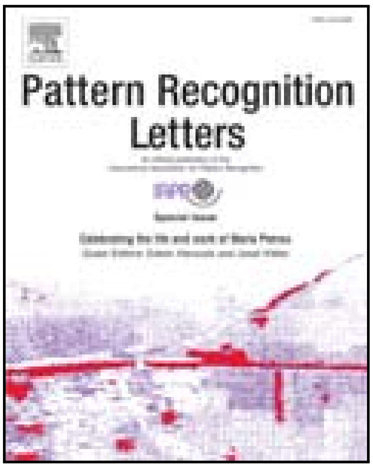


The MAHNOB Mimicry Database - a database of naturalistic human interactions

Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, Maja Pantic

PII: S0167-8655(15)00076-8
DOI: 10.1016/j.patrec.2015.03.005
Reference: PATREC 6185

To appear in: *Pattern Recognition Letters*

Received date: 27 February 2014
Accepted date: 5 March 2015

Highlights

- We present an audiovisual dataset for investigation of mimicry behaviour.
- We report baseline performances from per-session mimicry classification experiments.
- Performance is session-dependent, due to variability in subject expressiveness.
- Current mimicry classification methods need more development for spontaneous data.

Pattern Recognition Letters
journal homepage: www.elsevier.com

The MAHNOB Mimicry Database - a database of naturalistic human interactions

Sanjay Bilakhia[a,**], Stavros Petridis[a], Anton Nijholt[b], Maja Pantic[a,b]

[a] *Department of Computing, Imperial College London, UK*
[b] *EEMCS, University of Twente, The Netherlands*

ARTICLE INFO

Article history:
Received xxxx
Received in final form xxxx
Accepted xxxx
Available online xxxx

Keywords:
Behavioral mimicry
Motor mimicry
Temporal modelling
Social signal processing

ABSTRACT

People mimic verbal and nonverbal expressions and behavior of their counterparts in various social interactions. Research in psychology and social sciences has shown that mimicry has the power to influence social judgment and various social behaviours, including negotiation and debating, courtship, empathy and helping behaviour. Hence, automatic recognition of mimicry behaviour would be a valuable tool in various domains, and especially in negotiation skills enhancement and medical help provision training. In this work, we present the MAHNOB Mimicry database, a set of fully synchronised, multi-sensory, audiovisual recordings of naturalistic dyadic interactions, suitable for investigation of mimicry and negotiation behaviour. The database contains 11 hours of recordings, split over 54 sessions of dyadic interactions between 12 confederates and their 48 counterparts, being engaged either in a socio-political discussion or negotiating a tenancy agreement. To provide a benchmark for efforts in machine understanding of mimicry behaviour, we report a number of baseline experiments based on visual data only. Specifically, we consider face and head movements, and report on binary classification of video sequences into mimicry and non-mimicry categories based on the following widely-used methodologies: two similarity-based methods (cross correlation and time warping), and a state-of-the-art temporal classifier (Long Short Term Memory Recurrent Neural Network). The best reported results are session-dependent, and affected by the sparsity of positive examples in the data. This suggests that there is much room for improvement upon the reported baseline experiments.

1. Introduction

Research in psychology has found that people mimic postures, facial expressions, mannerisms and other verbal and nonverbal expressions of the counterpart in social interaction [6][16]. Contagious effects of laughter and yawning, mimicry of speech rate and rhythms, and imitation of smoking behaviour and mannerisms are just but a few examples [13]. Mimicry has been operationalized in varying ways and has overlap with related phenomena including interactional synchrony [9], and interactive alignment [28]. All of these phenomena (including mimicry) fall under the larger category of behavioural similarity. Mimicry behaviour can be divided into motor mimicry and emotional mimicry [15]. Motor mimicry constrains behaviours to be identical in expression (but not in duration, intensity or phase). In emotional mimicry, the displayed behaviours may not be identical, but have the same “functional value”, i.e. “convey the same message” in terms of the underlying affective state, including but not limited to, sadness, empathy, or trust. Note that motor mimicry may also be (a part of) an emotional mimicry episode. For example, an inner-brow raise

[**] Corresponding author: Tel.: +44-207-594-8195;
e-mail: sb1006@imperial.ac.uk (Sanjay Bilakhia)

displayed in sadness may be mimicked (and perhaps intensified by additional displays of chin raise and downwards head tilt). In this work we largely focus on motor mimicry, mainly because of its agnostic character. To wit, while emotional mimicry judgment is all about interpretation of what underlies the displayed behavioural expression and mimicry episode, motor mimicry judgment is objective, describing just the "surface" of the shown behaviour, such as which facial movement or speech mannerism has been mimicked, leaving inference about the conveyed message (emotion) to higher order decision making.

Research in psychology and social sciences has shown that presence or absence of motor mimicry behaviour can serve as a (positive or negative) indicator of co-operativeness [16], social judgment [13], presence of autism spectrum disorder [20], and even traumatic brain injury [18]. Hence the presence and characteristics of motor mimicry behaviour can serve as a useful step in higher-order behavioural inference. It is not surprising then that automated machine recognition of interpersonal mimicry behaviours could be of tremendous help to research and society. It could speed up research in behavioural, political, and social sciences. It could be of tremendous value for feedback provision in negotiation skills and medical help provision training. Crucially, it could revolutionise the way we interact with robots and avatars; such technology would enable these artefacts to mimic their human counterparts properly and show rapport and collaboration and evoke trust. Recently, few pioneering efforts towards machine analysis of mimicry behaviour have been reported, but the research on the topic is still in its infancy, partly because of the lack of suitable data to train machine learning algorithms on.

In this work we provide a comprehensive description of the MAHNOB Mimicry database, a collection of fully-synchronised multi-sensory audiovisual recordings of naturalistic dyadic interactions suitable for investigation of mimicry and negotiation behaviour. Although primarily intended for investigation of behavioural mimicry, the data can be used in studies of other social phenomena such as turn taking, rapport, and back channel communication. It is also suitable for evaluation of signal processing and machine learning techniques, for head pose estimation, facial expression tracking, automatic speech recognition, and similar. The database contains 11 hours of recordings, split over 54 sessions of dyadic interactions between 12 confederates and their 48 counterparts, being engaged either in a socio-political discussion or negotiating a tenancy agreement. Out of 54 sessions 15 have been fully annotated, in terms of facial points tracked for both session participants, millimetre-precision six-degrees-of-freedom (6-DOF) head pose for both participants, and human judgments of motor mimicry behaviours of head gestures, hand gestures, facial expressions, shoulder movements and postural shifts of the torso. The database is publicly available for non-commercial use at http:// mahnob-db.eu/ mimicry.

The MAHNOB Mimicry Database database is the first of its kind as it satisfies all of the following:

- contains fully-synchronised recordings of interpersonal dyadic interactions,
- contains recordings of a fairly large number of subjects and moderately wide range of ethnicities, and near-equal subject gender balance,
- is filmed in conditions allowing comprehensive research in computer vision and signal processing, in terms of range of views, amenable lighting conditions, good image resolution, and highly accurate synchronization of all recording sensors, and
- is annotated in terms of a large number of head, hand, shoulder and face gestures, and motor mimicry episodes involving these gestures.

The MAHNOB Mimicry Database has been partially presented at conferences (see [36][35][2]) but a complete description of the data, the recording protocol and the available annotations, has not been reported so far. The novelty of this work is not only in provision of a complete description of this database, it is also in provision of baseline experiments that could serve as benchmark for efforts in the field. We consider face and head movements tracked by the state-of-the-art-trackers (i.e. the face tracker described in [25] and the head pose estimator described in [17]) and report on binary classification of video sequences into mimicry and non-mimicry categories based on the following widely-used methodology: two similarity-based methods (cross correlation as used in [22] and Generalised Time Warping [40]), and the state-of-the-art temporal classifier, Long Short Term Memory Recurrent Neural Network (LSTM-RNN) [32]. Performance of the methods is evaluated against the ground truth, representing human annotations of motor mimicry behaviour.

