Temporal Models for Robot Classification of Human Interruptibility

Siddhartha Banerjee
Georgia Institute of Technology
801 Atlantic Dr NW
Atlanta, GA 30332, USA
siddhartha.banerjee@gatech.edu

Sonia Chernova
Georgia Institute of Technology
801 Atlantic Dr NW
Atlanta, GA 30332, USA
chernova@gatech.edu

ABSTRACT

Robots are increasingly being deployed in unstructured human environments where they will need to approach and interrupt collocated humans. Most prior work on robot interruptions has focused on *how* to interrupt a person or on estimating a human's awareness of the robot. Our work makes three contributions to this research area. First, we introduce an ordinal scale of interruptibility that can be used to rate the interruptibility of a human. Second, we propose the use of Conditional Random Fields (CRFs) and their variants, Hidden CRFs, and Latent-Dynamic CRFs, for classifying interruptibility. Third, we introduce the use of object labels as a visual cue to the context of an interruption in order to improve interruptibility estimates. Our results show that Latent-Dynamic CRFs outperform all other models across all tested conditions, and that the inclusion of object labels as a cue to context improves interruptibility classification performance, yielding the best overall results.

Keywords

Human-Robot Interaction; Interruptibility; Conditional Random Fields

1. INTRODUCTION

Robots are increasingly being deployed in unstructured human environments where they will need to approach collocated humans, whether to signal task completion, to report a problem or to offer a service. These interactions will often serve as interruptions for the humans involved, who might already be engaged in other tasks demanding high cognitive load. Research has shown that badly timed interruptions have the potential for negative impacts on human task performance [10]. Conversely, additional work has shown that in the right context at the right time, interruptions can actually be beneficial to human task performance [24]. This importance to the timing of an interruption is captured by the notion of *interruptibility*, which has been defined as a measure of the receptiveness of a human to receive external disturbances (interruptions) at any given moment [25]. Low interruptibility signifies the human's desire to not be disturbed, while high interruptibility signifies that the human could be amenable to an interruption.

Humans are very adept at gauging the interruptibility of others from observation [18]; our goal is to enable robots to similarly classify interruptibility given a short window of observation. Given multiple people in a typical office environment, some of whom may be working while others chat over coffee, we want to enable a mobile robot to determine which individuals it is most appropriate to approach in order to minimize the adverse effects of the interruption. This research problem is closely related to prior work in robotics on using a Hidden Markov Model (HMM) [17] to estimate a human's intention to engage in an interaction with a robot [15].

Our work makes three contributions to this research area. First, we introduce an ordinal scale of interruptibility that can be used to rate the interruptibility of a person and to influence decisions on whether or not to interrupt. Second, we explore the use of Conditional Random Fields (CRFs) [12] and their variants, Hidden CRFs (HCRFs) [26], and Latent-Dynamic CRFs (LDCRFs) [16], for classifying interruptibility. Using a dataset of person observations collected by a mobile robot, we compare the performance of these models against HMMs and show that the LDCRF consistently outperforms all other models across all tested conditions. Third, motivated by work on interruptibility in other areas of computing [28], we introduce the use of object labels as a visual cue to the *interruption context* in order to improve interruptibility estimates. Our results show that inclusion of object labels as a cue to context improves interruptibility classification performance, yielding the best overall results.

2. RELATED WORK

Research in psychology and human factors has long studied interruptions and their effects. Miyata and Norman [14] and Speier et al. [24] identify the appropriate timing of interruptions as a key factor in determining how an interruption affects humans. This work supports our motivation to accurately classify *interruptibility*, which is defined as the receptiveness of a person to interruptions at a point in time [25].

The fields of Human-Computer Interaction and Ubiquitous Computing have built upon the psychology literature in order to identify computable descriptors for reasoning about interruptions [28]. A key focus in these works has been estimating a person's *workload*, either through direct observation or knowledge of the person's tasks [10]. In robotics, task workload modeling has been accomplished though cognitive architectures and human performance moderator functions [8]. However, this approach requires detailed models of the human's tasks and extensive monitoring of the human's behavior, which are unavailable in our application.

