Modeling Hidden Dynamics of Multimodal Cues for Spontaneous Agreement and Disagreement Recognition

Konstantinos Bousmalis Louis–Philippe Morency Maja Pantic

***Abstract*— This paper attempts to recognize spontaneous agreement and disagreement based only on nonverbal multimodal cues. Related work has mainly used verbal and prosodic cues. We demonstrate that it is possible to correctly recognize agreement and disagreement without the use of verbal context (i.e. words, syntax). We propose to explicitly model the complex hidden dynamics of the multimodal cues using a sequential discriminative model, the Hidden Conditional Random Field (HCRF). In this paper, we show that the HCRF model is able to capture what makes each of these social attitudes unique. We present an efficient technique to analyze the concepts learned by the HCRF model and show that these coincide with the findings from social psychology regarding which cues are most prevalent in agreement and disagreement. Our experiments are performed on a spontaneous dataset of real televised debates. The HCRF model outperforms conventional approaches such as Hidden Markov Models and Support Vector Machines.**

I. Introduction

We have recently witnessed significant advances not only in the machine analysis of nonverbal cues, such as head and hand gestures, facial expressions, auditory cues, but also in the field of affect recognition [1]. However, only few works have so far attempted to recognize social attitudes like interest, politeness and flirting [2]. This is partly so because relevant research in social psychology, which would help identify discriminative combinations of multimodal cues, is at best scarce, and because of the fact that there is a gap of relevant annotated data that can be used for such analyses. Despite these difficulties, achieving such a goal is very important if we are to move towards a more naturalistic human–computer–interaction; machines who are able to detect social attitudes and react according to the needs of their user will be more efficient and welcomed for the rather more naturalistic experience they are bound to offer.

Such social attitudes are those of **agreement** and **disagreement**, which are inevitable in daily human–human interactions, from finding a location to dine, to discussions on notoriously controversial topics like politics. Existing work on the automatic recognition of agreement and/or disagreement (see Table I) has mainly used verbal and prosodic cues, e.g. pitch and energy. To the best of our knowledge, no work has managed to successfully recognize spontaneous agreement and disagreement based solely on nonverbal multimodal cues. However, although agreements and disagreements are frequently expressed verbally, the nonverbal behavioral cues that occur during their manifestation play a crucial role in their interpretation. Bousmalis et al. [3] have surveyed and identified such cues that seem to be relevant, as those are evident in social psychology literature (see Tables II and III for a summary). According to this survey, it is the temporal underlying dynamics of multimodal cues that will allow us to recognize agreement and disagreement.

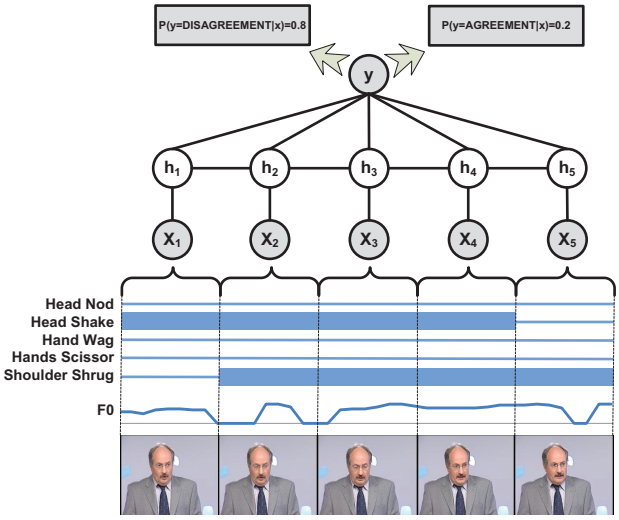

Fig. 1: HCRF for spontaneous agreement/disagreement recognition. h_i represents the hidden state that captures the underlying dynamics between features and labels at a given timestamp i. The HCRF model is able to capture fine–grain hidden multimodal dynamics better than other models by learning these hidden states and their relation to each class. Consequently, the HCRF model is able to learn a more suitable mapping between the observations **x** and each class label $y \in Y$.

