Neural Conditional Ordinal Random Fields for Agreement Level Estimation

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

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

***Abstract*—We present a novel approach to automated estimation of agreement intensity levels from facial images. To this end, we employ the MAHNOB Mimicry database of subjects recorded during dyadic interactions, where the facial images are annotated in terms of agreement intensity levels using the Likert scale (strong disagreement, disagreement, neutral, agreement and strong agreement). Dynamic modelling of the agreement levels is accomplished by means of a Conditional Ordinal Random Field model. Specifically, we propose a novel Neural Conditional Ordinal Random Field model that performs non-linear feature extraction from face images using the notion of Neural Networks, while also modelling temporal and ordinal relationships between the agreement levels. We show in our experiments that the proposed approach outperforms existing methods for modelling of sequential data. The preliminary results obtained on five subjects demonstrate that the intensity of agreement can successfully be estimated from facial images (39% F1 score) using the proposed method.**

***Keywords*—agreement analysis; neural networks; conditional ordinal random fields.**

I. Introduction

The amount and intensity of (dis)agreement one can express, as well as the frequency of its occurrence during interaction/discussion with others, can serve as a useful indicator of one's personality [1]. Based on this, an estimate of the relation, level of compatibility and cooperation between subjects can be determined. Machine analysis of agreement can thus provide a more objective approach (compared to humans) for personality assessment during, for instance, employment process and social interaction studies. Another important application of (dis)agreement estimation is in Human Computer Interaction (HCI) systems, endowing them with ability to automatically 'sense' users. However, existing approaches to analysis of agreement focus on its detection, instead of intensity estimation. Nevertheless, automated measurement of agreement on a fine-grained scale (i.e., its intensity levels) would allow HCI to better adapt its responses to target users.

Recognizing inter-personal concordance, i.e. agreement or disagreement, is a sequential and time dependent process. Moreover, the expression intensities follow the increasing monotonicity rule such that in order to pass from a negative intensity (disagreement) to a positive one (agreement), it must first go through a neutral state (neither agree nor disagree). Hence, it is important to take into account the time dependence

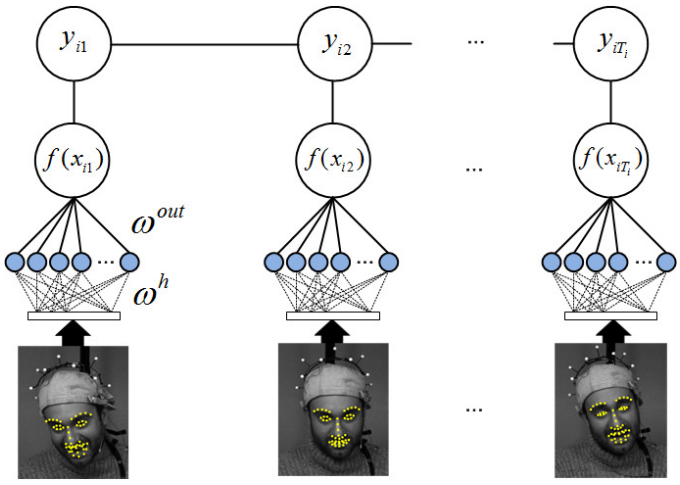

Fig. 1: Neural CORF model for intensity estimation of agreement intensity levels. The input to the model are the facial points extracted from the i-th input image sequence. These features are then projected via $f(\cdot)$ onto an ordinal line separating different levels, the temporal dynamics of which is modelled at the top level using the first-order transition model.

of each frame in the sequence. However, each of the facial features used in the recognition process is correlated with the others, therefore their independence should not be assumed. This inter-correlation and the relationship between the features extracted from facial images, and the labels is usually highly non-linear, especially when working with spontaneously displayed facial expressions (as in MAHNOB data) that may entail large head movements. In order to account for these artefacts, we introduce a non-linear transformation of features using the Feed-forward Neural Networks (NN) [2]. Specifically, the non-linear feature extraction is performed jointly with modelling of temporal and ordinal dependencies among the intensity levels of agreement within the framework of Conditional Ordinal Random Fields (CORF) [3]. The proposed Neural CORF (NCORF) method extends the linear CORFs by adding an additional non-linear hidden layer between the input features and the output node, thus, allowing for non-linear feature transformation and selection.

