Multi-Modal Neural Conditional Ordinal Random Fields for Agreement Level Estimation

Nemanja Rakicevic, Ognjen Rudovic and Stavros Petridis
Department of Computing
Imperial College London
{n.rakicevic, o.rudovic, stavros.petridis04}@imperial.ac.uk

Maja Pantic
Department of Computing
Imperial College London,
EEMCS University of Twente
m.pantic@imperial.ac.uk

Abstract—**The ability to automatically detect the extent of agreement or disagreement a person expresses is an important indicator of inter-personal relations and emotion expression. Most of existing methods for automated analysis of human agreement from audio-visual data perform agreement detection using either audio or visual modality of human interactions. However, this is suboptimal as expression of different agreement levels is composed of various facial and vocal cues specific to the target level. To this end, we propose the first approach for multi-modal estimation of agreement intensity levels. Specifically, our model leverages the feature representation power of Multi-modal Neural Networks (NN) and discriminative power of Conditional Ordinal Random Fields (CORF) to achieve dynamic classification of agreement levels from videos. We show on the MAHNOB-Mimicry database of dyadic human interactions that the proposed approach outperforms its uni-modal and linear counterparts, and related models that can be applied to the target task.**

I. Introduction

The characteristics that form one's personality are most noticeably exhibited through their interaction with other individuals [1]. Behavioural indicators such as the intensity of exhibited emotions, their frequency, duration, etc., can be useful for personality assessment during interviews, in social interaction studies, and also to measure the compatibility between individuals. It can also facilitate a more natural use of the computer agents in the human environment, among others. To this end, the first necessary step is to be able to automatically measure the human affective states such as agreement, confusion, liking, and so on.

In this paper, we focus on the expressions of agreement between subjects during dyadic conversations. Even though the process of distinguishing someone's (dis)agreement may seem easy to humans, it is rather challenging for a computer system. This is mainly due to the fact that (dis)agreement, like emotions, is a complex affective state expressed by verbal and non-verbal behaviour, also influenced by context (the person's age, gender, culture, etc.) [2]. Moreover, for computers (and also humans) to better apprehend the interlocutor's intentions or social attitudes, it is often necessary to determine not only the presence/absence of agreement, but also its intensity defined on a fine-grained scale. To this end, our goal is to detect specific agreement levels during dyadic conversations using audio-visual modalities (speech and facial expressions) as input, due to their propensity to convey complimentary information [3]. The agreement intensity levels within target videos follow a rising monotonic trend - going from neutral to higher (lower) levels, and back to neutral. Thus, considering the time dependence in agreement data can be of great importance for the level estimation performance. Another important aspect to consider is a non-linear inter-correlation between the input features (e.g., facial landmarks and speech features), especially in spontaneous expressions of agreement. To address this, we propose Multi-modal Neural Conditional Ordinal Random Fields (MM-NCORF), for intensity estimation of the (dis)agreement levels. In MM-NCORF, the non-linear feature extraction and fusion is attained by leveraging the modeling power of NNs [4], and the ordinal nature and temporal structures in the target data are accounted for via CORFs [5]. This is outlined in Fig. 1. The main contributions of this work can be summarized as follows:

- To the best of our knowledge, this is the first approach for multi-modal dynamic estimation of agreement intensity levels from audio-visual (A/V) modalities.
- We introduce a novel modeling framework for ordinal data (in our case, agreement intensity levels) that performs non-linear feature extraction and fusion of multiple modalities in a principled manner. We also introduce a 2-phase joint parameter optimisation approach, leading to efficient learning of the target model.
- We show on MAHNOB-Mimicry database of dyadic interactions that the proposed approach outperforms related unimodal and linear counterparts. It also outperforms the baseline models for dynamic modeling of sequential data.

The rest of the paper is organized as follows: Sec. II reviews the related work. In Sec. III, we describe the proposed approach, and in Sec.VI we show the experimental results. Sec. VII concludes the paper.