This calls for a model capable of capturing these complex dynamics and, based on them, distinguishing these social attitudes from each other. A Hidden Conditional Random Field (HCRF) [4], originally proposed for object recognition, is a model capable of not only capturing the underlying structure of events, but also of learning the combinations of features that are shared by each class and the ones that make each of them unique. Hence, HCRFs could be a good candidate for modeling agreement and disagreement.

This paper will show that *(i)* it is possible to recognize spontaneous agreement and disagreement without the use

K. Bousmalis is with the Department of Computing, Imperial College London, SW7 2AZ, London, UK k.bousmalis@imperial.ac.uk

L.–P. Morency is with the Institute for Creative Technologies, University of Southern California, Los Angeles, CA 90094, USA morency@ict.usc.edu

M. Pantic is with the Department of Computing, Imperial College London, SW7 2AZ, London, UK, and EEMCS, University of Twente, 7522 NB Enschede, The Netherlands m.pantic@imperial.ac.uk

Method	Features	Classifier	Data	Spontaneous
Hillard et al. [5] (2003)	Verbal, pause, fundamental frequency(F0), duration	Decision Tree	ICSI [6]	√
Galley et al [7] (2004)	Verbal	Bayesian Network	ICSI [6]	√
el Kaliouby et al. [8] (2004)	head nod, head shake, head turn, head tilt, AU1, AU2, AU12, AU16, AU19, AU20, AU25, AU26, AU27	HMM, DBN	Mind Reading DVD [9]	—
Hahn et al [10] (2006)	Verbal	Contrast Classifier, SVM	ICSI [6]	√
Sheerman–Chase et al. [11] (2009)	head yaw, head pitch, head roll, AU1, AU2, AU12, AU18, AU20, AU25, Gaze, head pose	AdaBoost	own	√
Germesin and Wilson [12] (2009)	Verbal, pitch, energy, duration, pauses, speech rate	Decision Tree, CRF	AMI [13]	√

TABLE I: Summary of the existing systems that have attempted agreement/disagreement classification.

of verbal cues (e.g. spoken words); *(ii)* HCRFs are indeed able to capture the underlying dynamics of multimodal cues and perform better than conventional models in this task (figure 1); and *(iii)* HCRFs are able to automatically identify groups of features specific to each attitude in a way that confirms the findings in social psychology literature regarding which cues are most prevalent during the expression of agreement and disagreement.

In the following section, we discuss agreement, disagreement and related work on their automatic recognition. In *Section III* we present Hidden Conditional Random Fields (HCRFs) and our technique to analyze the concepts learned by the HCRF. In *Section IV* we explain how our data was collected and what experiments we have conducted. Finally, in *Section V*, we present and discuss our results.

II. Agreement and Disagreement

A. Definitions and Associated Cues

Distinguishing between different kinds of agreement and disagreement is difficult, mainly because of the lack of widely accepted definitions of agreement and disagreement [3]. We can distinguish among at least three ways one could express agreement and disagreement with:

- **Direct Speaker's Agreement and Disagreement:** A speaker directly expresses his/her agreement or disagreement, e.g. "I (dis)agree with you".
- **Indirect Speaker's Agreement and Disagreement:** A speaker does not explicitly state her agreement and disagreement, but expresses an opinion that is congruent (agreement) or contradictory (disagreement) to an opinion that was expressed earlier in the conversation.
- **Nonverbal Listener's Agreement and Disagreement:** A listener nonverbally expresses her agreement or disagreement to an opinion that is currently or was just expressed. This could be via auditory cues like "mm hmm" or visual cues like a head nod or a smile.

It is important to mention at this point that in spontaneous direct and indirect speaker's agreement/disagreement, the speaker also exhibits nonverbal behavior which could perhaps be different than the one exhibited during nonverbal listener's agreement/disagreement.

Tables II and III present a full list of the nonverbal cues that can be displayed during agreement and disagreement [3][1]. The most prevalent and straightforward cues seem to be the **Head Nod** and the **Head Shake** for agreement and disagreement respectively, with nods intuitively conveying affirmation and shakes negation. However, simply the presence of these or any of the other cues alone cannot be discriminative enough, since they could have many other interpretations, as studied by Poggi et al. [14] and Kendon [15].