The contributions of this work can be summarized as follows.

- We propose the first approach to automated estimation of agreement intensity levels from images of spontaneously

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

displayed facial expressions.

- We propose a dynamic method for agreement intensity estimation that offers a simple but effective non-linear modelling of facial features for the target task by combining the NNs and CORFs. We show that this model outperforms its linear counterparts, as well as other non-linear models based on the kernel learning.

The outline of the proposed NCORF model is given in Fig.1.

II. Related Work

A. Agreement Detection

In the past, many approaches to quantitatively measuring character and personality traits have been proposed [4]. The main challenge in automated measurement of attitudes, character, and personality traits from faces is to identify the most relevant facial cues and perform their mapping into quantitative values such as the intensity of (dis)agreement facial expression. Therefore, in order to make quantitative measurements, the target intensity levels need to be properly defined. To the best of our knowledge, and as noted in [5], there is no formal definition and annotation procedure for agreement and disagreement intensity levels. Moreover, this type of social attitude can be inferred in multiple ways, from auditory information, visual (non-verbal) and a combination of these two. However, manual annotation of (dis)agreement in facial images is a tedious, time-consuming process. Furthermore, humans' reasoning about the agreement is the result of the person-specific cognitive process, the character of which may not be very detrimental (by typically developed people) as it involves subjective interpretation of the semantics of the discussion, subject's relationship, personality types, cultural and group climate affects the topic discussed [6]. Therefore, while automated intensity estimation of agreement could speed up this process significantly, it can also provide more consistent target annotations.

Based on conclusions in [7], there are several ways in which (dis)agreement can be expressed by subjects - direct (using specific words); indirect (not explicit, but through congruent or contradictory statements) and non-verbal (using auditory or visual non-verbal cues). This causes a very important issue with annotating (dis)agreement data - the inconsistencies between the modes in which (dis)agreement is conveyed. The labels obtained using either semantics obtained from the discussion (meaning) or non-verbal cues individually, could be discrepant to a large extent. A more comprehensive analysis of (dis)agreement expression modes and cues is presented in [5], including related work on (dis)agreement estimation on lexical [8] and text-based [9] data, auditory and prosodic data [10] and based on non-verbal cues [7].

B. Annotation Scale

Various rating scales have been developed to measure attitudes/agreement. The most accepted scale, Likert scale [11], relies on the principle of measuring attitudes by asking people to respond to a series of statements about a topic, in a way that they specify their level of agreement or disagreement on a symmetric agree-disagree scale, and thus tapping into the cognitive and affective components of attitudes. Likert scale assumes that the strength/intensity of agreement can be represented on a continuum from strongly agree to strongly disagree.

Likert scale defines intensity levels as follows: in the case of 5 levels, the first and last level (-2 and +2, or 1 and 5) should be assigned to the extremes of the attitude range (strongly disagree and strongly agree, respectively), and the middle level (0 or 3) should be assigned to the undecided position. The remaining two levels (-1 and +1, or 2 and 4) are then assigned to the intermediate positions between the extreme and neutral ones. Based on these guidelines, we apply a 5 level Likert scale to our case by defining the (dis)agreement levels as:

Neutral level $\{0\}$: Comprises of the frames where the subject is either making a new statement, listening to the collocutor without expressing its own opinion or contemplating about the topic, again, without expressing any distinguishable opinions, not verbally nor non-verbally.

(Dis)agreement level $\{-1,+1\}$: Corresponds to situations in which the subject has understood the collocutors statement, about which he has no previous opinion or has a weakly opposing one, but is willing to consider it as a valid point and maybe even change his previous view on this point. In case of disagreement, it would result in a counter argument, while for agreement an affirmative non-verbal cues.