The motor mimicry behaviours considered in our experiments include smiles, frowns, and eyebrow raises, as well as head nods, head shakes, and significant shifts in head posture. We say that a motor mimicry episode has occurred if one of the subjects displays a behaviour previously displayed by her counterpart and does so within a certain time limit. This time limit has both a lower and an upper bound. The former is set so as to distinguish between synchronicity and mimicry. The latter is set so as to distinguish between motor mimicry and behaviours that are identical in expression but displayed with large delay and, hence, having low likelihood that they represent motor mimicry. These thresholds are set to 0.04 seconds and 4 seconds. Research in psychology has shown that people need at least 40ms to recognise and start mimicking a facial movement [34], and hence we set the lower bound to 0.04 seconds. The upper boundary has been set in an experimental fashion, by reviewing all motor mimicry episodes annotated as such by the human annotators in the MAHNOB Mimicry database and setting the threshold to the duration of the longest delay.

The problem of automatic motor mimicry recognition is made difficult by the fact that the mimicked behaviour is identical in expression but may not be identical in duration, intensity and phase. For example, for nods, the primary rotation around transverse axis can be mixed with other rotations, and can vary in velocity, intensity, phase and number of periods. This variability implies that the events need to be related to each other through some non-trivial spatiotemporal transform. Most current methods for mimicry detection or classification rely on a pair-wise similarity measure, combined with a method to account for the delay in the reaction (via time-lags) and for the variability in the duration of the reaction (via temporal-window-based analysis). We use similar approaches in this work.

We conduct two sets of experiments. The first one is based on facial cues only, where positive examples consist of sequences containing motor mimicry of facial movements only. The second experimental setup is based on facial and head motions, and positive examples are sequences containing motor mimicry of facial and head movements. Of the tested methods, LSTM-RNNs gave the best performance due to the methods inherent ability to model well arbitrary spatio-temporal transformations. However LSTM-RNNs suffer from significant variance in classification performance, which we also observed in our experiments. The best reported results are session-dependent and affected by the sparsity of positive examples in the data. This suggests that there is much room for improvement upon the reported baseline experiments.

The paper is further organised as follows. Sec. 2 reviews prior work. Sec. 3 provides a complete description of the MAHNOB Mimicry Database. Sec. 4 and Sec. 5 detail the conducted baseline experiments. Sec. 6 concludes the paper.

2. Prior work

2.1. Other databases

Various databases containing audiovisual recordings of naturalistic human behaviour have been reported to date. These include databases of elicited naturalistic emotional responses to various video material (e.g., AM-FED [19], see [39], [27], for overviews), databases of broadcast material used in studies on (machine) analysis of social roles and personality (e.g., see [37] for an overview), databases of human-avatar interactions (e.g., SEMAINE database [21], the work provides also an overview of other such databases), and databases of interpersonal interactions where the involved subjects are co-located and recorded simultaneously (e.g., [12] provides an overview of group meetings data repositories and [9] provides an overview of data used in studies on (machine) analysis of interpersonal synchrony). However, most of these data repositories either publicly unavailable (e.g. [30], Spontal [10]) or suffer from some of the following limitations.

- The recordings are of professional actors and it is unclear whether the recorded interactions are acted or spontaneous (e.g., as in IEMOCAP [5]).
- The recordings are of short interactions. Research in psychology has shown that mimicry behaviour becomes more frequent as interaction progresses [4]; hence, short interactions (less than 5 minutes, e.g., as in NOMCO [26]) have significantly decreased likelihood of containing mimicry episodes.
- The recordings are technically suboptimal. The most common problem is the low accuracy of sensory synchronisation. To facilitate audio and/or visual analysis of mimicking behaviour, analysis of temporal-interdependencies of behavioural patterns shown by the interacting persons must be facilitated, and that is possible only if the utilised sensors (microphones and cameras) are synchronised to a high accuracy (of less than 40ms error, given that people need 40ms to recognise and start mimicking a behaviour [34]). However most of the currently available data on interpersonal interaction have an error in the synchronisation of the utilised cameras that is > 40 ms (which is the duration of one frame in 25fps temporal resolution of the camera, e.g., as in [6]). Other technical problems include low-resolution videos (e.g., as in NOMCO [26] and IFADV [33]), suboptimal view of the subjects (e.g., as in D64 [23]), and similar.
- The recordings have not been annotated by human experts in terms of motor mimicry episodes, facial movements, head and hand gestures, body postures, etc. (e.g., as is the case with IFADV [33]). Building effective and efficient machine learning algorithms for mimicry recognition depends on having suitable ground truth to learn from. Hence, an important aspect of making progress in the field lies in providing suitable datasets of enough labelled examples for building robust tools.

As explained in section 3, the MAHNOB Mimicry database has been collected as to address these limitations of the currently available databases of spontaneous dyadic interactions.

2.2. Machine recognition of motor mimicry

Previous works on mimicry detection have been based on a measure of correlation between the subjects' data. The methods usually rely on the construction of a control dataset, in which each subjects data is independently and randomly permuted. This permutation is applied to either windows of samples or individual samples. Generally a synchrony score is calculated on both the control and the original dataset, and a hypothesis test is used to either highlight informative variables, or temporal windows with significant similarity. Ramseyer et al. [29] used a proprietary dataset containing 104 sessions of cognitive behavioural therapy, in order to investigate the relationship between synchrony and clinical outcomes. They calculated windowed cross-correlation of motion energy, with time lags ranging from -4 to +4 seconds (i.e. for cross-correlation at time lag L, motion energy extracted from the first subject S_1 in the time interval $[t_1, t_2]$ is compared against that extracted from the second subject S_2 in the time interval $[t_1 - L, t_2 - L]$). These per-window scores were then aggregated into a global score. The global scores between original and (window-level) permuted data were compared for significance to distinguish those with synchrony behaviour. Boker et al. [3] investigated whether movement synchrony increased in the presence of acoustic noise. They used a proprietary dataset of 8 subjects in a conversational setting. Data was captured from body-mounted inertial sensors. They used windowed cross-correlation and peak-picking, with time lags ranging from -2 to +2 seconds, to estimate synchrony in head movement. Neither of the prior two works report any performance figures. Sun et al. [36] used the MAHNOB Mimicry Database to show that subjects tend to mimic body postures, head movements and hand gestures of their counterparts, and to investigate how mimicry evolves during an interaction. They used windowed cross-correlation of motion intensity histograms to