II. Related Work

A. Agreement Detection

In order to perform quantitative measurements of human affective states, the most informative behavioural cues need to be identified first. To the best of our knowledge, and as noted in [6], there is no formal definition and annotation procedure for (dis)agreement intensity levels. A challenge here is that there are many ways to define agreement levels: based on audio or

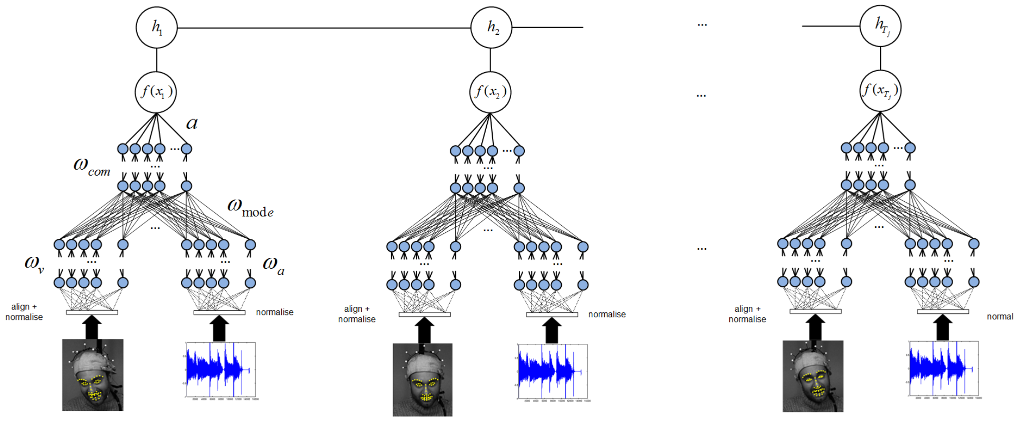

Fig. 1: A sample sequence, depicting tracked facial points in the corresponding images, and audio signals. The inputs are passed through non-linear NN feature extractors, which also perform fusion of the two modalities. The output of the NN is then passed through ordinal functions $f(\cdot)$ that map it onto an ordinal line, classifying the target signal into different intensity levels $h_t = 1, ..., L$ of agreement. Temporal dependencies between target levels (h_t and h_{t+1}.) are also modeled to smooth out the predicted intensity

visual (non-verbal) data, or their combination. To this end, [1] investigates different ways of encoding (dis)agreement: direct (using specific words), indirect (not explicit, but through congruent or contradictory statements) and non-verbal (using auditory or visual non-verbal cues). This, however, may cause ambiguities during the annotation process depending on the modality observed, as information from separate modalities can sometimes be even contradictory. Other lines of work deal with the analysis of (dis)agreement expression channels and verbal/non-verbal cues [6], [1], estimation based on lexical and text-based data [7], as well as audio and prosody cues [8]. Due to the variety of means by which the (dis)agreement can be communicated, we adopt the multi-level annotation scale introduced in [9]. The agreement levels are represented using Likert scale [10], where the intensity of agreement ranges from strong disagreement to strong agreement. In particular, the (dis)agreement levels are defined as: neutral {0}, (dis)agreement {-1,+1}, strong (dis)agreement {-2,+2} as defined in [9].

B. *Multimodal Learning*

For analysing human behaviour, human interactive modalities (i.e. audio, visual and tactile) and both verbal and non-verbal cues (speech, gestures, expressions, etc.) should be considered. However, not all of these are equally informative and/or can be measured reliably. For instance, [11] argue that the dominant channel used for conveying and inferring emotional states is the human face and facial expressions in particular. Many studies have integrated various modalities and reported the advantage of this approach in human emotion recognition, over using only single modalities [12]. This motivates our use of two distinct modalities - visual (facial landmarks) and audio.

A plethora of approaches for fusion of different input modalities exist in the literature (e.g. see [13], [14]). The simplest approach is the concatenation of input feature vectors, typically called early (feature-level) fusion [15]. The bottleneck of early fusion is that it increases the dimensionality of the input, making the model prone to overfitting in case of high-dimensional features. On the other hand, late (decision-level) fusion models each input stream independently and their predictions are then integrated on a higher level [16]. However, this approach fails to account for dependencies between the modalities. Several recent approaches employ graphical models [17] or deep learning [18] to perform multimodal learning, however they do not account for ordinal structure in the model output. In this work, we exploit the benefits of early and late fusion approaches by performing feature fusion through the intermediate layers of NNs.