Strong (dis)agreement level $\{-2,+2\}$: Occurs in circumstances in which the subject hears a statement that is completely concurring with his own point of view on that topic (strong agreement) and the subject expresses this. Strong disagreement occurs when the subject hears an opinion diametrically different to his own, and there is no willingness to consider it, neither partly agree with it.

In our case, the sessions were annotated by an expert annotator, using the scale definitions defined above. Fig. 2 shows the distribution of the levels in the data used.

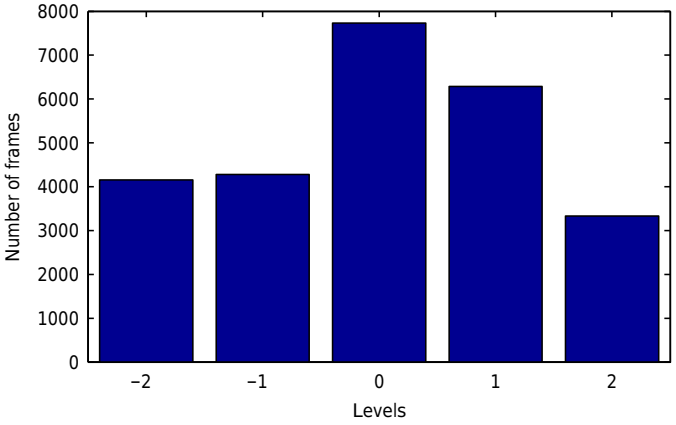

Fig. 2: Distribution of the agreement intensity levels.

C. Modelling Approaches

Existing methods applicable to the target task can be divided into static and dynamic. Furthermore, they can be divided into classification-based and regression-based methods. The former use classifiers for nominal data, such as the static methods based on the Support Vector Machine (SVM), to classify the intensity levels of facial action units (FAUs) [12]. The regression-based methods model the intensity of FAUs/pain on a continuous scale using static versions of the Relevance Vector Machine (RVM) regression [13], and Support Vector Regression (SVR) [14]. For instance, Kaltwang et al. [13] used the RVM model for intensity estimation of spontaneously displayed facial expressions of pain and 11 FAUs from the Shoulder-pain dataset [15]. The effectiveness of different image features such as Local Binary Patterns (LBPs), Discrete Cosine Transform (DCT) and facial landmarks, as well as their fusion, was evaluated for the target task. While these methods perform feature learning/selection using the kernel machines, other static methods are based on Neural Networks [2]. The advantage of using NNs is that they can perform non-linear feature learning from large amount of training data (in contrast to the kernel methods where a limited number of kernels has to be selected). Yet, learning of the structure of NNs is not trivial.

The temporal models aim at capturing the dynamics of the intensity levels. For instance, [16] expands the image features by adding features of the neighbouring frames, which are then fed into a static classifier such as SVM. On the other hand, to avoid modelling high-dimensional feature vectors, graphical models such as Hidden Markov Models and Conditional Random Fields [17] have been proposed to model first order dependences between image frames. To account for increasing monotonicity constraints in intensity level data, different variants of Conditional Ordinal Random Fields (CORFs) [18] and their kernel extensions [19] have been proposed for facial expression intensity estimation and its temporal segmentation.

Due to the learning and inference complexity when working with kernel extensions of CRF-based models [19], [20], several authors combined NNs with CRFs to perform non-linear feature extraction. For instance, the Conditional Neural Fields [21] implement a logistic gate function node hidden layer, which extracts the non-linear representation of the features opposed to their linear combination as in standard Conditional Random Fields. Due to a relatively small number of parameters, the optimization could be done jointly with Conditional Random Fields (CRFs). Another implementation of CRFs with NNs, a structured regression model using continuous outputs, called Continuous Conditional Random Fields (CCNFs), has been proposed in [22]. However, these methods fail to account for ordinal information inherent to the intensity levels.

