Multimodal Real-Time Contingency Detection for HRI

Vivian Chu, Kalesha Bullard, and Andrea L. Thomaz[1]

***Abstract*— Our goal is to develop robots that naturally engage people in social exchanges. In this paper, we focus on the problem of recognizing that a person is responsive to a robot's request for interaction. Inspired by human cognition, our approach is to treat this as a contingency detection problem. We present a simple discriminative Support Vector Machine (SVM) classifier to compare against previous generative methods introduced in prior work by Lee et al. [1]. We evaluate these methods in two ways. First, by training three separate SVMs with multi-modal sensory input on a set of batch data collected in a controlled setting, where we obtain an average F_1 score of 0.82. Second, in an open-ended experiment setting with seven participants, we show that our model is able to perform contingency detection in real-time and generalize to new people with a best F_1 score of 0.72.**

I. Introduction

The overall goal of our research is to develop social robots that interact with everyday people in human environments. In these situations, a person working with the robot should not be required to learn how to interact with it. We want to enable robots to take advantage of the ways people naturally engage in social exchanges.

Here we address one aspect of any social exchange, recognizing that the robot is in the presence of someone willing to interact. Our approach is inspired by human cognition. Watson has proposed that contingency detection is an important way for infants to recognize social agents [2]. In an experiment with 2-month old infants, when a toy is rigged to respond to the baby's movement, they exhibit a higher response rate and make social displays that are normally directed toward caregivers. He proposed that contingency is used by young infants to recognize social agents.

Specifically, a contingent response is a change in one agent's behavior *within a specific time window* in direct response to a signal from another agent. In the context of Human-Robot Interaction (HRI), when a social robot generates a signal to a human, contingency detection should help determine their willingness to interact (e.g., Fig. 1).

In this paper we build on the prior work of Lee et al. [1], [3], in which they characterize the problem as one of detecting a significant change between the time before and after the robot's signal across both visual and auditory channels. One limitation of this prior work is that it was never tested in a real-time interactive setting, only on datasets. Additionally, the approach has a computationally expensive graph analysis component that is a barrier to easily deploying the model in a real-time setting. Given these two observations, in this paper we explore an alternative approach to contingency detection, a simple discriminative Support Vector Machine (SVM) classifier. We present two experiments. The first is a data collection and model training experiment similar to that presented in [1], in which we show that the SVM classifier is able to achieve an average accuracy of 74% and an average F_1 score of 0.82, which is comparable to the performance reported in prior works. Then, in a second experiment with seven human participants, we show that our system is able to perform contingency detection in real-time with an accuracy of 67% and F_1 score of 0.72.

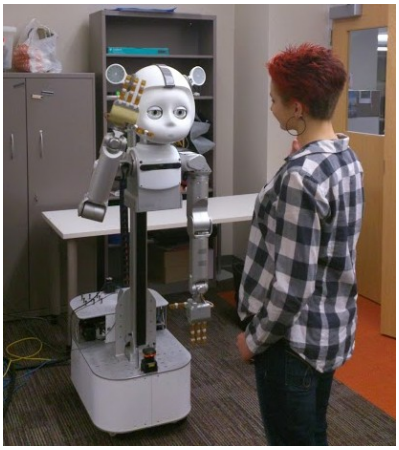


Fig. 1: Curi performing the Wave signal with a participant.

II. Related Work

There are other examples of computational models of contingency in prior work. In a seminal example, Movellan [4] developed the Infomax controller that optimally queried the environment with the motion of a single actuator, and determined if a detected audio signal was due to hearing the robot's own motion or to the presence of a social agent. In related works, a robot learns expected contingency windows for its own actions [5] including auditory responses from a social partner [6]. In all of these, the research question is around determining the expected timing windows for a contingent response, and is limited to a single channel of communication. Much of the previous HRI literature focused on contingency detection has targeted a single mode of communication, *i.e.* visual or auditory feedback [1], [7], [8]. We are looking at response detection and are including both visual and auditory cues in our approach to the problem.

One way to approach contingency detection is as an activity recognition problem. A good example of this is seen in the recent work of Rich et al. [9] that recognizes "engagement"

*This work was supported by NSF grant 0953181 and ONR grant N000141210484.