C. *Structure Modelling*

Temporal models have been shown to be very effective in automated analysis of human behaviour [12], especially for discriminating between posed and spontaneous expressions [19]. Several works perform temporal modeling through the expansion of input feature vectors, by stacking the features of neighbouring frames, which are then fed into a static classifier [11]. We adopt the linear-chain graphical models, such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF) [20]. To model the ordinal nature of intensity levels of facial expressions, extensions of CRFs have been introduced - Conditional Ordinal Random Fields (CORF) [21], [5], and Kernel CORF (KCORF) [22]. While standard CORFs rely on linear feature functions, KCORFs are limited by the number of kernels. Furthermore, they deal with a single modality only. MM-NCORF mitigates these limitations by introducing MM-NNs in the feature functions of these dynamic ordinal models. To perform non-linear selection of input features, [23], [24] combine NNs with CRFs. However, these methods fail to account for ordinal information in the target data.

III. Model Description

Let us denote the data set as $D = \{\mathbf{X}, \mathbf{Y}\}^N$, where for each of the N time instances, $\mathbf{X} = [\mathbf{X}_1, ..., \mathbf{X}_i, ...\mathbf{X}_N]$, the input $\mathbf{X}_i$ is comprised of multiple input vectors $\mathbf{X}_i =$

$\{\mathbf{x}_1, ..., \mathbf{x}_m, ..., \mathbf{x}_M\}$ corresponding to each modality, m, and having dimensions $\mathbf{x}_m \in \mathbb{R}^{D_m}$. The dimension of the m^{th} modality input vector is D_m. Furthermore, $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_i, ...\mathbf{y}_N]$ are the (dis)agreement level labels for each time frame, with $\mathbf{y}_i \in \{-2, -1, 0, +1, +2\}$ encoding expressions from strong disagreement, to strong agreement.

We extend the standard CORF model [21], and uni-modal Neural CORF [9] to design the multi-modal NCORF approach for the target task. In all these models, we first define the conditional distribution $P(\mathbf{Y}|\mathbf{X})$ of having a label sequence $\mathbf{Y}$, based on the observation sequence $\mathbf{X}$, as:

$$P(\mathbf{Y}|\mathbf{X}, \theta) = \frac{1}{Z(\mathbf{X}; \theta)} e^{s(\mathbf{X}, \mathbf{Y}; \theta)} \tag{1}$$

where $Z(\mathbf{X}; \theta) = \sum_{\mathbf{Y} \in \mathcal{Y}} \mathbf{e}^{\mathbf{s}(\mathbf{X}, \mathbf{Y}; \theta)}$ is the normalizing partition function ($\mathcal{Y}$ is a set of all possible output configurations), and θ are the parameters of the *score function* (or the negative energy)[1]. In the case of the linear-chain model with *node cliques* ($r \in V$, V is a set of nodes) and *edge cliques* ($e = (r, s) \in E$, E is a set of edges), the score function $s(\mathbf{X}, \mathbf{Y}; \theta)$ can be expressed as:

$$\begin{aligned} s(\mathbf{X}, \mathbf{Y}; \theta) = \sum_{r \in V} \mathbf{v}^\top \mathbf{\Psi}_r^{(V)}(\mathbf{X}, y_r) + \\ \sum_{e=(r,s) \in E} \mathbf{u}^\top \mathbf{\Psi}_e^{(E)}(\mathbf{X}, y_r, y_s) \end{aligned} \tag{2}$$

where $\theta = \{\mathbf{v}, \mathbf{u}\}$ are the parameters of the node features, $\mathbf{\Psi}_r^{(V)}(\mathbf{X}, y_r)$, and edge features, $\mathbf{\Psi}_e^{(E)}(\mathbf{X}, y_r, y_s)$, respectively. The score function in (2) provides high modelling flexibility via the task-specific node and edge features.

A. Node Potential

The node potentials in the CORF model are defined as:

$$\begin{aligned} \mathbf{v}^T \mathbf{\Psi}_r^{(V)}(\mathbf{X}, y_r) \to \sum_{c=1}^{R} I(y_r = c) \cdot \\ \left[\Phi\left(\frac{b_{y_r} - f(\mathbf{X})}{\sigma}\right) - \Phi\left(\frac{b_{y_r - 1} - f(\mathbf{X})}{\sigma}\right) \right] \end{aligned} \tag{3}$$

where $\Phi(\cdot)$ is the cumulative density function (CDF) of the standard normal distribution, $I(\cdot)$ is the indicator function that returns 1 (0) if the argument is true (false), and σ is usually set to 1 for the model identification purpose. The difference between the CDFs in (3) represents the probability of the observed features, in the ordinal regression framework, given by $\mathbf{X}$, belonging to class $y_r = c \in \{1, ..., R\}$ iff $b_{c-1} < f(\mathbf{X}) \leq b_c$, where $b_0 = -\infty \leq \cdots \leq b_R = \infty$ are (strictly increasing) thresholds or cut points.

