2013 Humaine Association Conference on Affective Computing and Intelligent Interaction

Audiovisual detection of behavioural mimicry

Sanjay Bilakhia
Department of Computing
Imperial College London
Email: sb1006@imperial.ac.uk

Stavros Petridis
Department of Computing
Imperial College London
Email: sp104@imperial.ac.uk

Maja Pantic
Imperial College London
Faculty of EEMCS
University of Twente
Email: m.pantic@imperial.ac.uk

***Abstract*—Human mimicry is a behavioural cue occurring during social interaction that can inform us about the participants' inter-personal states and attitudes. It occurs when a participant in an interaction exhibits some behaviour as a result of a co-participants prior display of that signal, and occurs on both short and long time-scales. To develop a detection method for such behaviour, we use a method based on feature prediction, where we train an ensemble of regression models from one subject's features to the co-subject's features, for each class. The ensemble of models with lowest reconstruction error is used to detect mimicry and non-mimicry, using continuous audiovisual streams. As mimicry events are dynamical phenomena, we use a temporal regression model (long short-term memory neural networks) to capture sequential dependencies in the data. On a data set of ten 12-minute dyadic interaction episodes, our method gave average positive and negative recall rates of 77.5% and 60.0% respectively, on data with significant class imbalances, due to the relative sparsity of mimicry samples when doing continuous detection.**

I. Introduction

Mimicry has been demonstrated to be an important part of human social interactions [1] and its implicit social signals have attracted increasing attention from not only psychologists but also from HCI researchers [2]. Mimicry has been operationalized in varying ways by psychology literature, and has overlap with the related concepts of interactional synchrony, interactive alignment, and convergence. It serves as an important indicator of cooperativeness and empathy during interaction. Recent research has explored whether people non-consciously exploit mimicry to gain a social advantage [3]. People can consciously or non-consciously mimic the behaviour of others, because their goals activate behavioural strategies which may aid in pursuing those goals [4], [5]. Individuals can mimic many different aspects of their interaction partners, including speech patterns, facial expressions, emotions, moods, postures, gestures, mannerisms, and idiosyncratic movements [6],[7]. While non-conscious mimicry may not be consciously perceived, it can be fostered or inhibited by social, motivational, cognitive, and affective conditions, and has been assumed to play a role in social glue [1], [8], to breed feelings of similarity, affiliation, and sympathy [9]. The more mimicry observed, the more smoothly an interaction is perceived, as participants or confederates who are being mimicked are more willing to alter the way in which they interact with others, in order to share similar affect, express understanding, and obtain more agreement[3],[10]. In this work we present a method to continuously detect mimicry episodes, using class-specific regression models to reconstruct the feature stream of a participant in a dyadic interaction, from the other participant's feature stream. A sample at time point t is given the label of the regression models with the lowest reconstruction error. We propose that the regression models of a particular class will more accurately model the relationship between the feature streams, when an episode of that class is occurring. For this study, we investigated mimicry of smiles, laughter, and linguistic vocalization (“hmm”,“yeah”). As such, we restricted the labelling of positive examples of mimicry to solely those episodes containing these behaviours. For this work, we define a mimicry episode as simple synchronous matching behaviour - an occurrence of a participant in an interaction exhibiting some behaviour as a result of their co-participants prior display of that behaviour, on a range of different time-scales. We used data from the MAHNOB Mimicry database [11]. To the best of our knowledge this is the first work that attempts continuous detection of behavioural mimicry, using data of natural interaction between minimally constrained subjects, and evaluates predictive performance against ground truth. We obtain promising results, with mimicry being detected successfully in many episodes, however the method is prone to variability in the reconstruction error, which can generate excessive false positives.