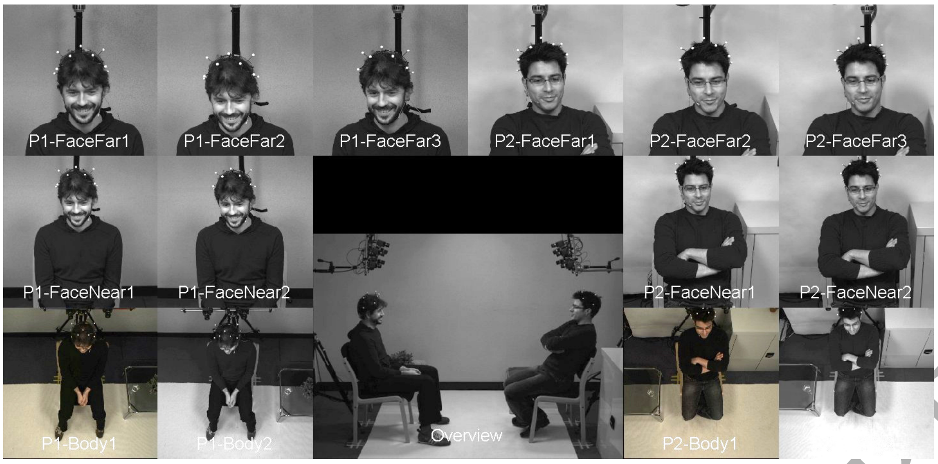


Fig. 1: Camera views available from the MAHNOB Mimicry database

provide a similarity measure for mimicry behaviour. They observed that the mean(s.d.) windowed cross-correlation across all negotiation scenarios rose from 0.53(0.01) at the start of a session, to 0.6(0.02) at the end of a session. For discussion scenarios, the average windowed cross-correlation rose from 0.3(0.02) to 0.5(0.03). These trends were also apparent in individual sessions. Altmann et al. [1] used a proprietary dataset of schoolchildren, to investigate how synchrony varies between friend and non-friend dyads, when placed in competitive and neutral scenarios. They used motion energy features to compute 2 linear regression models per window. The first model contained auto-regressive and cross-regressive components, whilst the second model contained an auto-regressive component only. If the difference in R^2 between the 2 models was significant (by F-test), it was inferred that the variance explained by the cross-regressive terms was significant, and hence an indicator of synchrony between the subjects. They quantify synchrony occurrence as the proportion of windows with statistically significant R^2 differences. Across all sessions, they report mean(s.d.) synchrony occurrences of 0.187(0.11) and 0.120(0.09) for neutral and conflict states respectively. Feese et al. [11] used a large, proprietary dataset to quantify behavioural mimicry. Subjects worked in small groups to rank fictional job candidates. Data was captured from body-mounted inertial sensors. They used gesture detectors for behavioural events, including head, arm and torso movements. They define positive output from the detectors closer than a temporal threshold to be mimicry. They report precision, recall and F1 scores for detection of their behaviour primitives, but they do not compare against human-rated annotations for mimicry behaviour specifically. They report F1 scores of e.g. 0.57 for posture changes, 0.67 for head nods, and 1.00 for face-touching. Delaherche et al. [8] used a proprietary dataset of students completing a cooperative task to quantify behavioural sychrony. They used motion energy, optical flow, and prosodic features, and calculated cross-correlation and magnitude coherence between all pairs of features, in one-second windows. Synchrony was assumed for a pair of features if the difference in cross-correlation between the real data and a sample-permuted control dataset was statistically significant. They quantify synchrony occurrence as the percentage of windows with statistically significant cross-correlation (compared to their control set). This percentage is high for some feature-pairs (e.g. 83.3% for both subjects' motion history image, 73.2% for both subjects' motion energy), and low for others (e.g. 28.9% for cross-correlation between one subject's acoustic pitch and the other subject's motion energy). None of the works mentioned above attempt to compare their methods to ground truth of mimicry behaviour. Michelet et al. [22] used a proprietary dataset for sequence classification into mimicry and non-mimicry classes. Their dataset contains 256 clips of posed, gross body movements, set against a uniform, static background. They used spatio-temporal interest points to extract HOG/HOF features, which were quantized into a dictionary. Windows within a sequence were then described by a bag-of-words. Cross-correlation and dynamic time warping (across these windows) were used as a similarity measure between entire sequences. For each sequence, a threshold was used to discriminate between mimicry and non-mimicry classes. ROC (receiver operating characteristic) curves were reported, giving a best performance of 0.920. Delaherche et al. [7] proposed a similiar approach to that in [22], using the likelihood ratios between one-class SVMs as a distance measure between class distributions. They used the same posed data from [22]. They reported a maximum area-under-curve of 0.92, similarly to the results reported in [22]. Bilakhia et al. [2] did preliminary investigation of long short-term memory networks. This is currently the only work on mimicry behaviour detection in continuous data (i.e. detection of multiple events in one entire sequence), as opposed to classification of pre-segmented sequences. They used facial animation parameters (FAPs) and cepstral features, extracted from the MAHNOB Mimicry database. For both the mimicry class and the non-mimicry class, an ensemble of regressors was learnt. These regressors learned to map the features from one subject to the features from their counterpart, and vice-versa. Given an unseen input sequence, the class-specific ensemble of regressors with the lowest reconstruction error was taken to be the one corresponding to the input sequence. Experiments were session-specific. The results show good recall, between 60%-90%, but poor precision (between 6%-80%), for each class. This is due to the class imbalance in the dataset.

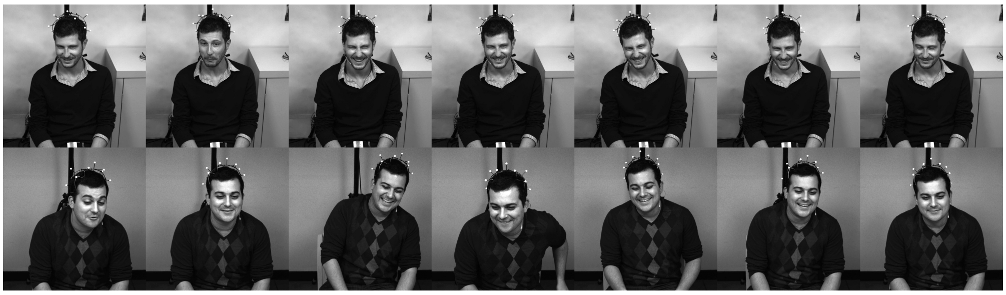


Fig. 2: Mimicry episode of laughter, S32, 9m06s-9m16s

3. Database

In this section we provide a comprehensive description of the MAHNOB Mimicry Database, a collection of annotated, accurately synchronized (to an error of less than 4ms), multisensorial, audiovisual recordings of naturalistic, dyadic interactions.

Protocol: The dataset consists of 54 recordings of face-to-face interactions. Of these, 34 are discussions on a contemporary sociopolitical topic, and 20 remaining are tenancy agreement negotiations. In both cases, while the participants are given the topic of discussion, no script was provided - participants were free to discuss at their leisure. In the latter case, session outcomes, i.e. whether participants decided to live together, are provided together with the rest of the data (of these, 18 are positive). Session lengths range between 5 and 20 minutes, with an average length of 12 minutes. In total, 11 hours and 40 minutes of recordings are available.

Recording setup: The corpus was recorded using the following sensors (see Fig. 1):

- Audio sensors: 1 far-field microphone, 1 head-mounted microphone per subject
- Visual sensors: 2 frontal cameras per subject, covering the head and torso (*FaceNear{1, 2}); 3 frontal cameras per subject, covering the head only (*FaceFar{1, 2, 3}); 2 downward facing cameras per subject, covering the entire body (*Body{1, 2}); and one profile-oriented camera covering both subjects (Overview).

Exact details of the utilized sensors, capture computers, and sensor fusion and synchronization procedures can be found in [35]. The frontal camera descriptors "Far" and "Near" refer to "far-field" and "near-field" measurement. The far-field camera has a longer focal-length than the near-field camera, hence the subject comprises more of the frame. Highly accurate synchronization (of less than 4ms) of 15 cameras, with 1024x1024 spatial- and 58Hz temporal resolution, and 3 microphones with 48kHz temporal resolution, is achieved by recording all trigger signals with a multi-channel audio interface. Each subject also wore a "tiara" with 9 white markers to provide accurate head pose estimates, as in Fig. 1 (see [17] for more details). The head pose data are available in each camera's respective reference frame, or the global reference frame.

Participants: 12 confederates and 48 counterparts took part in the study. Non-confederate subjects were told that the purpose of the recordings was automatic measurement of behaviour in debate and negotiation scenarios. Subjects were recruited from staff and students at Imperial College London, and span a range of ethnicities and primary languages (primarily Europe or the Near-East). Subject nationalities include Spanish, French, Greek, English, Dutch, Portugese, and Romanian. The subjects' ages range between 18 and 34 (μ=25, σ=4.8). Subjects' age and nationality are available for all subjects. Subjects were previously unacquainted. The database is recorded in English. There are 29 female and 31 male subjects, of which 15 wore eye-glasses.