CUE	KIND
Head Nod	Head Gesture
Listener Smile (AU12, AU13)	Facial Action
Eyebrow Raise (AU1+AU2)+Head Nod	Facial Action, Head Gesture
AU1 + AU2 + Smile (AU12, AU13)	Facial Action
Sideways Leaning	Body Posture
Laughter	Audiovisual Cue
Mimicry	Second–order Cue

TABLE II: Cues of Agreement. For relevant descriptions of AUs, see FACS [16].

B. Related Work on Automatic Recognition

There is no work, to the best of our knowledge, that has attempted agreement/disagreement classification on audiovisual spontaneous data. Table I summarizes the existing systems that have attempted classification of agreement and/or disagreement in one way or another. However, none of these systems is directly comparable with ours.

Hillard et al. [5] attempted speaker agreement and disagreement classification on pre–segmented 'spurts', speech segments by one speaker with pauses not greater than 500ms. The authors used a combination of word–based and prosodic cues to classify each spurt as 'positive–agreement', 'negative–disagreement', 'backchannel', or 'other'. Most of the results reported included word–based cues, however an overall classification accuracy of 62% was reported for a 17% confusion rate between the agreement and disagreement classes. Similar works by Galley et al. [7] and Hahn et al. [10] also deal with classifying spurts as disagreement and agreement, with [7] also dealing with finding the addressee of the action. Germesin and Wilson [12] also deal with these issues. However, the features used by these works included

[1]Our discussion of cues for agreement and disagreement is mostly relevant for cultures in Western Europe and North America. Further work might be needed to develop a similar system that targets other cultures.

CUE	KIND
Head Shake	Head Gesture
Head Roll	Head Gesture
Cut Off	Head Gesture
Clenched Fist	Hand Action
Forefinger Raise	Hand Action
Forefinger Wag	Hand Action
Hand Chop	Hand Action
Hand Cross	Hand Action
Hand Wag	Hand Action
Hands Scissor	Hand Action
Ironic Smile/Smirking [AU12 L/R(+AU14)]	Facial Action
Barely noticeable lip–clenching (AU23, AU24)	Facial Action
Cheek Crease (AU14)	Facial Action
Lowered Eyebrow/Frowning (AU4)	Facial Action
Lip Pucker (AU18)	Facial Action
Slightly Parted Lips (AU25)	Facial Action
Mouth Movement (AU25/AU26)	Facial Action
Nose Flare (AU38)	Facial Action
Nose Twist (AU9 L/R, AU10 L/R, AU11 L/R)	Facial Action
Tongue Show (AU19)	Facial Action
Suddenly Narrowed/Slitted Eyes (fast AU7)	Facial Action
Eye Roll	Facial Action/Gaze
Gaze Aversion	Gaze
Arm Folding	Body Posture
Large Body Shift	Body Action
Leg Clamp	Body Posture
Head/Chin Support on Hand	Body/Head Posture
Neck Clamp	Hand/Head Action
Head Scratch	Head/Hand Action
Self–manipulation	Hand/Facial Action
Feet Pointing Away	Feet Posture
Sighing	Auditory Cue
Throat Clearing	Auditory Cue
Delays	Auditory Cue
Utterance Length	Auditory Cue
Interruption	Auditory Cue

TABLE III: Cues for Disagreement. For relevant descriptions of AUs, see FACS [16]

lexical, structural and durational cues and are not comparable with other systems based on nonverbal cues.

The first such system is that by el Kaliouby and Robinson [8], which attempted agreement/disagreement classification of *acted* behavioural displays based on head and facial movements. They used 6 classes: 'agreeing', 'disagreeing', 'concentrating', 'interested', 'thinking', and 'unsure'. They tracked 25 fiducial facial points, out of which they extrapolated rigid head motion (yaw, pitch, and roll), and facial action units (eyebrow raise, lip pull, lip pucker), but also utilized appearance–based features to summarise mouth actions (mouth stretch, jaw drop, and lips parting). They used Hidden Markov Models (HMMs) to detect each head and facial action, and a Dynamic Bayesian Network (DBN) per class was trained to perform the higher–level inference of each of the 'mental states' mentioned above, allowing for the co–occurrence of states.