II. Prior work

Some initial work has been done on the learning of predictive models for mimicry detection. Multiple authors have developed methods for detection relying on the construction of a “junk” set, which randomly permutes each subject's data independently, either acting on windows of grouped samples, or as individual samples. This is then used to calculate an artificial baseline measure. A synchrony score is calculated on both the junk and real datasets, and some statistical test is used to either highlight variables that might be particularly informative, or temporal windows that have a particularly high similarity. Ramseyer et al. [12] calculated cross-correlation of motion energy features in one minute windows, with both positive and negative time lags; these scores were aggregated into a global score and compared between real and (window-level) permuted data. Delaherche and Chetouani [13] modelled global coordination between movement features (motion energy, motion history, hand trajectory) and prosodic features (pitch, energy, pause, and vocalic energy) for dyadic interactions, using Pearsons correlation and magnitude coherence (correlation of the signals' Fourier transform) between all feature-feature pairs. Synchrony was inferred for pairs of features where the correlation measures were further than 2 standard deviations from those produced using a sample-level random permutation. Michelet et al. [14] extracted spatio-temporal interest points, of which they calculated neighbourhood statistics. These histograms were concatenated, quantized into bags-of-words using k-means clustering, and locally-constrained dynamic time warping was used as a similarity measure. A threshold was then used to discriminate between mimicry and non-mimicry states in short windows. Barbosa et al. [15] used lagged cross correlation to measure motion co-ordination of lip and tongue movements during repetitious speech between two subjects, while [16] and [17] used windowed cross-correlation and peak-picking to investigate symmetry in head movement time

978-0-7695-5048-0/13 $26.00 © 2013 IEEE
DOI 10.1109/ACII.2013.27

series. Boker and Rotondo [18] used lagged windowed cross-correlation to investigate symmetry in full-body motion capture data of free-form spontaneous dance. Sun et al. [19] used the correlation of motion intensity histograms on video data of natural interactions, demonstrating that after some duration of face-to-face interaction, participants have the tendency to adopt body postures, head movements, hand gestures and linguistic idiosyncrasies of their co-participants. In an alternative approach to cross-correlation based models, [20] computed 2 linear regression models per window; one model contained both auto-regressive and cross-regressive components, while the other 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 reduction in unexplained variance by incorporation of cross-regressive terms was non-random, and hence an indicator of mimicry.

Recurrence analysis is another approach that has been used to detect movement synchronization. This applies a similarity metric to all pairs of samples from each sequence $\{(x_i, y_{j)} \mid x_i \in X, y_y \in Y, i = 1...N, j = 1...M,\}$ where X and Y represent each sequence of features, to give an NxM matrix, over which a threshold is taken to give a binary similarity matrix. The diagonal structures in this matrix then represent periods where the processes had a similar trajectory through the phase space, and its entropy can be used as a measure of process similarity. Such analysis has been used with motion capture data of body posture in dyadic interaction [21][22] where participants were asked to complete some joint task such as a co-operative visual puzzle, or the pronunciation of pairs of words subject to varying environmental conditions, to observe the effect on mimicry on postural sway.

Out of the studies above, only [14] develop a predictive model which is tested against ground truth, reporting AUC measure, while others generally investigate some similarity measure (assumed to be a suitable indicator of mimicry behaviour) in relation to some other variable (such as outcome success of therapeutic treatment) or hypothesis (such as presumed increase in mimicry over the course of an interaction).

III. Dataset

We used a relatively new multimodal database, containing mimicry episodes as they occur in naturalistic dyadic interactions, the MAHNOB Mimicry Database [11]. The experiments are designed to explore the relationship between the occurrence of mimicry and human affect. The corpus is recorded using ambient and individual close-talk fixed microphones, individual cameras from 6 frontal and 1 overhead view(s), and a profile-view wide-angle camera. All output signals were exactly synchronized using external triggers. Video data was recorded at 58 frames/second, and audio was sampled at 48kHz. The dataset consists of 54 recordings of dyadic face-to-face interactions: 34 are discussions on a political topic, and 20 are conversations situated in a role-playing game. Each session is between 5 and 20 minutes long. The subjects consist of 40 participants and 3 confederates, across a range of ethnic backgrounds and first languages. This data has been partially annotated for multiple behaviours, including dialogue acts, head gestures, hand gestures, body movement and facial expression, and mimicry episodes. Due to only partial availability of annotations, we used 10 sessions, with a session length median of 14 minutes. Table I presents statistics about the subjects used in this work. We define a mimicry episode as an occurrence of a participant in an interaction exhibiting a behaviour as a result of their co-participants prior display of that signal. The episode onset is taken to be the onset of the mimickee's action subsequently manifested by the mimicker, whilst the offset is taken to be the offset of the mimicker's display of that action. The upper bound on the time lag between mimickee action offset and mimicker action onset is set at 4 seconds - an even longer delay would be unlikely to be mimicry. Mimicry behaviours may occur multiple times within the same episode, either due to overlapping occurrences (ie. the onset of a behavior to be mimicked occurs before the offset of a previously mimicked behavior), or "reflective" mimicry, i.e. subject 2 mimicking an action of subject 1, which is subsequently mimicked by subject 1, as in contagious laughter.