Annotations: The data has been fully annotated for 15 out of 54 sessions, for motor mimicry behaviour of head gestures, hand gestures, body movement and facial expressions. The annotations in terms of gestures (i.e. which gesture occurs when in the data) have been attained in a semi-automatic manner - an automated detector of the target gesture has been run and the results have been manually inspected and corrected. For head gestures - nods, shakes, and tilts - the method in [14] has been used. Postural shifts of the head (in pitch, yaw and roll) have been tracked by the highly accurate methodology of [17]. For hand gestures - hand raising of left and right hand - the method in [24] has been used. For postural shifts in the torso and shoulder shrugs, optical flow methodology similar to that used in [36] has been used. Finally, for facial gestures - smiles, frowns, and raised eyebrows - we used the facial point tracker and facial action parameter coder in [25]. As already explained in the introduction, we focus on motor mimicry episodes in this work. The adopted definition of motor mimicry is similar to that used in [11], and is illustrated in Fig. 3. The episode onset is taken to be the onset of the mimickee's gesture, whilst the offset is taken to be the offset of the mimicker's identical gesture. Assume a behaviour instance $b_i^{S_m}$. We define the start and end times of each behavior instance to be $t_s(b_i^{S_m})$ and $t_e(b_i^{S_n})$, respectively. A behavior of subject S_m, $b_i^{S_m}$, is mimicked by subject S_n if a behavior instance $b_j^{S_n}$ exists that satisfies:

$$b_i^{S_m} = b_j^{S_n}, \quad t_s(b_j^{S_n}) > t_s(b_i^{S_m})$$

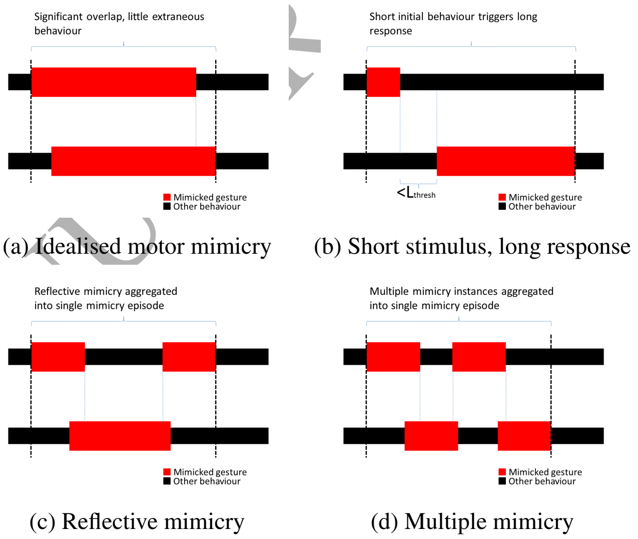
$$t_s(b_j^{S_n}) < t_e(b_i^{S_m}) + 4\text{sec}^{1}$$

The start time of a mimicking event is given by $t_e(b_i^{S_m})$, while the end time is given by $t_e(b_j^{S_n})$. Mimicry episodes have significant variability in their temporal structure, as seen in Fig. 3. These generally fall into four cases. Fig. 3a illustrates

[1] As explained in the introduction, the value of this threshold has been determined experimentally for the data in the MAHNOB Mimicry database.

Table 1: Per-session mimicry incidence statistics

Session #		1	2	3	4	5	6	11	21	30	32	33	35	42	44	53	Total
Mimicry type	Smile	10	2	5	10	16	10	11	7	11	23	6	14	28	18	5	176
	Laughter	6	0	0	5	4	0	1	7	8	7	0	0	13	3	1	55
	Frown	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	3
	Eyebrow raise	0	0	0	0	0	0	0	0	0	1	0	2	4	0	0	7
	Head nod	32	8	21	90	68	56	40	41	12	48	27	13	38	22	4	520
	Head shake	5	0	1	2	4	1	1	0	0	1	2	1	1	5	0	24
	Head pose shift	6	2	2	3	1	4	6	0	0	8	1	5	0	0	0	38
	Shoulder shrug	3	0	0	0	0	0	2	0	0	0	0	0	0	0	0	5
	Left hand movement	3	0	1	1	2	0	0	2	2	1	0	1	0	6	0	19
	Right hand movement	3	0	1	2	4	1	0	4	3	4	0	3	3	8	0	36
	Torso shift	1	0	3	0	0	0	3	1	2	2	0	3	0	1	0	16
	Total	69	13	34	114	99	72	64	62	38	95	36	42	88	63	10	900


Significant overlap, little extraneous behaviour
Mimicked gesture
Other behaviour
Short initial behaviour triggers long response
<Lthresh
Mimicked gesture
Other behaviour
Reflective mimicry aggregated into single mimicry episode
Mimicked gesture
Other behaviour
Multiple mimicry instances aggregated into single mimicry episode
Mimicked gesture
Other behaviour

(a) Idealised motor mimicry

(b) Short stimulus, long response

(c) Reflective mimicry

(d) Multiple mimicry

Fig. 3: Mimicry episode construction (black dashed lines define episode boundaries)

the case where only short time delay exists between the onsets and offsets of corresponding behaviours (e.g. in S16 at 13m07s). Fig. 3b illustrates the case where a short-duration initial behaviour (such as a monosyllabic laughter episode) triggers a much longer response by the counterpart (e.g. in S26 at 1m08s). The inverse can also occur, where a long-duration initial behaviour triggers a much shorter response by the counterpart. Several mimicry occurrences can also be aggregated into the same episode, as illustrated in Fig. 3c and Fig. 3d. Fig. 3c shows "reflective" mimicry, where subject 2 mimics an action of subject 1, which is subsequently mimicked by subject 1, such as in contagious laughter (e.g. in S28 at 2m7s). Episodes of reflective mimicry contain 3 or more displays of the same behaviour. Fig. 3d shows multiple mimicry, where the onset of a behaviour subsequently mimicked occurs before the offset of a previously mimicked behaviour (e.g. in S19 at 1m54s). These mimicry episodes are concatenated together due to temporal overlap. However, unlike reflective mimicry, they are not required to contain mimicry of the same behaviour display. For example, a multiple mimicry episode may contain an instance of smile mimicry concatenated to an instance of nod mimicry.

All motor mimicry episodes have been annotated by two annotators using the ELAN annotation software [38]. If discrepancies in annotation occurred, (e.g. in the exact timing of onset/offset of an episode), these were discussed to reach an agreement. Table 1 shows annotated session statistics for these motor mimicry episodes, per mimicked gesture. Mimicry episodes of head nods, smiles, and laughter are the most numerous - this is not surprising, as the confederates and their counterparts were previously unacquainted. Research in psychology has shown that people try to be liked by new acquaintances and tend to mimic positive emotions, characterised by laughter, smiles, and nods, more often [15]. An example of ideal motor mimicry of laughter is illustrated in Fig. 2.

4. Baseline experiments: Setup

Here we describe the setup for the conducted baseline experiments. We consider face and head movements, tracked by the state-of-the-art-trackers, in 10 annotated sequences (S3, S4, S5, S6, S11, S21, S32, S33, S42, S44) described in Table 1, and report on binary classification of these video sequences into mimicry and non-mimicry categories based on the following widely-used methodologies: two similarity-based methods and the state-of-the-art temporal classifier (LSTM-RNN). We conduct two sets of experiments. The first one is based on facial cues only, where positive examples consist of sequences containing motor mimicry of facial movements only. Hence (for the first experimental scenario only) an example sequence containing a mimicry episode of head nods is considered as a negative example. The second experimental setup is based on facial and head motions, and positive examples are sequences containing motor mimicry of facial and head movements. In both scenarios, all data from each session is used. Performance of the methods is evaluated against the ground truth, representing human annotations of motor mimicry behaviour.