[1]Authors are affiliated with the School of Interactive Computing, Georgia Institute of Technology, 801 Atlantic Dr., Atlanta, GA 30332. [vchu,ksbullard]@gatech.edu,athomaz@cc.gatech.edu

for a humanoid robot. Using both visual and auditory cues, they recognize four types of events to determine that a human is responsive to the robot: directed gaze, mutual facial gaze, conversational adjacency pairs, and backchannels. Bohus and Horvitz [10] similarly estimate engagement through the recognition of a variety of specific engagement cues (salutations, calling behaviors, specific approach trajectories and formations). However, natural human responses are often varied in real interaction scenarios. Rather than having to recognize specific behaviors, our goal is to generally classify any change in behavior as contingent.

To tackle real interaction scenarios, approaches using visual motion have been promising. Muller et al. [11] use motion trajectories and focused on selecting features from the raw information to classify engagment. More recent work from Lee at al. [1] build on this approach of using visual motion and construct graphs on the motion captured rather than just using feature selection methods. However, both methods used only visual inputs and were not shown to operate in real-time.

We frame the problem in the same way as Lee et al. [1], repeated here for convenience. When the robot makes a signal to a human at time t_s, there are two time windows of interest, W_B: the time window before the robot's signal; and W_A: the time window after the signal. Instead of attempting to detect specific events, or actions, that would reflect a person's level of engagement, the goal is to monitor change in the human's behavior generally between W_B and W_A.

The approach taken in [1] relies on building a similarity graph within the frames of W_B and between W_B and W_A, and analyzing these graphs to determine whether or not a significant change has occurred between W_B and W_A. This model showed performance with an accuracy of 79% on a collected dataset of contingent/non-contingent events in an HRI scenario. In follow-on work [3], this model was extended to include both visual and auditory cues, as well as prior knowledge about when to expect a human response to various robot signals, letting the robot more accurately decide what to include in the W_A window. This multimodal version with timing information showed nice performance of 91% (on a different dataset than [1]). Neither approach was ever shown to work in a real-time interactive setting. Our goal is to show accurate contingency detection in a real-time HRI setting. As such, in this paper we are proposing an alternative to this prior work. The graph building and analysis component are a computational bottleneck, and in this paper we explore the alternative of a simple discriminative classifier trained with positive and negative examples of contingency across the W_B and W_A windows.

III. Approach

A. Robot Platform

For our experiment we used "Curi", a humanoid mobile robot with two 7-degree-of-freedom (DOF) arms, an omni-directional mobile base, and a socially expressive head.

We used an Asus Xtion PRO Live RGB-D camera (ASUS) and a two-microphone stereo array mounted in the center of the robot's chest. We record amplitude values from the microphone array sampled at 44.1 KHz. We also use OpenNI's human tracker to record translation and rotation data of several different body parts for each time frame. The tracker contains information about the pose of the person's head, neck, torso, left and right shoulder, left and right elbow, left and right hand, left and right hip, left and right knee, and left and right foot.

For the purpose of experimenting with contingency detection we use a simple finite state machine (FSM) controller with six distinct states:

- **Idle** - Curi actively monitors the scene to detect if a person has entered her field of view.
- **PreSignal** - Curi enters this state once a person is detected. The person is tracked for the remaining duration of the interaction. Curi starts recording data from the ASUS and microphone sensors in this state, capturing the data for the before window of interest, W_B.
- **PerformSignal** - Curi performs a greeting once the person comes within a set threshold distance and t_s represents the moment Curi performs the signal.
- **PostSignal** - Curi observes the person's response for a specified time window, capturing sensor data for the W_A time window.
- **DetermineContingency** - Given the W_B and W_A data, a classification of contingent/not-contingent is made with the SVM model (to be detailed). If the model determines a contingent response, Curi moves to the engagement state. Otherwise, returns to **Idle**.
- **Engagement** - Curi continues interacting with the person, selecting one of several spoken utterances. Once the interaction is completed, Curi returns to **Idle**.

B. Contingency Cue Features

To detect a significant change between **PreSignal** and **PostSignal**, we select cues that can distill the sensor data into an aggregate measure, a holistic view of the user's current situation. We use the change in audio and body motion before and after the robot greeting; based on time windows reported in prior work we set the length of W_B to two seconds and W_A to three seconds. For each channel of audio (left and right), we compute the sound cue, c_s, as the average difference in amplitude between adjacent time frames:

$$c_s^{(j)} = \frac{1}{n-1} \sum_{t=1}^{n-1} a_{t+1} - a_t \tag{1}$$

where n is the number of time frames in the window and a_t is the amplitude of the audio signal, for channel j at time t. This results in two audio features for each time window.