Session #	Class size		Episode length mean/var		Session length
	Mimicry	Non-mimicry	Mimicry	Non-mimicry	
3	1273	24532	254/60	4583/6021	7m24s
4	2714	52685	226/126	4777/4466	15m54s
5	4040	52647	237/140	3284/7342	16m16s
6	2146	54016	214/94	5369/5277	16m07s
11	2350	52105	195/86	4299/4978	15m38s
21	1967	33057	281/126	4696/2393	10m03s
32	4800	32087	228/152	1515/1750	10m36s
33	1072	54826	172/62	9137/6093	16m03s
42	6009	36651	214/104	1307/1598	12m14s
44	3833	14384	212/124	845/1125	5m13s

TABLE I: Session statistics (class size and episode length reported as number of samples, session length as time in seconds)

IV. Features

Audio features: Cepstral features, such as MFCCs, have been widely used in speech recognition, language-identification, and discrimination between linguistic/non-linguistic vocalizations. We use the first 6 MFCCs, computed every 10ms, over a window of 100ms, giving a frame rate of 100 frames/second.
Visual features: Changes in facial expression are captured using the point tracker described in [23], which uses an online appearance model to track rigid head movements and non-rigid facial motion, using 113 landmark facial points. It also decouples this movements to output MPEG-4 facial animation parameter (FAP) estimates, corresponding to mouth width, mouth height, eyebrow pose etc.

V. Long short-term memory networks

The long short-term memory network (LSTM) is a neural network model that can preserve long-range dependencies and contextual information in sequential data. They are an advancement on standard recurrent neural networks trained with gradient methods, for which training is very difficult, especially with long input sequences. As the network is unrolled through time, the error signal tends to zero or divergence as it travels backward through the network layers. Each backwards pass through a "neuron" scales the error signal by the derivative of the neurons activation function multiplied by the neuron output's connection weight. As error is backpropagated, this scaling is repeatedly applied to the error term. If the scaling is consistently less than 1, the error will vanish, leading to negligible weight updates and extremely slow convergence; if consistently greater than 1, the error term will diverge. LSTMs preserve the error signal by forcing the scaling to 1, using a linear activation function with derivative equal to 1, and recurrent connection weight equal to 1. This allows them to maintain unscaled activation values and error derivatives across arbitrary time scales. As neural networks require nonlinear hidden unit activation functions to be able to represent arbitrary non-linear functions, each hidden memory cell′s state is squashed with a sigmoid before being passed on to the rest of the network. LSTM inputs/outputs are also "gated" to control internal state. If a memory cell stores information that is only useful later in a sequence, this (currently) irrelevant information may reduce performance in the interim, causing it to be discarded, which sacrifices overall performance. Memory cells may also be perturbed by irrelevant input, causing information relevant later in the sequence to be lost. Each cell therefore has its net input and output modulated by input and output gates, respectively, allowing a context-sensitive way to update its internal state, shield state information from interference, and protect downstream units from perturbation by (currently irrelevant) stored information. A forget gate scales the activation from the previous time-step, so memory cells can be cleared after the current state has become irrelevant.

VI. Methodology

We adapt a method first suggested in [24], where each feature vector is split into two disjoint subsets - one subset of features is reconstructed from the other using a class-specific regression model, and the model with minimum reconstruction error classifies the sample. In our case, our subsets are the subject-specific audiovisual features. For each of the two classes, mimicry and non-mimicry, we train a regression model from the first subject's features to the second subject's features, and vice versa. This is done for multiple time lags, both positive and negative, to account for subject reaction time, and directionality of mimicry. We use the long short-term memory network [25] as our underlying regression model to account for sequential dependencies in our data, without resorting to concatenation of multiple samples from a window into one very large feature vector. The relationship between the subject 1 and subject 2's features for both mimicry S_1^M, S_2^M and non-mimicry $S_1^{\bar{M}}, S_2^{\bar{M}}$ is modelled by $f^M_{S_1\to S_2}$, $f^M_{S_2\to S_1}$ for mimicry and $f^{\bar{M}}_{S_1\to S_2}$, $f^{\bar{M}}_{S_2\to S_1}$ for non-mimicry as follows:

$$f^M_{S_2\to S_1}(S_2^M) = \hat{S_1}^M \approx S_1^M \tag{1}$$

$$f^M_{S_1\to S_2}(S_1^M) = \hat{S_2}^M \approx S_2^M \tag{2}$$

$$f^{\bar{M}}_{S_2\to S_1}(S_2^{\bar{M}}) = \hat{S_1}^{\bar{M}} \approx S_1^{\bar{M}} \tag{3}$$

$$f^{\bar{M}}_{S_1\to S_2}(S_1^{\bar{M}}) = \hat{S_2}^{\bar{M}} \approx S_2^{\bar{M}} \tag{4}$$

Once the model parameters are learnt, an unseen example is given the label of the pair of class-specific models that produce the lowest reconstruction error. When new samples are available (for both subjects), the audio and visual features are computed, and are then fed to the models from eq. 1, 2, 3, 4, and 4 error values are produced. We use mean squared error (MSE) to scalarize the vector of reconstruction errors.

$$e^M_{S_2\to S_1} = MSE(\hat{S_1}^M, S_1^M) \tag{5}$$

$$e^M_{S_1\to S_2} = MSE(\hat{S_2}^M, S_2^M) \tag{6}$$

$$e^{\bar{M}}_{S_2\to S_1} = MSE(\hat{S_1}^{\bar{M}}, S_1^{\bar{M}}) \tag{7}$$

$$e^{\bar{M}}_{S_1\to S_2} = MSE(\hat{S_2}^{\bar{M}}, S_2^{\bar{M}}) \tag{8}$$

We then compute a weighted mean of the MSE, for each class, as shown in eq. 9 and 10, where $w_M, w_{\bar{M}}$ are parameters optimized using gridsearch during model selection on the validation set.

$$e^M = w_M \times e^M_{S_2\to S_1} + (1 - w_M) \times e^M_{S_1\to S_2} \tag{9}$$

$$e^{\bar{M}} = w_{\bar{M}} \times e^{\bar{M}}_{S_2\to S_1} + (1 - w_{\bar{M}}) \times e^{\bar{M}}_{S_1\to S_2} \tag{10}$$

A frame is classified as mimicry or non-mimicry depending on which pair of models (corresponding to a particular class) produced the best feature reconstruction, i.e. the pair with the lowest combined reconstruction error:

$$IF \;\; e^M > e^{\bar{M}} \;\; THEN \;\; \bar{\mathbf{M}} \;\; ELSE \;\; \mathbf{M} \tag{11}$$

VII. Experimental studies

A. Pre-processing steps

We split our data into training, validation and test sets on a per-session basis, as mimicry behaviours vary considerably between different pairs of subjects. The training set consisted of the first contiguous block of the session such that it contained half of all the mimicry episodes. This was then split into individual sequences and used for training. The contiguous block containing the next quarter of all the mimicry episodes formed the validation set, and the remaining data was used for testing and performance evaluation. Before training, all features are z-normalized (per session) to zero mean and unit standard deviation, and smoothed using a Savitzky-Golay filter of window size 15 and degree 3.

B. Training

Mimicry and non-mimicry models are trained with sequences from their respective classes only. We use an ensemble of the classifiers detailed above, with lags of {-24,0,24} samples, corresponding to time lags of {-0.5,0,0.5} seconds. This sits in between the time scales of low-latency motor mimicry, and emotional mimicry which involves higher-level (slower) cognitive processes [26]. As many mimicry episodes were short, models with longer time lags would have had even less training data per session than currently available, due to the need to clip the ends of each training sequence after time-shifting one relative to the other (for example, when using a 150 sample length sequence to train a model with a lag of 58 frames ≡ 1s, clipping the sequence after time-shifting would lose 40% of the data for that sequence!). Preliminary experiments also showed that including longer time lags had no meaningful effect on performance, for this model. We define lag relative to subject 1, hence a model with a negative lag implies that it models the relationship between data from subject 2 with earlier data from subject 1. So, as shown in Fig.1, for each class we train regression models to predict the audiovisual features at t in stream 2 based on the features at $t - 24$ in stream 1 (and vice versa), and models to predict the