Data: To account for the extreme inter-session and inter-subject variability, all experiments were performed on a per-session basis. We used 10 sessions in all experiments. As can be seen from Table 1, these sessions have enough positive examples of motor mimicry of both facial and head movements. The mimicry episodes in these sessions are also of reasonable intensity, in contrast to others where subject expressions are very subtle (such as session 1). Hence our choice to use these sessions. For each session, the data was split into 3 disjoint subsets for training, model validation, and testing. The training data consisted of the first part of the session such that it contained half of all mimicry episodes encountered during that session. The second part of the session containing the next quarter of all the mimicry episodes formed the validation set. The remaining data was used for testing. Splitting of the data per session into the training, validation and testing sets could have been done differently, e.g. based on the median and third-quartile of duration. However, such approaches would be suboptimal as positive examples are sparse and unevenly distributed over time. Hence, such an approach could cause one of the sets, e.g. the validation set, to have no positive examples. We also encountered another problem. As our data is fully spontaneous data, non-mimicry episodes in such data are significantly more frequent and longer than mimicry episodes. Hence, classification of such episodes into mimicry and non-mimicry classes is trivial using a simple threshold on episode length. To avoid this pitfall, we artificially segment long negative examples (i.e. non-mimicry episodes), with segment boundaries drawn from a Gamma distribution, fit to the empirical distribution of positive examples' (i.e. mimicry episodes') lengths. In that way we obtain positive and negative examples of comparable temporal lengths.

Features: For head movements, we use the millimetre-precision 6-DOF head pose estimator in [17] and calculate three rotation velocities as the features to be used in further processing. Velocities were obtained using finite-differences of the smoothed head pose and were then scaled to the interval [0,1]. Smoothing has been carried out using a cubic polynomial fit to a sliding window of 15 frames. For facial movements, we use facial point tracker in [25] and adopt Facial Animation Parameters (FAPs), directly calculated by the utilised tracker, as the features to be used in further processing. We chose not to use raw facial points and instead use the FAPs (corresponding to the upper lip position, jaw drop, lip width, inner brow height, lip corner position, and outer brow height), as they are representative of facial motion and descriptive of facial expressions. This results in 9 features (6 FAPs and 3 head rotation velocities) for each video frame of a target video sequence.

Cross-correlation: Analogous to the work in [22], we use a $D_{X_{s_1}}$ x $D_{X_{s_2}}$ cross-correlation matrix P, where $D_{X_{s_1}}$ and $D_{X_{s_2}}$ are the feature dimensionalities of data matrices X_{s_1} and X_{s_2}. These data matrices correspond to the features of subjects S_1 and S_2 respectively, for a given input sequence. A scalar measure of similarity is then calculated from this cross-correlation matrix. To this end, we take the trace of the cross-correlation matrix, as we are interested in (homogeneous) motor mimicry - i.e. mimicry of the same gestures, (i.e. the same features). We account for reaction delay by calculating this measure across different time-lags. We define time-lag relative to subject S_1. A cross correlation matrix with time-lag t, $P^t(X, Y)$, is defined element-wise as:

$$\rho^t_{ij} = \frac{\sum_{k=t}^{n}(x_{k-t,i} - \bar{x}_i)(y_{k,j} - \bar{y_j})}{\sqrt{\sum_{k=t}^{n}(x_{k,i} - \bar{x}_i)^2 \sum_{k=1}^{n-t}(y_{k,j} - \bar{x}_j)^2}} \tag{1}$$

$$1 \leq i \leq D_{X_{s_1}}, 1 \leq j \leq D_{X_{s_2}}, n = len(X_{s_1}) = len(X_{s_2})$$

The final similarity measure between the two sequences is then taken to be the mean of these traces:

$$xcorr\_dist(X_{s_1}, X_{s_2}) = \sum_{i=1}^{T} tr(P^{t_i}(X_{s_1}, X_{s_2})) \tag{2}$$

$$\forall i = 1..T$$

where T is the size of the set of time-lags, and tr() is the matrix trace function.

A decision threshold θ^* is then used to discriminate between mimicry and non-mimicry. This is empirically determined by optimizing over the range of similarity values calculated on a validation set S:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \qquad 2^*\tau^*\nu^*(\tau + \nu)^{-1} \tag{3}$$

where:

$$\tau = \frac{\sum_{i=1}^{N} y_i\hat{y}_i}{\sum_{i=1}^{N} \hat{y}_i} \qquad \nu = \frac{\sum_{i=1}^{N} y_i\hat{y}_i}{\sum_{i=1}^{N} y_i}$$

$$\theta^* \in [min(xcorr\_dist(X_{s_1,i}, X_{s_2,i}), i = 1...N),$$
$$max(xcorr\_dist(X_{s_1,i}, X_{s_2,i}), , i = 1...N)]$$

$$\hat{y}_i = I[xcorr\_dist(X_{s_1,i}, X_{s_2,i}) > \theta^*]$$

$$S = \{(X_{s_1,1}, X_{s_2,1}, y_1), (X_{s_1,2}, X_{s_2,2}, y_2)...(X_{s_1,N}, X_{s_2,N}, y_N)\}$$

where y_i is the true class label for example sequence i, $\hat{y}_i$ is the predicted class label for example sequence i, N is the number of example sequences, θ^* is the optimal decision threshold, τ is precision, ν is recall, and $I[]$ is the indicator function.

Generalized time-warping: Generalized time-warping (GTW) [40] aligns sequences of multivariate data. It does so by jointly finding a low-dimensional projection for each sequence that maximises the projections' cross-correlation, and a warping that maximises the alignment of these discovered projections. The GTW objective cost at convergence can be used as a similiarity measure between sequences, $gtw\_dist(X_{s_1}, X_{s_2})$. Note however, that in this case a higher score will indicate **dissimilarity** between sequences. As with the cross-correlation similarity measure above, we use a decision threshold θ^* given by Eq. 3 to discriminate between mimicry and non-mimicry. However, when using $gtw\_dist(X_{s_1}, X_{s_2})$, the comparison operator in the indicator function is reversed compared to Eq. 3, as a higher value for the $gtw\_dist(X_{s_1}, X_{s_2})$ score indicates that sequences are dissimiliar.

Long short-term memory: The long short-term memory recurrent neural network (LSTM-RNN) [32] is a recurrent neural network model that can preserve long-range dependencies in sequential data. They outperform standard recurrent neural networks, which suffer from gradient diffusion as the error is backpropagated during training. LSTM-RNNs preserve the error signal through a different choice of activation function and recurrent connection weight. We perform binary classification by concatenating the features of each subject together for the whole episode, and train with gradient-descent and resilient backpropagation using the ground-truth labels as output. We use the PyBrain [31] implementation of LSTM-RNN.