For the body motion cue, we compute a measure of the aggregate variance of translations and rotations of each joint. To reduce the dimensionality of the cue vector and efficiently generalize motion, we merge body parts into six larger connected components: head, torso, left arm, right arm, left leg, and right leg. For example, the left arm component includes left hand, left elbow, and left shoulder. The torso is the only body part that we include both the

TABLE I: Set of possible human actions during model training data collection

Behavior Before Signal	Behavior After Signal	Contingent Action
From far off, walk towards Curi	Stop in place and say a greeting	Yes
From far off, walk towards Curi	Stop in place and wave	Yes
From far off, walk towards Curi	Stop in place, wave, and say a greeting	Yes
Within Curi's field of view and talking in another direction	Stop talking and look at Curi	Yes
Within Curi's field of view and talking in another direction	Stop talking and look at Curi and say Yes	Yes
Within Curi's field of view and talking in another direction	Stop talking and look at Curi and wave	Yes
Within Curi's field of view and facing Curi	Walk up to Curi and say greeting	Yes
Facing away from Curi and in field of view	Turn and look at Curi	Yes
Facing away from Curi and in field of view	Turn and say greeting	Yes
Facing away from Curi and in field of view	Turn, look, and wave at Curi	Yes
Facing away from Curi and in field of view	Turn, look, wave, and say greeting at Curi	Yes
Facing away from Curi and in field of view	Walk up to Curi and say greeting	Yes
Tying shoe in Curi's field of view	Stop and look at Curi	Yes
Tying shoe in Curi's field of view	Stop, get up, and look at Curi	Yes
Tying shoe in Curi's field of view	Stop, get up, say a greeting, and wave	Yes
From far off, walk towards robot	Continue walking past robot	No
Facing away from Curi and in field of view	Do not move	No
Tying shoe in Curi's field of view	Continue tying shoe	No
Within Curi's field of view and talking in another direction	Continue talking	No

translation and rotation values. For the head component, we only include rotation, and for all other body parts, we only look at translation. We compute the motion cue, c_m, for each component j, using Equation 2:

$$c_m^{(j)} = \frac{1}{n-1}\sqrt{\sum_{t=1}^{n-1}(b_{t+1} - b_t)^2} \quad (2)$$

where b_t represents the body motion at time t. This results in seven features for W_B and seven features for W_A.

C. SVM Contingency Classifier

Our focus is on building a system that can use contingency detection in real-time. We believe that a discriminative approach could prove more efficient than a generative approach if we can show that performance is similar.

Given a dataset of positive and negative examples of contingency between W_B and W_A, we trained three separate SVMs: using audio cues only, body motion cues only, and using both. The input for this classification problem is the two windows W_B and W_A merged together to form a single input vector. Thus in the audio only case this input vector has 4 features (2 x 2 audio cues), in the body motion only case it has 14 (2 x 7 body motion cues), and the combined case has all 18 features.

In the next section we describe the training dataset. During both experiments described in the following sections, we used 5-fold cross validation on a training set to select the "best" SVMs for all three inputs (audio, body motion, and merged). We also cross validated on different kernels: RBF, Linear, and Polynomial. All "best" SVMs were selected using the scores found during cross validation and trained with weighted bias to account for our unbalanced training set as described in the next section. All SVMs used came from the python machine learning library scikit-learn [12].

IV. Training Contingency Detection Models

A. Data Collection

In order to collect a data set with a wide range of contingency behavior, we systematically collected a specified set of behaviors from three different people (the authors). We collected this data by having each person perform each of the short interactions listed in Table I. Each interaction began with the human starting the "before signal" behavior, and once the robot detected a person in the field of view it would provide one of the following signals: a verbal greeting of "Hi", or a waving gesture (see Figure 1). Each person did all of the interactions in Table I for each type of robot signal.