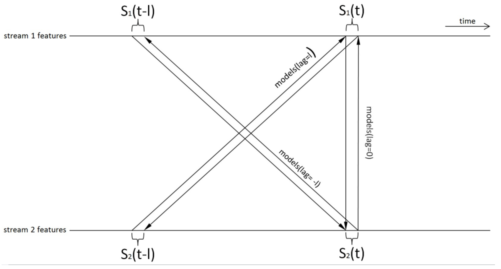

Fig. 1: Regression models at different time lags each generate a frame label for time t, which are combined using majority voting

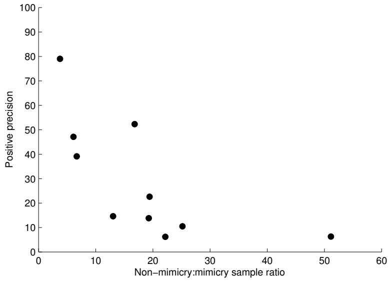

Fig. 2: Positive precision has an approximate inverse relationship to severity of class imbalance in the data

audiovisual features at t in stream 1 based on the features at $t-24$ in stream 2 (and vice versa), as well as models to predict the features at t in stream 1 from t in stream 2 (and vice versa). Models with time-lags suffer from inevitable edge effects (e.g. when training with the first sample in a session, there are no prior samples to train a time-lagged model with); rather than zero-pad the sequence ends, we clip those samples that have no corresponding samples (at the correct time) to train with. We use the LSTM implementation from the PyBrain neural network library [27].

C. Labelling procedure

After the regression models for each class have produced a reconstruction of their complementary features, the error values $e^{M}_{S_2\rightarrow S_1}$, $e^{M}_{S_1\rightarrow S_2}$, $e^{\bar{M}}_{S_2\rightarrow S_1}$, $e^{\bar{M}}_{S_1\rightarrow S_2}$ are smoothed using a Savitzky-Golay filter, with a window size of 29 frames, and degree 5. The reconstruction errors from each pair of regressors are then compared to generate a label prediction as per eq.11. As mentioned above, we use an ensemble of classifiers with different time lags, each of which produces a label for a given sample. Therefore each frame is labelled 3 times. These "votes" are then combined using a majority-voting decision rule. The performance measures we use are precision and recall. Note that we are not classifying pre-segmented sequences, rather we are performing classification on individual frames along the entire length of the sequence. Training of each regression model is performed using pre-segmented sequences (as they are only trained using data from their respective classes), however labelling of new frames is done continuously, to take advantage of the stateful LSTM model.

D. Model selection

We trained networks using only one hidden layer. The number of hidden neurons was optimized using a line search across the range [25-75] in steps of 10, where the hidden layer size for networks in both classes was constrained to be equal. Networks were trained using resilient back-propagation, with a training epoch limit of 500. Our method also requires optimization of the weights w_M and $w_{\bar{M}}$, with respect to classification performance. This is performed using a single-resolution grid-search in steps of 0.001, subject to $0 \leq w_M$, $w_{\bar{M}} \leq 1$. The best performing model is chosen (using f1 performance as a selection criterion), and tested as below.

E. Testing

The (resilient) back-propagation training algorithm when applied to an LSTM network, solves a local optimization of an error function with (potentially) many local minima - the local minimum the network finishes in is dependent on its initialization point in the weight-space. This initial point is instantiated randomly, so different training runs may end up in different minima. To account for this non-determinism, during testing we train and test each model 5 times, reporting mean and standard deviation.