Classifier training: As cross-correlation and GTW are unsupervised, we combine the training set and validation sets to estimate the hyperparameters. For the cross-correlation similarity measure, the trace of cross-correlation matrix for each sequence is calculated, for both positive and negative time-lags (in order to account for the fact that any subject could mimick their counterpart). We used time lags of {-24, 0, 24} samples, i.e. we define $\{t_i : i = 1...T\} \triangleq$ {-24, 0, 24} in *xcorr_dist*. Given that the utilised cameras record at 58Hz, a time lag of 24 frames amounts to approximately 0.5s. This value was determined in an experimental fashion by inspecting the annotated motor mimicry sequences. The majority of motor mimicry occurred within 0.5s of the display of the mimicked behaviour in our data. Hence we opted to use 0.5s as the longest time lag considered. Preliminary experiments showed that longer time lags had no effect on performance. The optimal hyperparameters are taken to be those which produce the best F1 performance on the validation set. Due to severe class-imbalance we use a skewed cost-matrix for model validation, to prevent selecting a classifier that just returns the negative class. The skew was set proportional to the class imbalance. For LSTM-RNNs, we use a single hidden layer. The only hyper-parameter is the number of LSTM-RNN blocks in this hidden layer (i.e. the rank of the hidden representation of the data), chosen experimentally by inspecting the performance for [15, 25, 35, ..., 75] blocks, and selecting the best performing one on the validation set. It is well known that the number of model parameters greatly increases with model depth. In turn, the risk of overfitting on small datasets is significantly augmented. Hence we chose to use a single hidden layer of moderate size, to prevent this risk. LSTM-RNNs were trained using resilient backpropagation, and training was stopped when the error gradient fell below some small threshold ϵ, or the number of training epochs (i.e. model parameter updates during learning) reached 500. LSTM networks are trained and tested 10 times with the hyper-parameters found during the validation procedure, to account for the stochastic learning procedure.

Evaluation measures: Classification performance on the test set is measured using sequence-level negative predicted value (NPV), specificity, precision and recall:

$$NPV = \frac{TN}{TN + FN} \qquad specificity = \frac{TN}{FP + TN}$$
$$precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}$$

where TP, FP, TN, FN correspond to true positive, false positive, true negative, and false negative respectively.

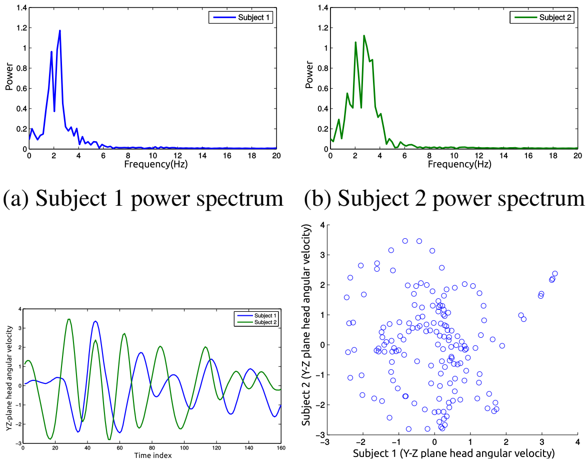
5. Baseline experiments: Results

As explained above, we decided on session-dependent experiments because of high inter-subject and inter-session variability. This would have adversely affected the simple cross-correlation-based and GTW-based classifiers due to their low learning capacity. For example, one could imagine a session with very expressive subjects, where excess extraneous motion drives down similarity globally. This would cause the learned classifier to perform even worse for some other session where subjects are less expressive and display only simple gestures (pushing up similarity globally).

Results for cross-correlation based classification in scenario 1 are shown in Fig. 5a. In most cases, precision is poor, and recall is highly variable, from 33% to 87% (precision/recall of 0 indicates no true-positive predictions). This is mainly due to the sparsity in positive examples and low class-separability that cross-correlation-based classification can achieve. Positive example sparsity depresses precision, as misclassifying a small proportion of negative examples inflates the number of false positives relative to true positives. Non-separability arises from cross-correlation's inadequacy for temporal data. For example, both subjects sitting motionless gives a high cross-correlation value, whereas complex movements (even when being a part of a mimicry episode) give a low correlation value. Examples can be seen in Fig. 6a, where negative examples give high similarity e.g. in S5, $7500 < t < 8000$, due to neutral facial expressions in both subjects. Sometimes cross-correlation performs reasonably well e.g. in S32, $4500 < t < 5000$ (Fig. 6b), corresponding to the intense laughter episode shown in Fig. 2. It also captures well motor mimicry in S32, $2700 < t < 3800$, where a series of intense smile-mimicry episodes occur with small time delay between behaviours. Both above-mentioned cases correspond to idealised motor mimicry (as in Fig. 2(a)), which cross-correlation can detect well. In the second experimental scenario (see Fig. 5b), cross-correlation has lower performance. It tends to give lower specificity and precision, and higher recall. This is due to the introduction of head nods and shakes into the task, greatly increasing the number of positive examples. However cross-correlation is unsuitable for periodic motion, as illustrated in Fig. 4, where an example of idealised mimicry is shown (low time delay, high overlap between actions). The bottom-left panel shows the first derivative of head-pitch for both subjects from a nod mimicry episode in S44. Periodicity is evident in the time domain, whilst their power spectra in Figs. 4a and 4b are similar, with a defined peak power between 3-5Hz. This corresponds to the frequency range for head nods and shakes. However cross-correlation gives very low similarity, as seen in Fig. 4d.

In the first experimental scenario, GTW has worse performance than cross-correlation in terms of specificity and precision (Fig. 5c), as its objective cost is also unrepresentative of gesture similarity. The sequence S5, $2000 < t < 3200$, (Fig. 6a) contains intense smile mimicry episodes, but their time-warping cost is high compared to other non-mimicry sequences in the test set (note that lower cost indicates higher similarity). As with cross-correlation, this is due to intra-gesture variability and weak coupling - one subject smiles with well-defined onset/offset (and constant apex), whilst the other smiles intensely but significant lip movement from enunciated speech is also present. Similarly to cross-correlation in the second experimental scenario, GTW proves inadequate for alignment of periodic data, as the performances in Fig. 5d show. Its requirement that the temporal warping be monotonic precludes alignment of signals with different numbers of periods. This reduces similarity for complex, semantically equivalent gestures, raising the optimal decision-threshold on the validation set. This leads to more false positive classifications, giving high recall and NPV.

In experimental scenario 1 (Fig. 5e) LSTM-RNNs have high



(a) Subject 1 power spectrum (b) Subject 2 power spectrum

(c) Head pitch, nod mimicry sequence (d) No significant cross-correlation (-0.0213)

Fig. 4: Cross-correlation based similarity measures are unsuitable for dynamic phenomena. Though the power spectra are very similar, especially when compared to the power spectrum of miscellaneous head motion as in (d), cross-correlation in the time domain is sensitive to e.g. phase, changing intensity over the episode duration, and non-stationarity.

NPV, and specificity performance between 50-70%, with low precision. As the LSTM-RNN is not constrained to a linear separator, it does not need to learn a decision boundary that misclassifies large volumes of the negative-class to get a few true positives. However, whilst they make positive predictions in temporal proximity to mimicry episodes, they suffer from large numbers of false positives. Their performance is highly variable when trained repeatedly. This is because the training data contains mimicry episodes with no significant overlap between gestures, due to reaction-delay or duration difference. Hence samples corresponding to a behaviour in one subject are aligned with samples corresponding to a different behaviour (or a neutral state) in the other subject (similarly to Fig. 3a). The classifier learns to associate these non-matching behaviours with mimicry, causing false positives at the onset of a non-reciprocated gesture. In scenario 2, precision and specificity improve significantly across all sessions, though recall decreases.

The LSTM-RNNs generally have better precision, with more stable recall across the sessions. Session-specific performance is consistent across all models, e.g most methods perform better on S42 and S44, whilst all perform poorly on S6. Table 2 shows average performance across all sessions for scenario 1, revealing the classifiers' positive bias. For scenario 2, despite the reduced class imbalance, performance is degraded for correlation and GTW. This is because they are unsuitable for characterising oscillatory motion. LSTM-RNN performance does not degrade as significantly. Table 3 shows the average performance across all sessions for scenario 2. We see that NPV and specificity drop for nearly all classification methods compared scenario 1, whilst precision and recall increase. The performances shown in these experiments indicate that more advanced learning methods are needed to accommodate for the variability in human motor-mimicry behaviour.