In the proposed A/V MM-NCORF model, we adopt a non-linear feature transformation learned, as opposed to using a linear projection of the observed features to obtain the node potentials $f(\mathbf{X}) = \beta^T \mathbf{X}$, where β are the projection weights. This is done by means of a non-linear hidden layer in target NN with sigmoid activation functions for each modality. These outputs are then fused through appropriate stream weights of the common layer and a linear output layer. A simple, 1-layer per modality and 1 common layer, can be written as:

$$f(\mathbf{X}) = \omega_{com}^T \left[\sigma\left(\omega_{mode}^T \left[\sigma(\omega_a^T x_a) + \sigma(\omega_v^T x_v)\right]\right)\right]) \tag{4}$$

where σ is the sigmoid function, defined as $\sigma(x) = \frac{1}{1+e^{-x}}$, and ω_{com}, ω_{mode}, ω_a and ω_v are the weights of the common to output layer, modality stream to common and weights for audio and video streams, respectively (see Fig. 1). The bias parameters associated with each of the layers are included in the weight matrices.

B. Edge Potentials

We use the standard CRF/CORF edge potentials, given by the transition model:

$$\mathbf{\Psi}_e^{(E)}(y_r, y_s) = \Big[I(y_r = k \ \wedge \ y_s = l) \Big]_{R \times R} \tag{5}$$

where R is the number of intensity levels. The role of this potentials is to achieve smooth intensity predictions in the model output.

C. Parameter Optimisation

With the node/edge features defined above, the regularized cost function of the MM-NCORF model is given by:

$$\arg\min_{\theta} \sum_{i=1..N} - \ln P(\mathbf{Y}|f(\mathbf{X}), \theta) + \Omega(\theta) \tag{6}$$

where $\theta = \{\omega, b_1, \ldots, b_{R-1}, \mathbf{u}\}$ are the model parameters[2], and $\Omega(\theta) = \rho_1 \|\mathbf{u}\|^2 + \rho_2 \|\omega - \omega_o\|^2$, is the L_2 regularization used to avoid overfitting of the model parameters[3]. To condense the notation $\omega = \{\omega_{com}, \omega_{mode}, \omega_a, \omega_v\}$, and ω_o are their initial values. In order to produce a good starting point for the NN weights [13], during decoupled 2-phase optimisation, we perform multimodal unsupervised pre-training based on Restricted Boltzmann machines (RBM) as described in [18].

1) Joint Optimisation: We propose a joint optimisation procedure for the CORF parameters together with the NN weights. We use Stochastic Gradient Descent (SGD) with a modified backpropagation algorithm, which considers the mode-specific branches. The gradients for each of the updated parameters are obtained w.r.t. the minimised log-likelihood function (Eq. 6). The CORF parameters' gradients are defined as in [21], and for NN are obtained using the derivative chain rule and the back-propagation approach, similarly as in [23]:

$$\frac{\partial \mathcal{L}(\mathbf{w}, x)}{\partial \mathbf{w}} = \frac{\partial \mathcal{L}(\mathbf{w}, x)}{\partial \tilde{X}} \frac{\partial \tilde{X}}{\partial \mathbf{w}} \tag{7}$$

Here, $\tilde{\mathbf{X}}$ is the output of the NN which is taken as the input of the CORF part, and can be considered as a high-level transformation of the inputs. The derivative of the conditional log-likelihood function w.r.t. a certain weight, $\frac{\partial \mathcal{L}(\mathbf{w}, x)}{\partial \mathbf{w}}$, is obtained by propagating the derivative w.r.t. the inputs to

[1] For simplicity, we drop the dependency on θ in notations.

[2] For more information about the CORF parameters, please refer to the original paper [21].

[3] Note that the second term controls how far are the learned weights of the NN from those learned during pre-training stage, as described in [24].