III. Method

We introduce an approach to sequence prediction that combines the artificial neural networks' non-linear feature representation abilities with the CORF model, which performs ordinal classification of temporal data. To this end, we assume we are given n image sequences $\mathcal{D} = \{(\mathbf{x}^l, \mathbf{y}^l)\}_{l=1}^n$, where $\mathbf{x}$ denotes the the location of a set of facial points extracted from facial images, and used for predicting the intensity levels of agreement as encoded per frame by the labels $\mathbf{y}$.

A. Neural Conditional Ordinal Random Fields (NCORF)

In this section, we extend the linear CORF [3], [18] model for dynamic, non-linear estimation of facial expression intensity levels. The CORF model is an adaptation of the linear-chain CRF [17] model, obtained by setting CRF's node features using the ordinal regression [23] modelling framework. In this way, the monotonicity constraints are imposed on the ordinal labels (in our case, (dis)agreement levels). Formally, given the i-th image sequence, $\mathbf{x}_i = \{x_{i1}, \ldots, x_{iT_i}\}$, and the corresponding intensity labels, $\mathbf{y}_i = \{y_{i1}, \ldots, y_{iT_i}\}$, the conditional distribution $P(\mathbf{y}|\mathbf{x})$ of the CORF model can be written as the Gibbs form clamped on the observations $\mathbf{x}$:

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

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]. By assuming the linear-chain model with *node* cliques ($r \in V$) and *edge* cliques ($e = (r, s) \in E$), the score function $s(\mathbf{x}, \mathbf{y}; \theta)$ can be expressed as the sum:

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

where $\theta = \{\mathbf{v}, \mathbf{u}\}$ are parameters of 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) has a great modeling flexibility, allowing the node and edge features to be chosen depending on target task.

1) **Node features:** In the CORF model [3], the node features are defined using the homoscedastic ordinal regression model [23] (i.e., with the constant variance σ) as:

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

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. In ordinal regression, the difference between the CDFs in (3) is the probability of the observed features, given by x_r, belonging to class $y_r = c \in \{1, ..., R\}$ iff $b_{c-1} < f(x_r) \leq b_c$, where $b_0 = -\infty \leq \cdots \leq b_R = \infty$ are (strictly increasing) thresholds or cut points.

In the standard CORF model [3], $f(x_r) = \beta x_r$, where β is the (linear) ordinal projection. In the proposed NCORF model,

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

instead of using a linear projection of the observed features x_r, we adopt a non-linear feature transformation learned by means of a non-linear hidden layer with sigmoid activation functions and a linear output layer, which is given by:

$$f(x_r) = \omega_{out}^T(\sigma(\sum_{h=1}^{H} \omega_h^T x_r + bias_{in})) + bias_{out}, \quad (4)$$

where σ is the sigmoid function, defined as $\sigma(x) = \frac{1}{1+e^{-x}}$, and ω_h and ω_{out} are the weights of the hidden and output layer, respectively. The bias parameters $bias_{in}$ and $bias_{out}$, are associated with the input and hidden layer, respectively. The number of nodes in the hidden layer is given by H.

2) **Edge features:** The edge features are defined using the transition model as in the standard CRF:

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

enforcing the smoothness of the predicted intensity levels of agreement along the target image sequence.

3) **Learning and Inference:** Using the node and edge features defined above, we arrive at the regularized objective function of the NCORF model:

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

where $\theta = \{\omega_{out}, \omega_{in}, b_1, \ldots, b_{R-1}, \mathbf{u}\}$ are the model parameters, and $\Omega(\theta) = \rho_1\|\mathbf{u}\|^2 + \rho_2(\|\omega_{out}\|^2 + \|\omega_{in}\|^2)$, is the L_2 regularization used to avoid overfitting of the model parameters. The parameters are separated into two sets, the network weights and CORF parameters, and are assigned a different regularization parameter for each group (ρ_1/ρ_2). The weights for each term in the regulizer are found using a cross-validation procedure based on a grid search. To ensure that the threshold parameters b satisfy the ordinal constraints, the displacement variables δ_l are introduced, where $b_l = b_1 + \sum_{n=1}^{l-1}\delta_n^2$ for $l = 2, \ldots, R-1$.