Sheerman–Chase et al. [11] are, to our knowledge, the only research group who have attempted recognition of agreement based on non–verbal cues in spontaneous data. However, they did not include disagreement as a class, because of the lack of data. They instead distinguished between 'thinking', 'understanding', 'agreeing' and 'questioning'. Their spontaneous data was obtained by capturing the four 12–minute dyadic conversations of 6 males and 2 females. 21 annotators rated the clips with each clip getting on average around 4 ratings that were combined to obtain the ground truth label. For the automatic recognition, they used no auditory features and the tracking of 46 fiducial facial points was used. The output of the tracker was then processed to obtain a number of static and dynamic features to be used for classification. Principal Component Analysis (PCA) was performed on the tracked points in each video frame, and the PCA eigenvalues were used as features. Similarly to el Kaliouby and Robinson [8], the head yaw, pitch and roll, the eyebrow raise, lip pucker and lip parting were calculated as functions of these tracked facial points. Gaze was also estimated in a similar fashion —the eye pupils were among the points tracked.

III. HCRFs for Multimodal Gesture Recognition

Hidden Conditional Random Fields —discriminative models that contain hidden states— are well–suited to the problem of multimodal cue modeling for agreement/disagreement recognition. Quattoni et al. [4] presented and used HCRFs to capture the spatial dependencies between hidden object parts. Wang et al. [17] used them to capture temporal dependencies across frames and recognize different gesture classes. They did so successfully by learning a state distribution among the different gesture classes in a discriminative manner, allowing them to not only uncover the distinctive configurations that uniquely identifies each class, but also to learn a shared common structure among the classes. Moreover, as a discriminative model, HCRFs require a fewer number of observations than a generative model like a Hidden–Markov Model (HMM). These were all qualities that prompted us to select HCRFs, as a model to experiment with, in our attempt to recognize agreement and disagreement.

A. Model

Following the notation of Quattoni et al. [4], [17], we represent m local observations by a vector $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$. Each local observation x_j is represented by a feature vector $\phi(x_j) \in \Re^d$ which includes all input features (e.g., the presence of a head nod or the value of F0-pitch). We wish to learn a mapping between the observations $\mathbf{x}$ and the class label $y \in Y$. The class label can be 'agreement' or 'disagreement'. An HCRF models the conditional probability of a class label given an observation sequence by:

$$P(y \mid \mathbf{x}, \theta) = \sum_{\mathbf{h}} P(y, \mathbf{h} \mid \mathbf{x}, \theta) = \frac{\sum_{\mathbf{h}} e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}}{\sum_{y' \in Y, \mathbf{h}} e^{\Psi(y', \mathbf{h}, \mathbf{x}; \theta)}}. \quad (1)$$

where $\mathbf{h} = \{h_1, h_2, \ldots, h_m\}$, each $h_i \in H$ captures certain underlying structure of each class and H is the set of hidden states in the model. The potential function $\Psi(y, \mathbf{h}, \mathbf{x}; \theta) \in \Re$ is an energy function, parameterized by θ, which measures the compatibility between a label, a sequence of observations and a configuration of the hidden states.

$$\Psi(y, \mathbf{h}, \mathbf{x}; \theta) = \sum_j \phi(\mathbf{x}, j) \cdot \theta_h(h_j) + \sum_j \theta_y(y, h_j) + \sum_{(j,k) \in E} \theta_e(y, h_j, h_k) \quad (2)$$

The graph E is a chain where each node corresponds to a hidden state variable at time t. The paremeter vector θ is made up of three components: $\theta = [\theta_e \ \theta_y \ \theta_h]$. We use the notation $\theta_h[h_j]$ to refer to the parameters θ_h that correspond to state $h_j \in H$. Similarly, $\theta_y[y, h_j]$ stands for parameters that correspond to class y and state h_j and $\theta_e[y, h_j, h_k]$ measures the compatibility between pairs of consecutive states j and k and the gesture y.