All of the contingent interactions were collected twice and the non-contingent interactions were collected three times. This was done to balance the number of positive and negative examples of contingency. Originally, we had more non-contingent interactions than the ones listed, but have since realized that many were actually "subtly contingent" (e.g. the person looks at the robot, says "hold on" and looks away). We have removed these examples from the dataset, resulting in an unbalanced set of contingent vs. non-contingent trials. The final set seen in Table I contain 15 examples of contingency vs. only 4 examples of non-contingency, resulting in 30 trials of contingent reactions and 12 trials of non-contingent reactions per person and per robot signal.

We created multiple training and test sets from this data. To simulate a real-world scenario, we used a leave-one-user out approach for splitting the training and testing sets. Thus, we have three separate train/test sets, each with 84 training examples and 42 testing examples for each robot signal.

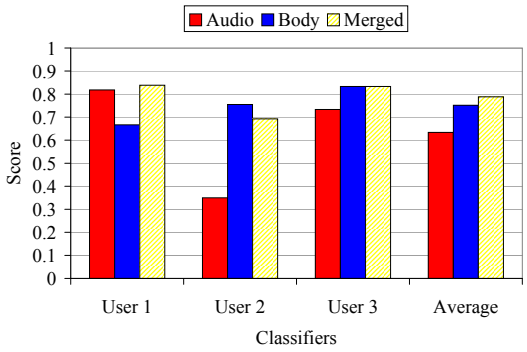
B. Model Performance

We used the standard metrics of precision, recall, and F_1 score, as defined in Table II, where tp is true positives, tn is true negatives, fp is false positives, and fn is false negatives. We also computed accuracy across the entire test set.

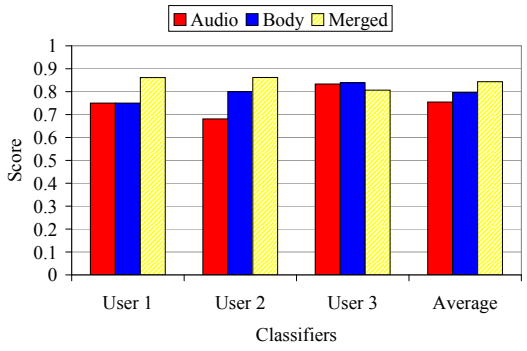
The results of testing on each separate test set can be seen in Figure 2, which shows the F_1 scores by cue type and test set. The **Hi** and **Wave** signal has an average F_1 scores of 0.79 and 0.84 respectively. When the robot performs the **Hi** signal, merging the two cues (audio and body motion)

TABLE II: Metric Equations

Precision	Recall	Accuracy	F_1
$\frac{tp}{tp+fp}$	$\frac{tp}{tp+fn}$	$\frac{tp+tn}{tp+fn+tn+fp}$	$2 \cdot \frac{precision \cdot recall}{precision + recall}$



(a) **Hi** Signal. Average F_1 scores of 0.79 for merged cues



(b) **Wave** Signal. Average F_1 score of 0.84 for merged cues

Fig. 2: F_1 scores, training on 2 people and testing on 1, for both robot signals. See also Table III

does not seem to help the classifier's performance. However, in the wave case, we clearly see an improvement of scores when using audio and body cues.

To gain a better understanding of the performance of the classifiers, we broke down the results to show how the classifiers performed on the contingent vs. non-contingent experiments. These values as well as the accuracy, precision, and recall scores are shown in Table III. The **Hi** and **Wave** signal has an average F_1 scores of 0.79 vs. 0.35 and 0.84 vs 0.56 respectively. All classifiers perform better on the contingent examples than non-contingent examples, which we expected given the unbalanced nature of our training data. However, we would like to note that the classifiers do not label everything as contingent to achieve these scores.

To understand how the models performed compared to the prior work, we computed the results of the different cues using accuracy, the metric used by Lee et al. [3] [1]. These values can be seen in Table III. The **Hi** and **Wave** signals obtained an average accuracy of 70% and 77% respectively. To understand these scores, we first need to look at the experiment data collected by both prior works.