The quasi-Newton limited-memory BFGS method can then be used to find new (unconstrained) parameters θ by jointly optimizing the top layer weights and CORF parameters. The optimization is done using the gradients of the parameters w.r.t. the objective function in (6). The derivation of the CORF gradients can be found in [3]. The gradients of the projection function f w.r.t. the weights of the top layer of the NN projection are:

$$\frac{\partial f}{\partial \omega_{out_{Lk}}} = \sigma_k(\sum_{j=1}^{X} \omega_{h_{kj}} x_j), \qquad \frac{\partial f}{\partial bias_{out}} = 1, \quad (7)$$

Where L stands for the output node, $k \in H$ for the corresponding nodes in the hidden and $j \in X$ in the input layer. For the weights of the hidden layer:

$$\begin{aligned} \frac{\partial f}{\partial \omega_{h_{kj}}} &= \omega_{out_{Lk}}\sigma_k(\omega_{h_{kj}}x_j)(1-\sigma_k(\omega_{h_{kj}}x_j))x_j, \\ \frac{\partial f}{\partial bias_{in_k}} &= \omega_{out_{Lk}}\sigma_k(\omega_{h_{kj}}x_j)(1-\sigma_k(\omega_{h_{kj}}x_j)) \end{aligned} \quad (8)$$

The most critical aspect of the NCORF model is optimization of the NN weights's as their number scales with the input dimension D_x as $(D_x + 1) * H + (H + 1)$. Thus, when D_x and H are both large, a careful regularization of the model parameters is needed to avoid overfitting. Once the model parameters are estimated, inference of test sequences is carried out using Viterbi decoding [3].

IV. Experimental Results

We conducted experiments using facial images from the MAHNOB-Mimicry [24] dataset, which consists of 54 sessions of dyadic discussions, by 40 subjects in total. The subjects discussed either one of the topics such as money, television, women & men, book, smoking, etc. (34 sessions), or participated in the "landlord - student looking to rent" role playing game (20 sessions). Since the former last longer and contain more instances of both agreement and disagreement, they have been considered in these experiment. Furthermore, for this study, we selected 5 subjects, annotated by an expert using the Likert scale (in terms of the agreement level). Although the number of subjects considered is low, the target image sequences were 15 minutes long on average ($\sim$ 55K frames), providing sufficient amount of training/test data for the model evaluation.

The features used are the (x, y) coordinates of 49 tracked facial points (shown in Fig. 5), obtained using the facial point tracker [25]. The facial points have been properly aligned to account for the head position and orientation. The resulting feature vector was of dimension $D_x = 98$. Such features were pre-processed using Principal Component Analysis, preserving $\sim$ 98 % of energy, resulting in $D_x^{pca} = 30$. These features were used as input to the models evaluated. Furthermore, due to the high sampling rate (58 fps), the sequences have been down-sampled by the rate of 4. 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 statement himself. For this reason, each session was pre-segmented into a number of small sequences which contain at least one non-neutral level. These sequences were then used to perform a subject-independent 5-fold cross-validation of the model. Each of the folds contained a similar number of sequences on average.

The proposed approach is compared with other baseline methods, such as static classifiers - artificial NNs with sigmoid function, SVMs (using LIBSVM [26]) (Linear and Radial Basis Function kernel), and also sequence classifiers - CRF and the standard linear CORF, using the Matlab DOC toolbox[2] [27], [28]. Moreover, we compare the performance of the proposed NCORF to the non-linear extension of the CORF model, i.e., kernel CORF (KCORF) [19]. The measures used to show the prediction performance are: F1 score, Mean Absolute Error (MAE), and the commonly used measure in behavioural sciences, Intraclass Correlation Coefficient (ICC) [29], which measures the agreement between the annotations and model predictions.