B. Training

Given a new test sequence $\mathbf{x}$, and parameter values θ^* induced from training examples, we will take the label for the sequence to be:

$$\arg\max_{y \in Y} P(y \mid \mathbf{x}, \theta^*). \quad (3)$$

The following objective function is used in training the parameters:

$$L(\theta) = \sum_i \log P(y_i \mid x_i, \theta) - \frac{1}{2\sigma^2} ||\theta||^2 \quad (4)$$

The first term in Eq. 4 is the log-likelihood of the data. The second term is the log of a Gaussian prior with variance σ^2, i.e., $P(\theta) \sim \exp\left(\frac{1}{2\sigma^2} ||\theta||^2\right)$. We use gradient ascent to search for the optimal parameter values, $\theta^* = \arg\max_\theta L(\theta)$, under this criterion. For our experiments we used a Quasi-Newton optimization technique to minimize the negative log–likelihood of the data.

C. Analysis

The HCRF model is a powerful sequential discriminative model. It can learn the hidden dynamic of a signal using the latent variable h_j. For multimodal gesture recognition, this hidden dynamic is usually related to the synchrony and asynchrony between speech and gestures. While previous work has shown the efficiency of HCRF for learning visual gestures [4], [17], none of them described or analysed what the HCRF model learned. In this paper we are presenting an efficient approach to analyze the concepts learned by the HCRF model. This analysis tool enables a new direction of research where machine learning is not simply used as a black box but instead is there to help understand human interactions.

To analyze the HCRF model, one has to understand the optimized parameters $[\theta_h \ \theta_e \ \theta_y]$:

- θ_h models the relationship between observations $\mathbf{x_j}$ and hidden states h_j. If the HCRF model has 10 input features and 3 hidden states, then the θ_h parameter will be of length 30 (10x3). By analysing the amplitude of each weights in θ_h, it is possible to learn the relative importance of each input feature for each hidden state.

(a) Forefinger Raise (b) Forefinger Wag

(c) Hand Wag (d) Hands Scissors

Fig. 2: Some of the gestures used as cues for the experiments.

- The parameter θ_y models the relationship of the hidden states h_j and the label y. If the model contains 3 hidden states and 2 labels, then the θ_y will be of length 6 (3x2). By analyzing the weights of θ_y, it is possible to see which hidden states are shared and which ones are not.
- The parameter θ_e represents the links between hidden states. It is similar to the transition matrix in a Hidden Markov Model. An important difference is that the HCRF model keeps a transition matrix for each label. If the HCRF model contains 3 hidden states and 2 labels, then the θ_e parameter will be of length 18 (3x3x2).

The procedure for analyzing the HCRF model contains three steps: *(1)* identify the relevant features for each hidden state using θ_h, *(2)* determine which hidden states are shared and which ones are not using θ_y, and *(3)* analyze the possible transitions between hidden states using θ_e. In our experiments (see Figure 5), we apply this procedure to identify the relevant concepts learned by the HCRF model to recognize agreement and disgreement behaviors.

IV. Experiments

A. Dataset and Cues

Our dataset originated from the *Canal 9 Database of Political Debates* [18], one that comprises of 43 hours and 10 minutes of 72 real televised debates on *Canal 9*, a local Swiss television station. The debates are moderated by a presenter, and there are two sides that argue around a central issue, with one or more participants on each side. Hence, the database is rather rich in episodes of spontaneous agreement and disagreement.

The dataset we used comprises of 53 episodes of agreement and 94 episodes of disagreement, which occur over a total of 11 debates. These episodes were selected on the

basis of verbal content, and thus, only episodes of direct and indirect agreement/disagreement were included (see Section II-A). As the debates were filmed with multiple cameras, and edited live to one feed, the episodes selected for the dataset were only the ones that were contained within one personal, close–up shot of the speaker.