TABLE III: Detailed Model Training Classification Results

Signal	Data	Acc.	Label	P	R	F_1
Hi	Audio	0.57	C	0.73	0.62	0.63
			NC	0.37	0.44	0.36
			A	0.63	0.57	0.63
	Body Motion	0.66	C	0.81	0.82	0.82
			NC	0.54	0.53	0.54
			A	0.74	0.74	0.82
	Merged	0.70	C	0.78	0.82	0.79
			NC	0.33	0.39	0.35
			A	0.65	0.70	0.79
Wave	Audio	0.67	C	0.82	0.74	0.75
			NC	0.29	0.50	0.36
			A	0.67	0.67	0.75
	Body Motion	0.70	C	0.80	0.80	0.80
			NC	0.46	0.47	0.46
			A	0.70	0.71	0.80
	Merged	0.77	C	0.82	0.87	0.84
			NC	0.62	0.53	0.56
			A	0.77	0.77	0.84

Key: (Acc)uracy, (P)recision, (R)ecall, (C)ontingent, (N)ot (C)ontingent, (A)verage
Note: (A)verage is the weighted average of contingent and noncontingent for precision, recall, and F_1. The darker the color, the better the score

In [1], the problem statement was most similar to ours because the authors used contingency detection for response detection. The dataset collected by this work matched the type of dataset used in our model training. The work looked at using two different visual cues to improve contingency detection, but did not use audio cues. Furthermore, the work only tested robot cues that were physical, similar to **Wave**, but did not have any audio robot signals. The highest accuracy obtained by [1] was 79%.

Lee et al. [3] extended the prior model and integrated various cues including audio. Additionally, the follow-on work included prior knowledge about the typical response delay for the various robot signals used in the task. The evaluation dataset for this work was from sessions of people playing the game Simon Says with a robot. This dataset breaks down into two components: negotiation phase, where the robot listens for audio cues, and a game phase, where users mainly use physical motion to mimic the robot. This is arguably an easier dataset for merging audio and visual cues because the interaction often was entirely audio or entirely visual. Our training set including many examples where both audio and visual cues were necessary to determine contingency. The highest average accuracy obtained by [3] was 91%. This in part goes to show that the staged contingency events that we use for evaluation (also used in [1]) may just be a harder problem and more similar to a realistic interactive task scenario.

The results of training our classifiers fill in the missing gap of how using audio and motion cues effect the accuracy of using contingency detection for response detection. We obtain a comparable accuracy to [1], especially when we compare the results of **Wave** to their physical robot cues.

V. Interactive Contingency Detection

After we trained our contingency detection classifiers, our second experiment tests their ability to be used in real-time situations and to generalize to new users.

A. Extending to Real-Time

To classify users in real-time, we use the same sensors, controller, and FSM discussed in Section III-A. However, instead of just recording the data, Curi processes the entire interaction example in **DetermineContingency** as soon as the **PostSignal** state completes. During **DetermineContingency**, Curi calculates the audio and body motion cues and uses the best merged SVM classifier from training to decide contingency. Curi performed a follow-up sentence if contingency was detected, or otherwise said "Bye."[1] The average time for Curi to respond was less than 2 seconds.

B. Experiment Design

In our validation experiment we test the system with seven new humans (5 male, 2 female) that were not part of the training set. The interaction was similar to that used for collecting our training examples. We designed three different "before signal" scenarios for the experiment:

- Within view of the robot and talking on the phone
- Within view of the robot and reading a book silently
- From far off, walk towards Curi

All participants did each scenario four times, 2 contingent and 2 non-contingent, resulting in 84 different interactions. We randomly assigned the scenario order, as well as the order of responding contingently or not.

Participants were instructed before each scenario, to respond contingently or not to the robot's greeting. In the non-contingent trials, they were told to ignore the robot. For the contingent runs, there was no behavior specification given, the participants was asked to respond naturally.

In this experiment, we only used the **Hi** signal from Curi. After doing pilot runs with users, we discovered the **Wave** signal was too slow for people to react and engage naturally. Participants were confused as to why the robot was not responding until after the wave was done. This can be addressed in future work by speeding up the gestures, and by allowing a cross modal response (i.e., responding in the speech channel before the body gesture has finished). This was not done for this experiment because the change in the signal would invalidate the data collected during contingency modeling (Section IV) and require new data to be collected and new SVMs to be trained.

C. Results

The results from this experiment used the same evaluation metrics in Table II and were obtained by testing with the three separate classifiers trained using leave-one-user-out testing (C_{louo}) in Section IV. The individual and average F_1 scores can be seen in Figure 3 displayed by cue type. The detailed scores can be seen in Table IV.

The average F_1 score is 0.54 when using both audio and body motion cues. However, if we look at the just using audio cues, the average F_1 score jumps to 0.72.