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

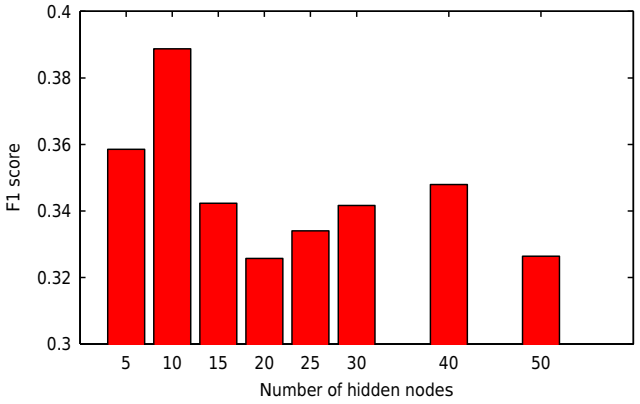

Fig. 3: F1 score for different hidden layer sizes.

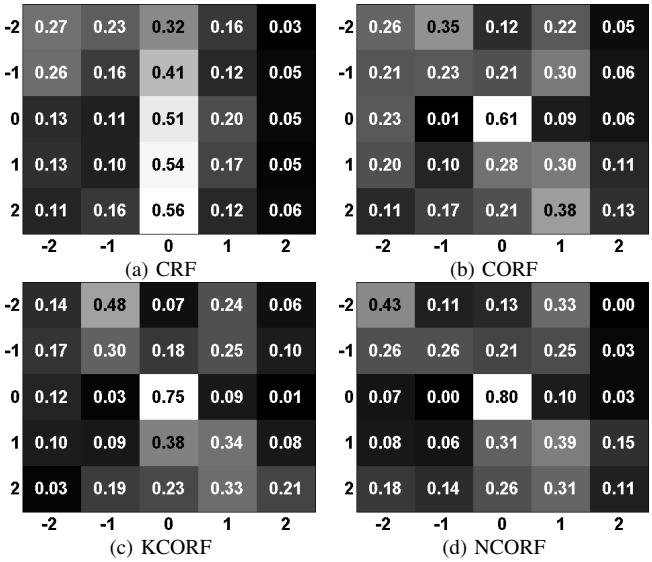

Fig. 4: Confusion matrices of the compared methods. The rows represent the actual, and columns the predicted labels.

We first investigate the performance of the NCORF model w.r.t. to the number of hidden nodes. From Fig. 3, we observe that, as expected, the architecture of the employed NN affects the performance of the model. In other words, the number of hidden nodes (HN) in the hidden layer plays an important role since it performs the non-linear projection of the input features onto the CORFs ordinal classification line. The results suggest that for the target task, having 10 HN is optimal for modelling the non-linear feature projection. On the other hand, by increasing the number of HNs, also the number of NN's weight parameters to optimize significantly increases. This makes the optimization process harder and leads to a worse performance.

The performance of different models is reported in Table I. For the NN-based models, we show the results for the best number of hidden nodes (50 for NN, and 10 for NCORF), found by a cross-validation over the number of nodes. The number of kernel bases in the kernel methods was found by another cross-validation, where, for instance, for KCORF we found 100 bases (we evaluated 50, 100, 200 and 300), to result in the best performance by this model. As can be observed from Table I, the proposed NCORF model outperforms the other models across all three measures. Note that although KCORF achieves comparable ICC as NCORF, the difference in F1 score is 5%. This indicates that both models predict well the trend of the target signal, however, KCORF fails to predict intensity levels per frame at the same accuracy level as NCORF. This can also be noted from Fig. 4, where NCORF predicts the -2 level (strongly disagree), much better than KCORF (43% versus 14%). This is attributed to the fact that while both methods tend to misclassify mainly neighbouring intensity levels, KCORF is more prone to overfitting of the lower intensities. We also note that temporal linear models, CORF and CRF outperform static models (SVM and NNs), with CORF outperforming CRF due to its modelling of the ordinal node features.