We automatically extracted nonverbal auditory features used in related work, specifically the fundamental frequency (F0) and energy, by using a freely–available tool, *OpenEar*[19]. Since our main goal is to analyze dynamics of nonverbal cues during agreement/disagreement recognition, our dataset was manuallly annotated to gather as accurate temporal information about the gestures as possible. Based on the results presented in this paper, our future work will evaluate the recognition performance using our automatic nonverbal gesture annotation [20], [21]. The hand and head gestures we included were based off the relevant list of cues from the Social Psychology literature (see Section II-A), with the exception of a number of head and hand gestures that never appeared in the dataset, and the addition of the 'Shoulder Shrug' and the 'Forefinger Raise-Like' gestures. The latter is a 'Forefinger Raise' without an erect index finger. The cues we finally extracted and used in our experiments are listed in Table IV; the visual cues that may not be self–explanatory from their title are depicted in figure 2.

CUE	KIND
Head Nod	Head Gesture
Head Shake	Head Gesture
Forefinger Raise	Hand Action
'Forefinger Raise'–Like	Hand Action
Forefinger Wag	Hand Action
Hand Wag	Hand Action
Hands Scissor	Hand Action
Shoulder Shrug	Body Gesture
Fundamental Frequency (F0)	Auditory Cue
Energy	Auditory Cue

TABLE IV: The list of features we used in our experiments.

B. Methodology

We conducted experiments with Support Vector Machines (SVMs), as our baseline static classifiers, Hidden–Markov Models (HMMs), the most–commonly used dynamic generative model, and Hidden Conditional Random Fields (HCRFs), the dynamic discriminative model we believe is most appropriate for such a task. We conducted different experiments for three groups of cues: only auditory, only visual, and both auditory and visual ones.

Our cues were encoded differently for our static and dynamic classifiers, but the same information was available to all classifiers. For SVMs, the features of each gesture were the start frame and the duration (total number of frames) of the gesture within the segment of interest. For the auditory features we used the mean, standard deviation, and the first, second(median), and third quartiles of each. The later values did not take into account the undefined areas of F0, and all values were scaled from -1 to 1. For the experiments with HMMs and HCRFs, we encoded each gesture in a binary manner (1 if the gesture is activated in a certain frame, 0 otherwise), and used the raw values of our auditory features, normalized per subject. Figure 1 allows the reader to visualize the process of reaching a classification decision from our data by using an HCRF.

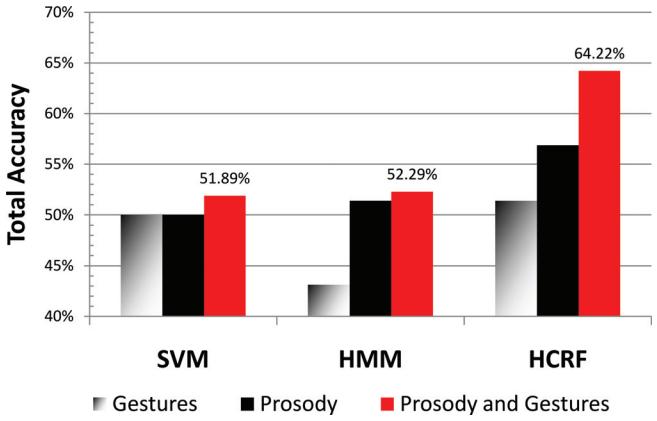

Fig. 3: Comparisons of recognition performance (total accuracy) by the classification methods we explored on the three different groups of features used.

All our experiments were run in a leave–one–debate–out fashion, i.e. the testing set always comprised of examples from the one debate which was not included in the training and validation sets. The optimal model parameters for each test set were chosen by a three–fold validation on the remaining debates. Those were cost and gamma for SVMs, number of mixtures of Gaussians for HMMs, regularization factor for HCRFs and number of hidden states for both HMMs and HCRFs. The HMM and HCRF experiments were run with 10 different random initializations, the best of which was chosen each time during the validation phase (i.e., based on performance on the validation sets). The evaluation metric that we used for all the experiments was the total accuracy in a balanced dataset, i.e. percentage of sequences for which the correct label was predicted in a test set that contains an equal number of agreement and disagreement examples.

V. Results and Discussion

Figure 3 summarizes the results of the experiments on spontaneous agreement and disagreement classification using auditory, gestural and both auditory and gestural features. It is clear that:

(a) It *is* possible to perform the task of spontaneous agreement and disagreement classification without the use of any verbal features.
(b) The temporal dynamics of the cues are vital to the task, as it is evident that SVMs are not able to perform well by using static information alone.
(c) HCRFs outperform SVMs and HMMs, especially when the cues used are multimodal and the underlying dynamics of the different modalities need to be learned.