To see if we could improve these results, we built a model (C_{all}) with all of the data from the first experiment. Recall,

[1]This interaction can be seen in the video included with the paper.

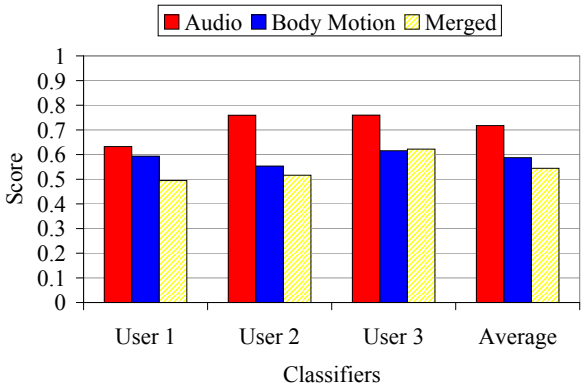


Fig. 3: F_1 scores of Leave-One-User-Out on test set. For exact numbers refer to Table IV

TABLE IV: Average Leave-One-User-Out (C_{louo}) Results

Data	Acc.	Label	P	R	F_1
Audio	0.66	C	0.65	0.83	0.72
		NC	0.68	0.48	0.53
		A	0.67	0.66	0.72
Body Motion	0.51	C	0.51	0.71	0.59
		NC	0.47	0.31	0.36
		A	0.49	0.51	0.59
Merged	0.43	C	0.47	0.67	0.54
		NC	0.39	0.25	0.28
		A	0.43	0.46	0.54

Key: (Acc)uracy, (P)recision, (R)ecall, (C)ontingent, (N)ot (C)ontingent, (A)verage
Note: (A)verage is the weighted average of contingent and noncontingent for precision, recall, and F_1. The darker the color, the better the score

the previously built models only used data from two users at a time. C_{all} was trained using the same cross validation method described in Section III-C, except we removed cross validation over different kernels. Whenever the model used a kernel other than the linear kernel, it would label all runs in the cross validation test set as contingent. The results of C_{all} is seen in Table V. Again, none of the trained classifiers saw any examples in the newly collected test set, even when selecting the classifier to use.

The resulting F_1 score improves to 0.63 when using C_{all}. However, the F_1 score when using just the audio cue drops to 0.61. When looking closer, this occurs because the C_{all} classifier does better at the non-contingent cases, but drops slightly for the contingent cases. Figure 4 shows the breakdown in F_1 scores and accuracy for both C_{all} and C_{louo} on the different cues. The highest accuracy obtained (67%) is by the audio cue classifier C_{all}.

Almost all of the SVMs trained with just the audio cues outperformed the other classifiers regardless of training dataset or score metric. During both the training and test experiments, we noticed a tendency of people to respond with the same modality that the robot used to initiate the interaction. When Curi provided a verbal greeting, people were inclined to respond verbally. When Curi waved, people were inclined to respond with a gesture. We believe this largely contributed to the superior performance of the audio classifier in the real-time validation set. This is a promising opportunity for future work where a classifier could be

TABLE V: All Users Trained (C_{all}) Results

Data	Acc.	Label	P	R	F_1
Audio	0.67	C	0.73	0.52	0.61
		NC	0.63	0.81	0.71
		A	0.68	0.67	0.61
Body Motion	0.52	C	0.52	0.71	0.60
		NC	0.54	0.33	0.41
		A	0.53	0.52	0.60
Merged	0.57	C	0.55	0.74	0.63
		NC	0.61	0.40	0.49
		A	0.58	0.57	0.63

Key: (Acc)uracy, (P)recision, (R)ecall, (C)ontingent, (N)ot (C)ontingent, (A)verage
Note: (A)verage is the weighted average of contingent and noncontingent for precision, recall, and F_1. The darker the color, the better the score

selected based on the modality of the robot's signal to boost the overall performance.