TABLE I: PERFORMANCE COMPARISON OF DIFFERENT MODELS APPLIED TO (DIS)AGREEMENT INTENSITY LEVEL ESTIMATION.

Methods	F1	MAE	ICC
NN (50HN)	0.15	0.99	0.07
SVM (rbf)	0.19	1.09	0.15
SVM (lin)	0.20	1.23	0.12
CRF	0.22	1.18	0.14
CORF	0.30	1.15	0.19
KCORF (100 bases)	0.34	0.97	0.26
NCORF (10HN)	**0.39**	**0.94**	**0.28**

Finally, we show in Fig. 5 a sample sequence of the annotated and estimated labels. It can be seen that the proposed model performs well in distinguishing between neutral, agreement and disagreement, and follows the trend of the target labels. However, as we can judge from the models scores, particularly low F1 and ICC, there is still room for improvement in order to achieve precise discrimination between the intensity levels of agreement.

V. CONCLUSION

We propsoed a model for dynamic estimation of agreement intensity levels from image sequences. The proposed model takes advantages of NN's non-linear feature transformation and CORF's dynamical ordinal modelling capabilities. Our preliminary results on data of a small number of subjects show that taking into account the dynamics and, especially the ordinal nature of the data, helps to better distinguish among the agreement levels, compared to other approaches applicable to the target task. Furthermore, the non-linear feature transformation results in the model's ability to better discriminate the more subtle intensity levels (e.g. strong (dis)agreement vs (dis)agreement), when compared to existing models. In future work, we plan to extend the model using the notion of deep neural networks for the feature selection, and perform evaluation using a significantly larger amount of training data. Moreover, we are also planning to investigate the use of the audio modality which may also contain useful information for estimating the level of agreement. This can be further combined with our video-only approach in order to build