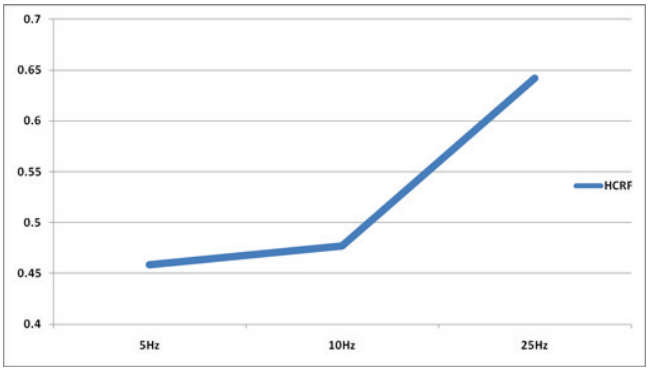

Fig. 4: The performance (total accuracy) of HCRFs increases proportionally to the sampling rate of the multimodal data.

Figure 4 demonstrates the complexity of the task at hand, and the importance of fine–grain multimodal dynamics to its solution, by summarizing the accuracies achieved with HCRF models when sampling our data at different rates. The fact that the higher the sampling rate, the higher the accuracy achieved by the HCRF models, also demonstrates the ability of the HCRFs to cope with such fine–grain dynamics.

We applied our model analysis technique described in Section III-C to the optimal HCRF selected during our experiments. By examination of the weights learned by the HCRF for each of its cues θ_h, hidden states θ_y, and transitions θ_e, we were able to rank, according to importance, the information that the model used. Figure 5 shows the automatically learned concepts of the optimal HCRF model

In Figure 5, each hidden state (represented by the white circles) is linked to its highest ranked observed features in a descending order of importance. The relationships between these observed features and the hidden states were identified using the parameter θ_h. The highest ranked features in these hidden states show that the Head Nod and the Head Shake, which are considered, by social psychologists, the most prevalent cues in agreement and disagreement respectively (see Section II-A), are also the most discriminative cues here. It could be the case that 'Forefinger Raise-Like' gestures might in fact play no role in discriminating between the two attitudes.

By analyzing the parameter θ_y, we can see that the HCRF model assigned one state as prevalent for each of the two class labels, and one state as shared between them. The analysis of the transition parameter θ_e shows that different transitions are learned for each class label. The Figure 5 marked the most likely transitions associated to each attitude (class label): green for agreement and red for disagreement. Disagreement will usually end with hidden state $h = 2$ (middle circle) while the agreement can transition directly to a head shake (hidden state $h = 1$ depicted on the left).

VI. Conclusion and Future Work

Related work on spontaneous agreement/disagreement classification has used verbal (e.g. spoken words) and prosodic features. We have shown, in this work, that the task

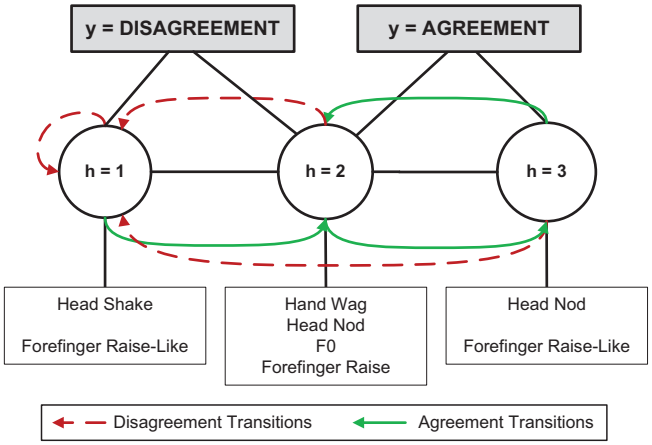

Fig. 5: The features learned for each state by a three-state HCRF model. The green and red connections correspond to the highest–ranked transition from each state in the cases of agreement and disagreement respectively. The middle state is shared among the two classes.

is possible without the use of verbal features. Furthermore, we have shown that HCRFs are a good choice for this task, as they outperform SVMs and HMMs, demonstrating the advantages of joint discriminative learning and their ability to model the hidden fine–grain dynamics of the multimodal cues related to agreement and disagreement. Finally, we have shown that HCRFs can be automatically analysed to identify what groups of features are the most discriminative in each class.