VIII. Results

TABLE II: Class-specific precision and recall measures for detection of mimicry of laughter, smiles, and linguistic vocalization

Session #	Non-mimicry		Mimicry	
	precision	recall	precision	recall
3	93.8 (1.1)	56.2 (1.4)	13.8 (1.2)	65.4 (6.6)
4	95.4 (1.5)	63.4 (1.2)	22.6 (1.3)	77.8 (8.0)
5	98.3 (1.0)	56.1 (2.4)	14.6 (1.5)	88.4 (6.0)
6	98.9 (3.6)	61.3 (1.2)	10.5 (6.8)	86.8 (4.1)
11	97.8 (4.8)	53.5 (4.7)	6.2 (0.5)	71.4 (6)
21	83.3 (4.2)	69.9 (3.2)	52.3 (5.4)	70.2 (7.7)
32	95.2 (0.8)	63.9 (0.9)	39.1 (0.7)	87.7 (2.2)
33	98.9 (4.9)	49.2 (3.6)	6.3 (0.5)	84.4 (8.2)
42	91.2 (1.9)	63.3 (1.5)	47.1 (1.9)	84.2 (3.4)
44	40.6 (2.0)	63.9 (3.0)	79 (1.8)	59.2 (1.8)

Table II shows the experimental results on 10 full sessions of the MAHNOB Mimicry database. We can see that the performance is highly session dependent, however the models have a bias towards labelling a frame as mimicry, as shown by the generally high positive recall performance. This may be due to the significant class imbalance in the data.

Although our method is not directly discriminating between the two class distributions in the feature space, the abundance of non-mimicry data may allow the non-mimicry model to learn a smoother approximation between the two sets of features, allowing better generalization. Even after filtering, the high-frequency noise in the mimicry model error is more prominent than in the non-mimicry model error. This noise seems to cause the false positives when reconstruction error is low for both models, examples of which can be seen in Fig. 3. For our experiments we limited filtering to a 29 frame window, to avoid removing small variations on the order of 0.5s that could be informative of real expressions. However the persistent presence of noise suggests that a more aggressive smoothing of the reconstruction errors might improve performance; it may prove useful to include the smoothing parameters in the optimization of the weights w_M, and $w_{\bar{M}}$ during model selection. We can also see in Fig. 2 that there is an inverse relationship between the ratio of the class sizes, and the positive precision. The variances in performance for positive precision and recall is also significantly higher than for negative precision and recall, suggesting that the solutions found by the networks for the mimicry models are relatively unstable. The networks for both mimicry and non-mimicry were constrained to have an equal number of hidden neurons, to reduce the size of the model search space, however this may have lead to overfitting of the mimicry models, as the ratio between the number of parameters to optimize (network weights) and available data becomes too high. Resampling to artificially balance class-sizes would be of little use in this case, since we are not directly discriminating in the feature space, whilst subsampling the non-mimicry data may lead to instability and overfitting for the non-mimicry models as well.

However we can see that our method can successfully detect boundaries between mimicry and non-mimicry in some cases, as in Fig.3. The sequence corresponding to those frames is shown in Fig.4. This sequence has very obvious vocalized laughter, smile and linguistic mimicry, and furthermore the difference between mimicry and non-mimicry states for these subjects is well defined, and hence is detected relatively easily from the surrounding non-mimicry states. Furthermore, the false positives later on in the sequence, between frame indices 2000-3000, are in this case not entirely wrong; during this segment there is mimicry of head nods, with very significant head movement from subject 2. Since the face tracker used is not perfect at decoupling head movement from facial expressions, and the subjects in session 32 nod very vigorously, these nods are (though heavily damped) still present in the features for normalized facial movements. Hence the models may have unintendedly learnt how to detect nod mimicry, as nod mimicry frequently co-occurs with smile mimicry in the training data (and hence with positively labelled training data, as we only labelled data as positive if they contained smile, laughter or linguistic mimicry, for this work). In other sessions, such as session 11, one subject had an extremely wide variety of highly animated expressions while discussing a political topic, and the difference in expression between mimicry and non-mimicry states was poorly defined; another subject in session 3 had an extremely small amplitude of expression and unvarying style of speech throughout. These, coupled with the class imbalance (e.g. for session 11, the negative-to-positive ratio is 25), may have depressed the positive precision significantly.