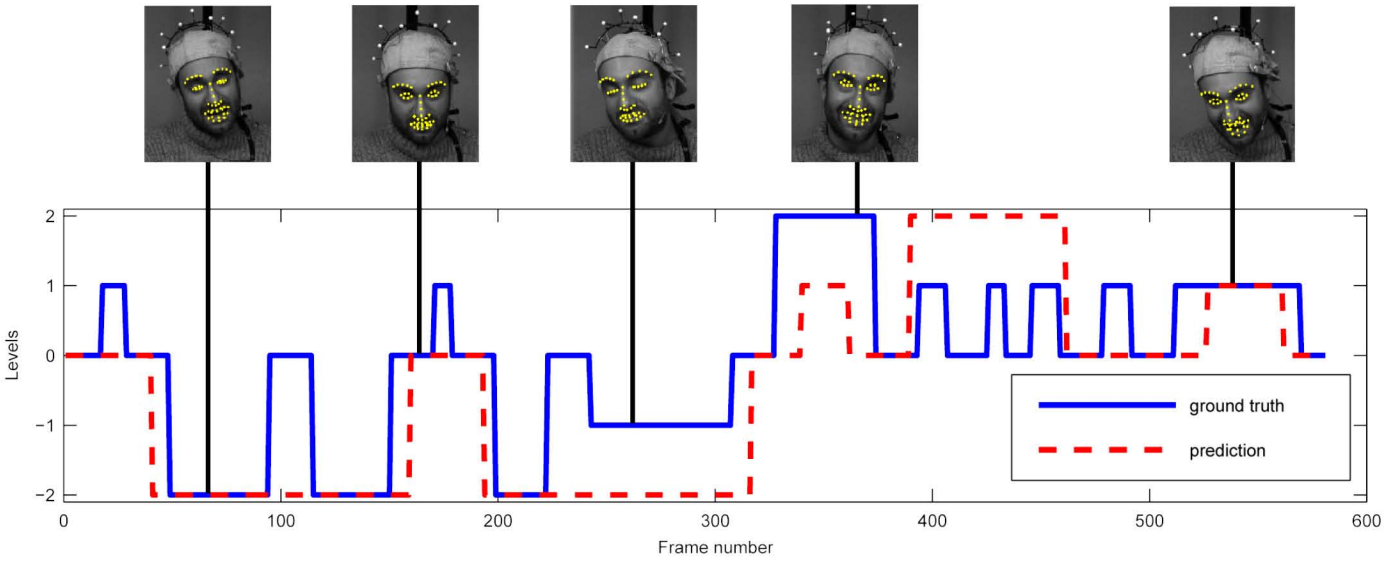

Fig. 5: Examples of facial expressions with tracked facial points connected with corresponding annotated frames from the sample sequence of annotations vs. predictions. The presented predictions are obtained using the NCORF model with 10 hidden nodes.

an audiovisual intensity level estimator, which is expected to further improve the agreement level estimation.

Acknowledgment

This work has been funded by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA). The work of Stavros Petridis is also funded in part by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA).

References

[1] A. S. Gerber, G. A. Huber, D. Doherty, and C. M. Dowling, "Disagreement and the avoidance of political discussion: Aggregate relationships and differences across personality traits," *AJPS*, vol. 56, no. 4, pp. 849–874, 2012.

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

[3] M. Kim and V. Pavlovic, "Structured output ordinal regression for dynamic facial emotion intensity prediction," in *ECCV*. Springer, 2010, pp. 649–662.

[4] W. Bilsky and S. H. Schwartz, "Values and personality," *European journal of personality*, vol. 8, no. 3, pp. 163–181, 1994.

[5] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *ACII*, 2009, pp. 1–9.

[6] F. Johnson, "Agreement and disagreement: A cross-cultural comparison," *BISAL*, vol. 1, pp. 41–67, 2006.

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

[8] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg, "Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies," in *ACL*, 2004, p. 669.

[9] K. Allen, G. Carenini, and R. T. Ng, "Detecting disagreement in conversations using pseudo-monologic rhetorical structure," *EMNLP*, 2014.

[10] W. Wang, S. Yaman, K. Precoda, C. Richey, and G. Raymond, "Detection of agreement and disagreement in broadcast conversations," in *ACL*, 2011, pp. 374–378.

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

[12] M. F. Valstar, I. Patras, and M. Pantic, "Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data," in *CVPRW*, 2005, pp. 76–76.

[13] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," in *Advances in Visual Computing*. Springer, 2012, pp. 368–377.

[14] L. Jeni, J. M. Girard, J. F. Cohn, F. De La Torre *et al.*, "Continuous au intensity estimation using localized, sparse facial feature space," in *FG*, 2013, pp. 1–7.

[15] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *FG*, 2011, pp. 57–64.

[16] J. Mourao-Miranda, K. J. Friston, and M. Brammer, "Dynamic discrimination analysis: a spatial–temporal svm," *NeuroImage*, vol. 36, no. 1, pp. 88–99, 2007.

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

[18] O. Rudovic, V. Pavlovic, and M. Pantic, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *TPAMI*, August 2014.

[19] ——, "Kernel conditional ordinal random fields for temporal segmentation of facial action units," in *ECCV*. Springer, 2012, pp. 260–269.

[20] J. Lafferty, X. Zhu, and Y. Liu, "Kernel conditional random fields: representation and clique selection," in *ICML*, 2004, p. 64.

[21] J. Peng, L. Bo, and J. Xu, "Conditional neural fields," in *NIPS*, 2009, pp. 1419–1427.

[22] T. Baltrušaitis, L.-P. Morency, and P. Robinson, "Continuous conditional neural fields for structured regression," in *ECCV*, 2014.

[23] R. Winkelmann and S. Boes, *Analysis of microdata*. Springer Science & Business Media, 2006.

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

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

[26] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM TIST*, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[27] 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.

[28] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," in *FG*, 2015, pp. 1–8.

[29] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability." *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.