The next step is to evaluate our recognition algorithm using automatically annotated head and hand gestures[20], [21]. Furthermore, a rating study, which is already underway, will exhibit how human raters perform at classifying these clips. Finally, another possible research avenue is the inclusion of other groups of cues associated with agreement/disagreement (see Tables II and III), especially facial actions.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0917321 and the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. This work has been funded in part by the European Community's 7^{th} Framework Programme [FP7/2007-2013] under the grant agreement no 231287 (SSPNet). The work of Maja Pantic is further funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). Finally, the authors would like to thank Dr. Marc Mehu for his help in creating the dataset used in this work.

References

[1] M. Pantic, A. Nijholt, A. Pentland, and T. S. Huang, "Human–Centred Intelligent Human–Computer Interaction (HCI^2): How far are we from

attaining it?" *Journal of Autonomous and Adaptive Communications Systems*, vol. 1, no. 2, pp. 168–187, 2008.

[2] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[3] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *Proc. IEEE Int'l Conf. Affective Computing and Intelligent Interaction*, 2009, pp. 1–9.

[4] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *Proc. Conf. Neural Information Processing Systems*, 2004, pp. 1097–1104.

[5] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: training with unlabeled data," in *Proc. Conf. North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, pp. 34–36.

[6] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, and E. Shriberg, "ICSI meeting corpus," in *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, 2003.

[7] 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 *Proc. Meeting Association for Computational Linguistics*, 2004, pp. 669–676.

[8] R. el Kaliouby and P. Robinson, "Real–time inference of complex mental states from facial expressions and head gestures," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 3, 2004, pp. 154–154.

[9] S. Baron-Cohen, O. Golan, S. Wheelwright, and J. J. Hill, *Mind Reading: The Interactive Guide to Emotions*. London:Jessica Kingsley, 2004.

[10] S. Hahn, R. Ladner, and M. Ostendorf, "Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers," in *Proc. Human Language Technology Conf. of the NAACL*, 2006, pp. 53–56.

[11] T. Sheerman-Chase, E.-J. Ong, and R. Bowden, "Feature selection of facial displays for detection of non verbal communication in natural conversation," in *Proc. IEEE Int'l Workshop on Human–Computer Interaction*, 2009.

[12] S. Germesin and T. Wilson, "Agreement detection in multiparty conversation," in *Proc. Int'l Conf. on Multimodal Interfaces*, 2009, pp. 7–14.

[13] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi–everything AMI Meeting Corpus," *Language Resources and Evaluation Journal*, vol. 41, no. 2, pp. 181–190, 2007.

[14] I. Poggi, F. D'Errico, and L. Vincze, "Types of nods. the polysemy of a social signal," in *Proc. Int'l Conf. Language Resources and Evaluation*, 2010.

[15] A. Kendon, "Some uses of the head shake," *Gesture*, vol. 2, no. 2, pp. 147–182, 2002.

[16] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system," Salt Lake City: Research Nexus, 2002.

[17] S. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1521–1527.

[18] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," in *Proc. IEEE Int'l Conf. Affective Computing and Intelligent Interfaces*, vol. 2, 2009, pp. 96–99.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR — Introducing the Munich open-source emotion and affect recognition toolkit," in *Proc. IEEE Int'l Conf. Affective Computing and Intelligent Interaction*, 2009, pp. 1–6.

[20] L.-P. Morency, J. Whitehill, and J. Movellan, "Generalized adaptive view–based appearance model: Integrated framework for monocular head pose estimation," in *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, 2008.

[21] A. Oikonomopoulos, I. Patras, and M. Pantic, "An implicit spatiotemporal shape model for human activity localisation and recognition," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, vol. 3, 2009, pp. 27–33.