Table 2: Per-method performance (mean/std) averaged over all sessions, scenario 1 (facial mimicry only experiments)

Model	negative predicted value	specificity	precision	recall
Cross-corr	82.2(8.3)	81.4(6.4)	29.0(18.4)	45.9(21.4)
GTW	84.7(15.4)	51.9(28.5)	11.8(8.0)	47.5(29.3)
LSTM	87.7(8.9)	59.9(6.0)	17.7(12.8)	47.7(9.7)

Table 3: Per-method performance (mean/std) averaged over all sessions, scenario 2 (facial+head mimicry experiments)

Model	negative predicted value	specificity	precision	recall
Cross-corr	56.3(30.8)	24.9(27.3)	26.1(8.9)	76.1(24.0)
GTW	68.1(28.3)	30.7(31.2)	22.0(10.6)	66.6(34.2)
LSTM	76.1(8.3)	67.9(5.53)	28.8(10.0)	37.8(7.9)

6. Conclusion

In this paper we present the MAHNOB Mimicry database, a set of highly-accurately synchronised multi-sensory audio-visual recordings of naturalistic dyadic interactions, suitable for investigating mimicry and negotiation behaviour. The amount of mimicry data captured and annotated is significant. The database can be used for other applications as well, including facial point tracking, continuous interest prediction, or automatic speech recognition. The database is not yet fully annotated, however raw data and current annotations are publicly available for non-commercial use, at http://mahnob-db.eu/mimicry. In this paper we also presented experimental studies considering motor mimicry of facial and head movements and widely-used classifiers including two similarity-based methods (cross-correlation and generalized time warping), and a state-of-the-art temporal classifier (LSTM-RNN).

Acknowledgment

Current work of the team is supported by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA).

References

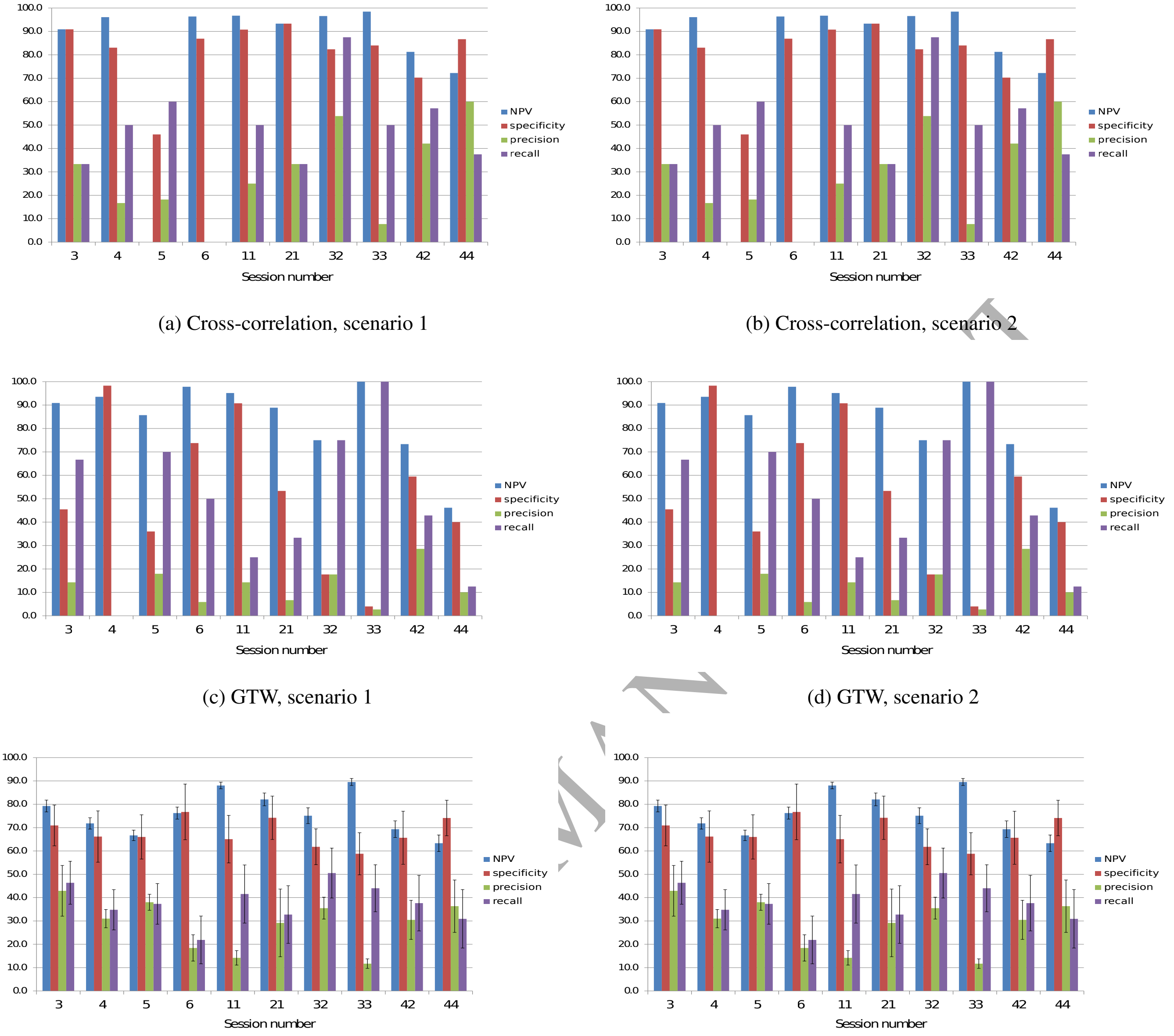
[1] Altmann, U., 2011. Investigation of movement synchrony using windowed cross-lagged regression, in: Analysis of Verbal and Nonverbal Communication and Enactment.. Springer, pp. 335–345.

[2] Bilakhia, S., Petridis, S., Pantic, M., 2013. Audiovisual detection of behavioural mimicry, in: IEEE Affective Computing and Intelligent Interaction, IEEE.

[3] Boker, S., Xu, M., Rotondo, J., King, K., 2002. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. Psychological Methods 7(3), 338–355.

[4] Burgoon, J., Stern, L., Dillman, L., 2007. Interpersonal adaptation: Dyadic interaction patterns. Cambridge University Press.

[5] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S., 2008. Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation 42(4), 335–359.



(a) Cross-correlation, scenario 1

(b) Cross-correlation, scenario 2

(c) GTW, scenario 1

(d) GTW, scenario 2

(e) LSTM, scenario 1

(f) LSTM, scenario 2

Fig. 5: Experimental results for cross-correlation, time-warping, and LSTM classifiers. Scenario 1 tested classification of facial expression mimicry, whilst scenario 2 tested for mimicry of both facial expressions and head movement. Error bars show standard-deviation over 10 independent experiment repetitions)

[6] Chartrand, T., Bargh, J., 1999. The chameleon effect: The perception-behavior link and social interaction. Journal of personality and social psychology 76, 893–910.

[7] Delaherche, E., Boucenna, S., Karp, K., Michelet, S., Achard, C., Chetouani, M., 2013. Social coordination assessment: Distinguishing between shape and timing, in: Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction. Springer, pp. 9–18.

[8] Delaherche, E., Chetouani, M., 2010. Multimodal coordination: exploring relevant features and measures, in: Proc. 2nd international workshop on Social signal processing, ACM. pp. 47–52.

[9] Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D., 2012. Interpersonal synchrony: A survey of evaluation methods across disciplines. Affective Computing, IEEE Transactions on 3(3), 349–365.

[10] Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., House, D., 2010. Spontal: A swedish spontaneous dialogue corpus of audio, video and motion capture., in: LREC, pp. 2992–2995.

[11] Feese, S., Arnrich, B., Trosboker2002windowedter, G., Meyer, B., Jonas, K., 2012. Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion, in: IEEE International Conference on Social Computing (SocialCom), pp. 520–525.