Algorithm 1: MM-NCORF 2-phase optimisation approach

```
Initialisation:
NN weights ← unsupervised pre-training
CORF parameters ← random initialisation, but $a$ from NN weights
// $a$ is optimised in both phases
$\varepsilon$ ← tuned threshold
Optimisation:
repeat
    // 1st phase
    begin CORF optimisation
        Get $\tilde{X}$ // NN output
        Calculate gradients of cost function w.r.t $\tilde{X}$
        Optimise using LBFGS (20 iterations)
        Update CORF parameters, except $a \not\leftarrow a_{CORF}$
        return $a_{CORF}$
    end
    // 2nd phase
    begin NN weight optimisation
        Calculate error function:
          $Error = (a_{CORF}^T \cdot \tilde{X} - a_{NN}^T \cdot f_{NN}(X))^2$
        Backpropagation (30 epochs, batch=sequence)
        Update NN weights, including last layer $a \leftarrow a_{NN}$
        Re-compute $\tilde{X}$ using updated weights at each step.
        return $a_{NN}$
    end
until $|a_{NN} - a_{CORF}| < \varepsilon$ and $\mathcal{L}_i \geq \mathcal{L}_{i-1}$;
```

the CORF part, $\frac{\partial \mathcal{L}(\mathbf{w},x)}{\partial \tilde{X}}$, using the backward pass of the backpropagation procedure. The main difference lies within $\frac{\partial \tilde{X}}{\partial \mathbf{w}}$, where the adjustment to the back-propagation algorithm is applied to accommodate the derivative propagation to an arbitrary number of modality input branches. The optimisation is done in batches, where each sequence represents one batch.

2) 2-phase Optimisation: In order to decouple the CORF and NN parameter optimisation, the sizes of which are imbalanced, we introduce a 2-phase optimisation procedure. This is to cancel the negative effects of the NN parameters outnumbering the CORF ones, and to reduce the computation time. To keep the optimisation joint, the parameter vector ω_{com} is updated in both phases. In the CORF phase this represents the projection weight β. We denote ω_{com} further as $\mathbf{a}_{NN}$ and $\mathbf{a}_{CORF}$, depending on the phase executed. Therefore, we can define the error function for the NN as the difference in the projections using the common parameter obtained from the CORF phase and NN phase, $\left\|\mathbf{a}_{NN}^T\tilde{\mathbf{X}} - \mathbf{a}_{CORF}^T\tilde{\mathbf{X}}\right\|^2$. Thus, the goal is to align the feature mapping within the CORF and NN's models. The overview of the 2-phase optimisation algorithm procedure is presented in Alg. 1. The first iteration computes $\tilde{\mathbf{X}}$ using the randomly initialised weights, and it is used as input features for the CORF model. The CORF parameters, including $\mathbf{a}_{CORF}$ are updated through the LBFGS method during 20 iterations. In the second step, the error is backpropagated to calculate the NN weight gradients and update them. This phase is performed using SGD, specifically, the ADADELTA approach [25] described earlier. After the NN weight updates, the new $\tilde{\mathbf{X}}$ can be obtained and the method returns to the first phase. The stopping criteria is met when the change in the parameters is not greater than a predefined threshold and if the cost function is not improving, or when the maximum number of iterations of each phase has been achieved. The minimisation of the log-likelihood function is shown in Fig. 2. over the 12 model iterations.

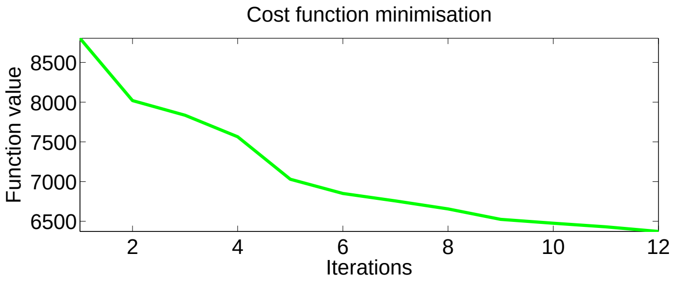

Fig. 2: Log-likelihood function values during the 2-phase optimisation for the best performing NCORF model

IV. Dataset

To train and evaluate the performance of the model we used the MAHNOB-Mimicry database [26] for which we performed the agreement intensity level annotation, as in [9]. The database consists of video recordings of 54 dyadic discussion sessions. We selected videos of 38 subjects, an extension of [9], where authors used only 5 subjects. Labelling was done using segments, which could be defined as a generalisation of 'spurts' - periods of speech by one speaker that have no pauses greater than 0.5 second (similarly as defined in [27]) - where both audio and visual modality were considered. However, intensity labelling was done per frame. Moreover, as in [28], we took semantics into account, not just the literal meaning of the phrase. This means that a sarcastic episode of agreement was labelled as disagreement, regardless of the affirmative wording.