IX. Conclusion

We presented a method to detect mimicry behaviour in audiovisual data of naturalistic dyadic interaction, using a temporal regression model, long short-term memory networks, to reconstruct one subject's behaviour from the other. Our method can perform reasonably well, with promising positive and negative recall rates of 77.5% and 60% respectively, but is sensitive to a negative-to-positive sample ratio that is extremely large, or insufficient amplitude of gesture and expression. Class imbalance will be a significant problem to overcome for future mimicry detection methods, and model complexity will be have to be controlled to prevent noise overpowering subtle behavioural cues.

X. Acknowledgements

This work has been supported by the European Communitys 7th Framework Programme [FP7/20072013] under grant agreement no. 231287 (SSPNet). The work of Maja Pantic was supported by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

References

[1] Tanya L Chartrand, John A Bargh, et al., "The chameleon effect: The perception-behavior link and social interaction," *Journal of personality and social psychology*, vol. 76, pp. 893–910, 1999.

[2] Jeremy N Bailenson and Nick Yee, "Digital chameleons automatic assimilation of nonverbal gestures in immersive virtual environments," *Psychological science*, vol. 16, no. 10, pp. 814–819, 2005.

[3] Tanya L Chartrand and Valerie E Jefferis, "Consequences of automatic goal pursuit and the case of nonconscious mimicry," *Social judgments: Implicit and explicit processes*, pp. 290–305, 2003.

[4] Baumeister et al., "Effects of social exclusion on cognitive processes: Anticipated aloneness reduces intelligent thought," *Journal of personality and social psychology*, vol. 83, no. 4, pp. 817–827, 2002.

[5] Nicolas Gueguen, Celine Jacob, and Angelique Martin, "Mimicry in social interaction: Its effect on human judgment and behavior," *European Journal of Social Sciences*, vol. 8, no. 2, pp. 253–259, 2009.

[6] Jessica L Lakin and Tanya L Chartrand, "Exclusion and nonconscious behavioral mimicry," *The social outcast: Ostracism, social exclusion, rejection, and bullying*, pp. 279–296, 2005.

[7] Jessica L Lakin, Valerie E Jefferis, Clara Michelle Cheng, and Tanya L Chartrand, "The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry," *Journal of nonverbal behavior*, vol. 27, no. 3, pp. 145–162, 2003.

[8] Jessica L Lakin and Tanya L Chartrand, "Using nonconscious behavioral mimicry to create affiliation and rapport," *Psychological Science*, vol. 14, no. 4, pp. 334–339, 2003.

[9] Harald G Wallbott, "Congruence, contagion, and motor mimicry: Mutualities in nonverbal exchange," *Mutualities in dialogue*, pp. 82–98, 1995.

[10] R.Vonk M.Stel, "Mimicry in social interaction: benefits for mimickers, mimickees, and their interaction," *British Journal of Psychology*, vol. 101, no. 2, pp. 311–323, 2010.

[11] Xiaofan Sun, Jeroen Lichtenauer, Michel Valstar, Anton Nijholt, and Maja Pantic, "A multimodal database for mimicry analysis," in *ACII*, pp. 367–376. 2011.

[12] Fabian Ramseyer and Wolfgang Tschacher, "Synchrony: A core concept for a constructivist approach to psychotherapy," *Constructivism in the human sciences*, vol. 11, no. 1, pp. 150–171, 2006.

[13] Emilie Delaherche and Mohamed Chetouani, "Multimodal coordination: exploring relevant features and measures," in *Proc. of 2nd International workshop on Social signal processing*. ACM, 2010, pp. 47–52.