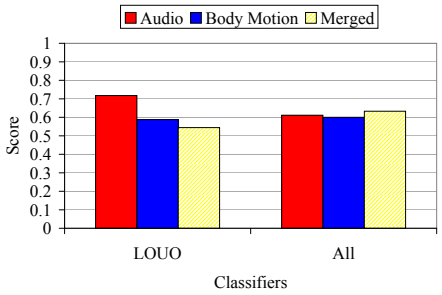
The performance drop between our trained model F_1 score of 0.79 to the real-time experiment dataset F_1 score of 0.72 was expected due to the difficulty of generalizing to the real-time experiment where participants were free to do any contingent reaction. Some particularly difficult cases were participants who continued their body behavior and simply verbalized a responses to Curi (e.g. asking the person to "hold on" when on the phone before beginning a conversation with Curi). From the perspective of the model, there was no distinctive difference in sensory input before and after the signal even though the participant was contingent.

To improve the real-time classifier results, there are some avenues of future work. First, collecting a more balanced dataset with a wider range of reactions from participants. Another is to use a second level Bayesian inference model similar to the approach used in [3] to combine cues on a decision level, when each individual cue classifier has returned a prediction. Currently, we have merged at the module level by putting the cues directly together into one feature vector. A final improvement for future work is to explore incremental recognition, allowing the robot to signal again in order to determine, with higher confidence, the user's intent.

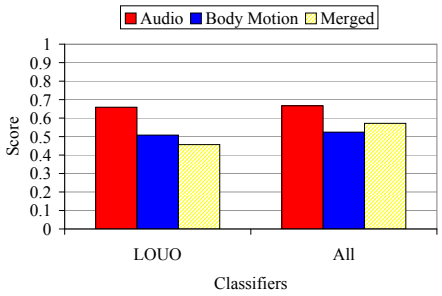
VI. Conclusion

We tested a discriminative supervised classifier approach for using contingency to solve the engagement detection problem. We trained SVM models using two different sensory inputs for two different robot signals. In training, the models were able to successfully detect engagement on a completely different user with an average F_1 score of 0.79 for the **Hi** signal and 0.84 for the **Wave** signal. These results are comparable to previous approaches.

We successfully implemented the engagement detection system in real-time, a task that the prior work did not attempt. We tested the system with seven new users and showed that the trained model was able to generalize to these new users with a best average F_1 score of 0.72. These results show that contingency can be used to successfully detect user engagement in real-time.



(a) Average F_1 Scores



(b) Average Accuracy Scores

Fig. 4: Average F_1 score and Accuracy for C_{louo} and C_{all}

References

[1] J. Lee, J. Kiser, A. Bobick, and A. Thomaz, "Vision-based contingency detection," in *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, 2011, pp. 297–304.

[2] J. S. Watson, "Smiling, cooing, and" the game"," *Merrill-Palmer Quarterly of Behavior and Development*, vol. 18, no. 4, pp. 323–339, 1972.

[3] J. Lee, C. Chao, A. F. Bobick, and A. L. Thomaz, "Multi-cue contingency detection," *International Journal of Social Robotics*, vol. 4, no. 2, pp. 147–161, 2012.

[4] J. R. Movellan, "An infomax controller for real time detection of social contingency," in *Development and Learning, 2005. Proceedings. The 4th International Conference on*. IEEE, 2005, pp. 19–24.

[5] A. Stoytchev, "Self-detection in robots: a method based on detecting temporal contingencies," *Robotica*, vol. 29, no. 1, pp. 1–21, 2011.

[6] K. Gold and B. Scassellati, "Learning acceptable windows of contingency." *Connect. Sci.*, vol. 18, no. 2, pp. 217–228, 2006.

[7] M. Michalowski, S. Sabanovic, and R. Simmons, "A spatial model of engagement for a social robot," in *Advanced Motion Control, 2006. 9th IEEE International Workshop on*, 2006, pp. 762–767.

[8] N. J. Butko and J. R. Movellan, "Detecting contingencies: An infomax approach," *Neural Networks*, vol. 23, no. 8, 2010.

[9] C. Rich, B. Ponsler, A. Holroyd, and C. Sidner, "Recognizing engagement in human-robot interaction," in *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, 2010, pp. 375–382.

[10] D. Bohus and E. Horvitz, "Models for multiparty engagement in open-world dialog," in *Proceedings of the SIGDIAL 2009 Conference*, ser. SIGDIAL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 225–234.

[11] S. Muller, S. Hellbach, E. Schaffernicht, A. Ober, A. Scheidig, and H. Gross, "Whom to talk to? estimating user interest from movement trajectories," in *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, 2008, pp. 532–538.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.