V. Feature Extraction and Synchronisation

Video features. We used the location of 49 facial points (Fig. 1), obtained using the facial point tracker [29]. The facial points have been aligned to the average face from the dataset using an affine transform.
Audio features. We used the OpenSMILE software [30] to extract audio features. The features extracted per frame are the 65 used in [31] (including MFCC, zero-cross rate, jitter, etc.) together with their derivatives, resulting in 130-D feature vectors. A voice activation mask has been applied to each subject's data, which assigns 0 to all instances during which the subject in focus is silent. This avoids the confusion due to the background signal from the other person.

In order to synchronise the data, due to different sampling rates, the audio data were down-sampled to match the video data's frame rate of 59 fps. All features have been z-normalised per-subject to mitigate speaker variation. Still, most parts of the target sessions contained mainly neutral level of agreement, because the subject recorded is either listening to his collocutor making a statement, or is making a neutral statement himself. For this reason, each session was pre-segmented into a number of small sequences which contain at least one non-neutral level. Finally, 909 sequences containing (dis)agreement expressions have been produced. The final annotated agreement level distribution is depicted in Fig. 3.

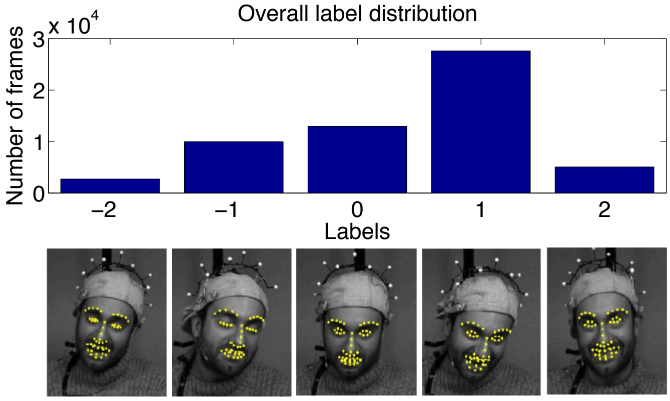

Fig. 3: The agreement level distribution with corresponding example images and tracked facial points.

Note that the levels are highly imbalanced, since there are fewer strong (dis)agreement cases, which can be explained as a ‘conversational preference’ [2] (i.e., when people are more inclined to agree than disagree when talking to strangers).

VI. Experimental Results Analysis

The goal of our evaluation procedure is to investigate: (i) to what extent the proposed fusion scheme outperforms its unimodal counterparts, and (ii) the benefit of non-linear feature extraction and fusion, as achieved by the two proposed learning approaches in MM-NCORF. To this end, we compare MM-NCORF to the baseline models - artificial NN and Support Vector Machines (SVM) - and the the related uni-modal methods - CRF, CORF[4] and KCORF are used for comparison. The cross-validation has been performed for each of the methods in order to find the optimal hyperparameters producing the best results which are presented. For instance, for KCORF using 50 RBF kernels resulted in the best performance. Similarly, for SVM, Linear and RBF kernels were used. In the case of CRF, CORF and also KCORF, L2 regularisation parameters were cross-validated. Furthermore, a comparison between unimodal and multimodal cases is given, where in the former, the model is trained and tested on either audio or video (landmarks) features. For the evaluation measures, we employ the F1 score, as it is insensitive to class imbalance, and the Mean Absolute Error (MAE), as an ordinal measure [21]. In all our experiments, we adopt 5 fold subject independent cross validation (three folds for training, one for parameter validation, and one for testing).