[12] Gatica-Perez, D., 2009. Automatic nonverbal analysis of social interaction in small groups: A review. Image and Vision Computing 27(12), 1775–1787.

[13] Gueguen, N., Jacob, C., Martin, A., 2009. Mimicry in social interaction: Its effect on human judgment and behavior. European Journal of Social Sciences 8(2), 253–259.

[14] Gunes, H., Pantic, M., 2010. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners, in: Intelligent virtual agents, Springer. pp. 371–377.

[15] Hess, U., Fischer, A., 2013. Emotional mimicry as social regulation. Personality and Social Psychology Review .

[16] Lakin, J., Jefferis, V., Cheng, C., Michelle, M., Chartrand, T., 2003. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. Journal of nonverbal behavior 27(3), 145–162.

[17] Lichtenauer, J., Pantic, M., 2011. Monocular omnidirectional head motion capture in the visible light spectrum, in: IEEE ICCV Workshop,

2011, pp. 430–436.

[18] McDonald, S., Li, S., De Sousa, A., Rushby, J., Dimoska, A., James, C., Tate, R.L., 2011. Impaired mimicry response to angry faces following severe traumatic brain injury. Journal of clinical and experimental neuropsychology 33(1), 17–29.

[19] McDuff, D., El Kaliouby, R., Senechal, T., Amr, M., Cohn, J.F., Picard, R., 2013. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected" in-the-wild", in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on, IEEE. pp. 881–888.

[20] McIntosh, D.N., Reichmann-Decker, A., Winkielman, P., Wilbarger, J.L., 2006. When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism. Developmental science 9(3), 295–302.

[21] McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M., 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. Affective Computing, IEEE Transactions on 3(1), 5–17.

[22] Michelet, S., Karp, K., Delaherche, E., Achard, C., Chetouani, M., 2012. Automatic imitation assessment in interaction, in: Human Behavior Understanding. Springer, pp. 161–173.

[23] Oertel, C., Cummins, F., Edlund, J., Wagner, P., Campbell, N., 2013. D64: A corpus of richly recorded conversational interaction. Journal on Multimodal User Interfaces 7(1-2), 19–28.

[24] Oikonomopoulos, A., Patras, I., Pantic, M., 2011. Spatiotemporal localization and categorization of human actions in unsegmented image sequences. Image Processing, IEEE Transactions on 20(4), 1126–1140.

[25] Orozco, J., Rudovic, O., Gonzalez, J., Pantic, M., 2013. Hierarchical online appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises. Image and Vision Computing .

[26] Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K., 2010. The nomco multimodal nordic resource–goals and characteristics .

[27] Petridis, S., Martinez, B., Pantic, M., 2013. The mahnob laughter database. Image and Vision Computing 31(2), 186–202.

[28] Pickering, M., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behavioral and brain sciences 27(2), 169–189.

[29] Ramseyer, F., Tschacher, W., 2006. Synchrony: A core concept for a constructivist approach to psychotherapy. Constructivism in the human sciences 11(1), 150–171.

[30] Rozgic, V., Xiao, B., Katsamanis, A., Baucom, B.R., Georgiou, P.G., Narayanan, S.S., 2010. A new multichannel multi modal dyadic interaction database., in: INTERSPEECH, pp. 1982–1985.

[31] Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., Rückstieß, T., Schmidhuber, J., 2010. PyBrain. Journal of Machine Learning Research 11, 743–746.

[32] Schmidhuber, J., Hochreiter, S., 1997. Long short-term memory. Neural Comput 9, 1735–1780.

[33] van Son, R., Wesseling, W., Sandersa, E., van den Heuvel, H., 2008. The ifadv corpus: a free dialog video corpus., in: LREC, pp. 501–508.

[34] Sonnby-Borgström, M., Jönsson, P., Svensson, O., 2003. Emotional empathy as related to mimicry reactions at different levels of information processing. Journal of Nonverbal behavior 27(1), 3–23.

[35] Sun, X., Lichtenauer, J., Valstar, M., Nijholt, A., Pantic, M., 2011a. A multimodal database for mimicry analysis, in: Affective Computing and Intelligent Interaction. Springer, pp. 367–376.

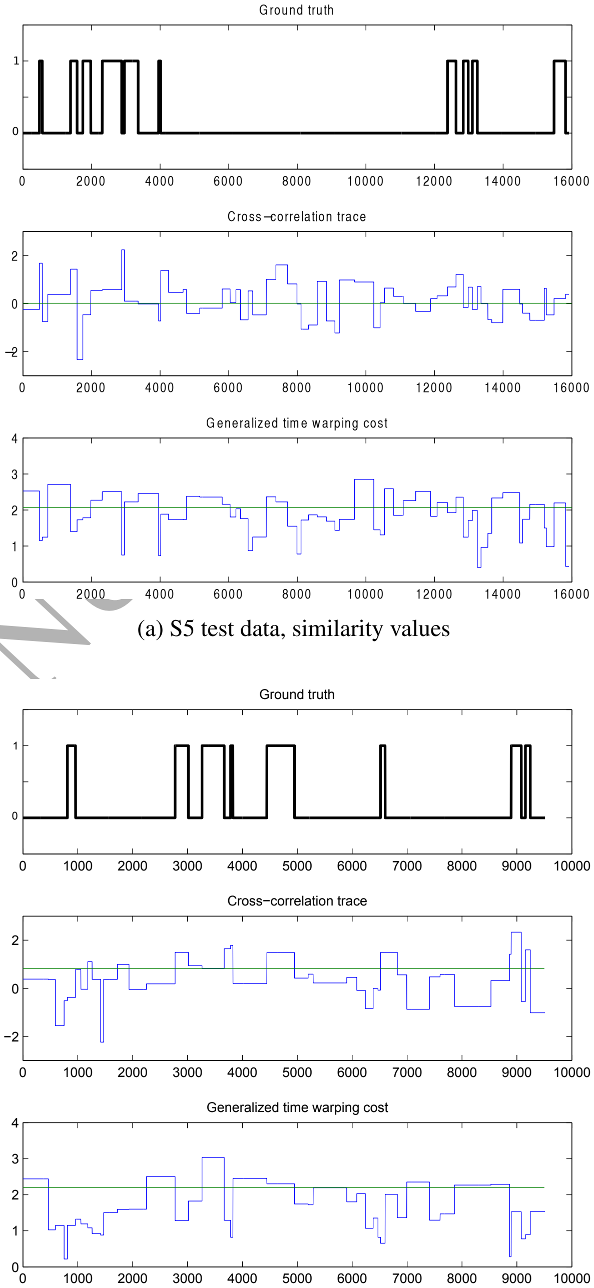
[36] Sun, X., Truong, K., Nijholt, A., Pantic, M., 2011b. Automatic visual mimicry expression analysis in interpersonal interaction, in: IEEE CVPR Workshop, 2011, pp. 40–46.

[37] Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., Schröder, M., 2012. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. Affective Computing, IEEE Transactions on 3(1), 69–87.

[38] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H., 2006. Elan: a professional framework for multimodality research, in: Proceedings of LREC.

[39] Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(1), 39–58.

[40] Zhou, F., Torre, F.d.l., 2012. Generalized time warping for multi-modal alignment of human motion, in: IEEE CVPR 2012, IEEE. pp. 1282–1289.



(a) S5 test data, similarity values

(b) S32 test data, similarity values

Fig. 6: Similiarity values for test data from S5 and S32 (scenario 1). Note that high similarity values can occur for sequences with no (facial) mimicry present, e.g. for sequences where both subjects have static facial pose. High values for cross-correlation, and low values for time-warping cost, indicate "similar" seqences. A ground truth value of 1 indicates a sequence containing a mimicry episode. Green horizontal lines represent decision boundaries.