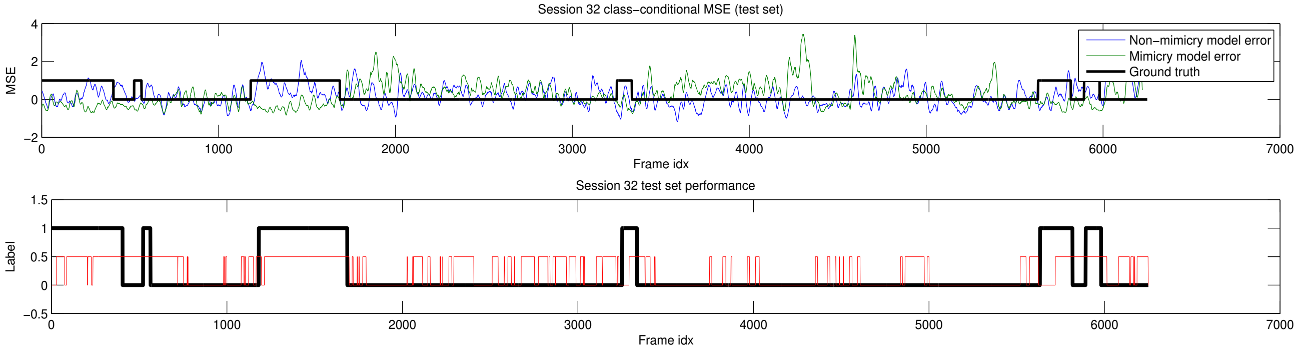

Fig. 3: Model error and subsequent frame classification on session 32's test set.

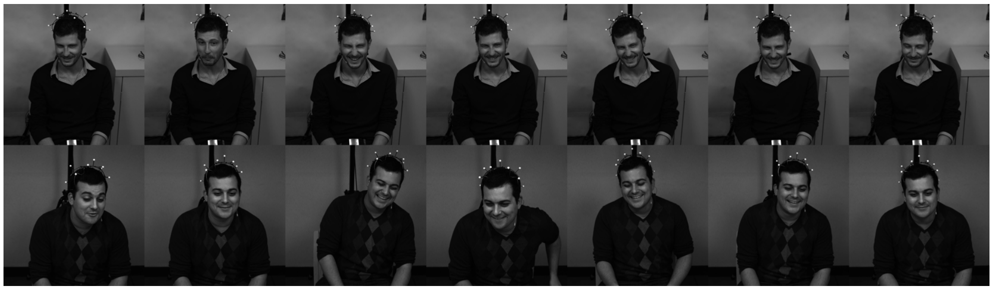

Fig. 4: Sequence corresponding to frame indices 1100-1700 in Fig. 3 (sampled every 100 frames)

[14] Stéphane Michelet, Koby Karp, Emilie Delaherche, Catherine Achard, and Mohamed Chetouani, "Automatic imitation assessment in interaction," in *Human Behavior Understanding*, pp. 161–173. Springer, 2012.

[15] Adriano V Barbosa et al., "An instantaneous correlation algorithm for assessing intra and inter subject coordination during communicative behavior," in *Modeling Human Communication Dynamics Workshop, NIPS*, 2010, p. 38.

[16] S.Boker et al., "Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series," *Psychological Methods*, vol. 7, no. 3, pp. 338–355, 2002.

[17] K. Ashenfelter et al., "Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation," *Journal of experimental psychology. Human perception and performance*, vol. 35, no. 4, pp. 1072, 2009.

[18] J. Rotondo S.Boker, "Symmetry building and symmetry breaking in synchronized movement," *ADVANCES IN CONSCIOUSNESS RESEARCH*, vol. 42, pp. 163–174, 2002.

[19] Xiaofan Sun, KP Truong, A Nijholt, and Maja Pantic, "Automatic visual mimicry expression analysis in interpersonal interaction," in *CVPR Workshops, 2011*. IEEE, 2011, pp. 40–46.

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

[21] K. Shockley et al., "Mutual interpersonal postural constraints are involved in cooperative conversation," *Journal of Experimental Psychology-Human Perception and Performance*, vol. 29, no. 2, pp. 326–332, 2003.

[22] K.Shockley, "Articulatory constraints on interpersonal postural coordination," *Journal of experimental psychology. Human perception and performance*, vol. 33, no. 1, pp. 201, 2007.

[23] J. Orozco, O. Rudovic, J. Gonzlez, and M. Pantic, "Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises," *Image and Vision Computing*, February 2013.

[24] S. Petridis, A. Asghar, and M. Pantic, "Classifying laughter and speech using audio-visual feature prediction," in *Proc. ICASSP'10*, Dallas, USA, March 2010, pp. 5254–5257.

[25] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] J. Burgoon et al., *Interpersonal adaptation: Dyadic interaction patterns*, Cambridge University Press, 2007.

[27] Tom Schaul, Justin Bayer, Daan Wierstra, Yi Sun, Martin Felder, Frank Sehnke, Thomas Rückstieß, and Jürgen Schmidhuber, "PyBrain," *Journal of Machine Learning Research*, 2010.