In the unimodal case, the audio modality produces better results with all the models, as can be seen from Table I. The performance gap between the static and dynamic models is noticeable, which is caused by the temporal structure in the data that the static models fail to consider. Another interesting observation can be made by examining the CRF model, which gives inferior results compared to its ordinal counterpart. This shows the importance of modeling the ordinal structure in the data Furthermore, we compare the performance for the NCORF model with two optimisation approaches, joint (NCORF [JNT]) and 2-phase (NCORF [2PH]). The best NCORF model outperforms the other compared models on both measures, with the difference from the second best performing approach with F1 score of 5.7% and 6%, for the audio and visual features, respectively. Here, the performance difference between the two optimisation approaches is less significant, but the 2-phase approach gives better results when more hidden layers are used. The values in the brackets indicate the architectures of the models. Fig. 4 depicts a comparison between several architectures of NCORF [JNT] (left) and NCORF [2PH] (right), and their corresponding F1 scores. The best performing architecture for the video modality is a single hidden node, non-linear layer, while for the audio modality are 2-layer consisting of 10 hidden nodes. By looking into the the performance of the models when using both input streams, from Table II we see the varying performance. Firstly, the difference between static and dynamic models, and also CRFs is pronounced here as well. Concatenating the features (early fusion) leads to worse results in some models, because

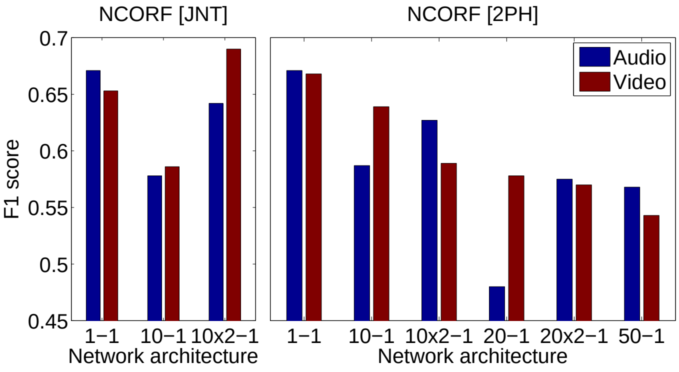

Fig. 4: F1 score for various network architectures of the NCORF, for both audio and visual inputs.

TABLE I: Performance comparison for different methods using unimodal data applied to (dis)agreement intensity level estimation.

	Methods	F1	MAE
Audio	NN (100-100-100-100-1)	0.294	0.777
	SVM (lin)	0.273	0.713
	CRF [20]	0.267	0.717
	CORF [21]	0.629	0.574
	KCORF [22](50 bases)	0.633	0.567
	NCORF [JNT] (10-10-1)	**0.690**	**0.494**
	NCORF [2PH] (1-1-1)	0.671	0.515
Video	NN (100-100-100-1)	0.266	0.774
	SVM (lin)	0.223	0.871
	CRF [20]	0.229	0.828
	CORF [21]	0.504	0.713
	KCORF [22](50 bases)	0.607	0.576
	NCORF [JNT] (1-1)	0.649	0.532
	NCORF [2PH] (1-1-1)	**0.668**	**0.516**

[4] using http://ibug.doc.ic.ac.uk/resources/DOC-Toolbox/

increasing the input feature dimensionality makes the optimisation more difficult and deteriorates the models' performance. MM-NCORF shows that the fusion of the modalities results in significantly better estimates than the early fusion, and it outperforms the KCORF model's F1 score by 7%. The latter evidences that the proposed NN feature extraction is more effective, for the target task, than that achieved by the kernel approach. Moreover, we conclude that the decoupling of the parameters leads to more robust optimisation. When compared to unimodal models, we obtain a 1.3% increase in the F1 score, than each of the single modalities separately. Furthermore, note that the best performing 2 phase MM-NCORF performs 3% better in F1 than its unimodal counterparts, while we obtain a 1.3% increase in the F1 score when compared to the best performing model with joint (uni-modal) optimisation. This indicates that the non-linear modelling of the feature mappings plays the major role in this application. However, note that the MM-NCORF [2PH] uses far fewer parameters, and its learning is faster due to the decoupling of the NN and CORF parameters.

TABLE II: PERFORMANCE COMPARISON OF DIFFERENT MODELS USING MULTIMODAL DATA APPLIED TO (DIS)AGREEMENT INTENSITY LEVEL ESTIMATION.

