

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Real-time /**  **Offline** |  | **Model** | **Gestures** | **Modalities** |
| 1 | RT |  | HMM | Nod, shake |  |
| 2 |  |  | HMM | Nod, shake |  |
| 3 | RT |  | HMM | Nod, shake | Video, depth |
| 4 | OFF |  | MS-CNN-LSTM | Nod, shake, tilt, ... | Video |
| 5 | RT |  | LSTM, GRU | Nod, shake | Video, depth |
| 6 | RT |  | SVM | Nod | Video, depth |
| 7 |  |  | SVM | Nod | Video, depth |
| 8 |  |  | HMM | Nod, shake |  |
| 9 |  |  | SVM | Nod | Audio, video |
| 11 | RT |  | SVM | Nod, shake | Video |
| 13 |  |  | SVM, MLP |  | Audio, video |
| 14 |  |  | Rule-based, HMM | Nod |  |