Methods	F1	MAE
NN (100-100-100-100-1)	0.246	0.788
SVM (lin)	0.216	0.876
SVM (rbf)	0.250	0.740
CRF [20]	0.227	0.831
CORF [21]	0.507	0.805
KCORF [22](50 bases)	0.603	0.595
MM-NCORF [JNT]	0.677	0.505
MM-NCORF [2PH]	**0.703**	**0.487**

VII. Conclusions

In this paper, we introduced a multimodal dynamic method for automatic agreement intensity estimation, from A/V data of naturalistic human-human interactions. From the conducted experiments, we conclude that by using the NN feature extraction within our approach, we outperform the compared kernel-based approaches. Furthermore, both audio and visual modalities, when used separately, exhibit comparable performance in the target task. The fusion of these using the proposed approach is more pronounced when two step learning is employed. While small improvements are achieved over the audio modality only, the benefits of the proposed multimodal approach are that it requires far fewer parameters and it can be learned faster. We expect that by extending our method using facial texture features and Convolutional Neural Networks (CNN) will result in higher contribution from the visual modality, and also overall better feature fusion results.

Acknowledgment

This work has been funded by the European Community Horizon 2020 under grant agreement no. 645094 (SEWA). The work of Stavros Petridis is also funded in part by the European Community FP7 under grant agreement no. 611153 (TERESA).

References

[1] I. Poggi, *Mind, hands, face and body: a goal and belief view of multimodal communication*. Weidler, 2007.

[2] F. Johnson, "Agreement and disagreement: A cross-cultural comparison," *BISAL*, 2006.

[3] R. W. Picard, *Affective computing*. MIT press Cambridge, 1997.

[4] D. Rummelhart, "Learning representations by back-propagation errors," *Nature*, pp. 533–536, 1986.

[5] O. Rudovic, V. Pavlovic, and M. Pantic, "Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation," in *CVPR*, 2012, pp. 2634–2641.

[6] K. Bousmalis *et al.*, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *ACII*, 2009.

[7] K. Allen *et al.*, "Detecting disagreement in conversations using pseudo-monologic rhetorical structure," *EMNLP*, 2014.

[8] W. Wang, S. Yaman, K. Precoda, C. Richey, and G. Raymond, "Detection of agreement and disagreement in broadcast conversations," in *ACL*, 2011.

[9] N. Rakicevic, O. Rudovic, S. Petridis, and M. Pantic, "Neural conditional ordinal random fields for agreement level estimation," in *WASA*, 2015.

[10] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[11] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[12] J. A. Russell, J.-A. Bachorowski, and J.-M. Fernández-Dols, "Facial and vocal expressions of emotion," *Annual review of psychology*, 2003.

[13] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey," *Proceedings of the IEEE*, 2010.

[14] P. K. Atrey, M. A. Hossain, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, 2010.

[15] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *ACM Multimedia*.

[16] S. E. Kahou, C. Pal, and R. Chandias, "Combining modality specific deep neural networks for emotion recognition in video."

[17] C. Cao, Y. Zhang, and H. Lu, "Multi-modal learning for gesture recognition," in *ICMCS*, 2015.

[18] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *NIPS*.

[19] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," *International Journal of Wavelets, Multiresolution and Information Processing*, 2004.

[20] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[21] M. Kim and V. Pavlovic, "Structured output ordinal regression for dynamic facial emotion intensity prediction," in *ECCV*, 2010.

[22] O. Rudovic *et al.*, "Kernel conditional ordinal random fields for temporal segmentation of facial action units," in *ECCV*, 2012.

[23] T. Do, T. Arti *et al.*, "Neural conditional random fields," in *Int'l Conf. on Artificial Intelligence and Statistics*.

[24] A. Vinel *et al.*, "Joint optimization of hidden conditional random fields and non linear feature extraction," in *ICDAR*, 2011.

[25] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[26] X. Sun, J. Lichtenauer, M. F. Valstar, A. Nijholt, and M. Pantic, "A multimodal database for mimicry analysis," in *ACII*, 2011.

[27] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation." in *INTERSPEECH*, 2001.

[28] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proc. of the Conf on Human Language Technology*.

[29] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *CVPR*, 2014.

[30] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *ACM Multimedia*.

[31] B. Schuller, S. Steidl, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," *Interspeech*, 2013.