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

Within robotics, interruptibility has been considered directly. Rosenthal et al. [19] accumulated knowledge on the occupancy schedule of people in their offices and used that to predict their availability, under the assumption that a person was interruptible if their door was open. However, this approach did not take into account the social cues that signaled interruptibility. Satake et al. [21] used SVMs trained on trajectories of shoppers within a shopping mall to predict interruptibility, and Shi et al. [23] created and verified a state-based model for engaging a target human. But these works relied on extensive instrumentation of the environment (building-mounted lasers and motion capture system), and are thus not directly applicable to a mobile robot operating in an unstructured environment.

Prior research has also focused on detecting human *intent-to-engage* or on estimating the human's level of awareness of the robot in the context of companion robots [15, 3], shopping mall assistants [11], receptionists [2] and bartenders [6]. Estimation of engagement and awareness is closely related to our work on interruptibility, and we build on this prior work, particularly in the choice of perception features used to monitor the person's behavior. However, classifying interruptibility poses its own research problem because interruptibility can be high even when the person shows neither intent-to-engage nor awareness of the robot. Most closely related to our work in this area is that of Mollaret et al. [15] and Chiang et al. [3]. Mollaret et al. present an approach for estimating human intent-to-engage when modeled by an HMM using audio features, body position features, and the head gaze features. Chiang et al. also use HMMs and a similar set of features to estimate a human's awareness of the robot. In our work we use HMMs and features derived from [15, 3] as a baseline.

Additional research has focused on determining the best way to perform the interruption once a human is defined as interruptible. Saulnier et al. [22] have explored the most appropriate set of nonverbal behaviours for interruptions while Chiang et al. [3] have used Reinforcement Learning to personalize interruption behaviors. Our work focuses on the first part of the problem, which is to classify whether the person is interruptible at a point in time.

Finally, prior work highlights the importance of *interruption context* [18, 28]. Computationally, the interruption context broadly consists of features that describe the user (e.g., personality traits) [25, 27], the task [1, 10], the environment [5, 27], the interruption [9], and the relationships between these [13, 20]. In our work, we focus on garnering environment context: we hypothesize that the labels of objects that a person is interacting with can serve as valuable contextual cues to improve classification of interruptibility.

3. DEFINITION OF INTERRUPTIBILITY

Interruptions are defined as "externally generated, randomly occurring, discrete events that break the continuity of cognitive focus on a certain task" [24], and the *interruptibility* of a person at any given point in time is defined in terms of their receptiveness to interruptions at that moment [25]. When a person is focused on their current task and not amenable to an interruption, they are said to have low interruptibility; meanwhile a person amenable to interruptions is said to have high interruptibility.

It is important to distinguish interruptibility from the decision to interrupt. The interruptibility of a person tries to quantify the *disturbance* that a person might experience as a result of an interruption, while the decision to interrupt depends upon a person's interruptibility as well as other factors such as the urgency and characteristics of the interrupting task [18, 20]. In this work, we focus on the classification of interruptibility with the goal of using the results later within a broader framework for deciding when to interrupt.

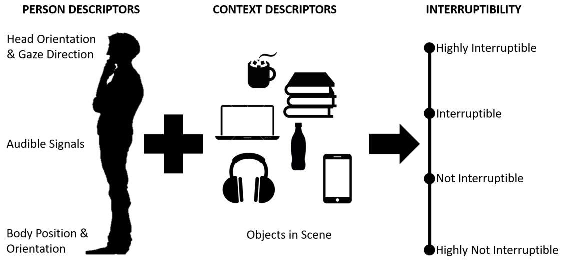


Figure 1: The level of interruptibility of a person is represented on a four point scale. In order to arrive at a value on this scale, we use information about person state and interruption context. In this paper, we use object labels as a cue to the context.

To characterize an individual's interruptibility, we propose the following interruptibility scale:

INT-4 **Highly Interruptible**. The person is not busy and they are aware of the robot's presence.

INT-3 **Interruptible**. The person is not busy, but they are unaware of the robot's presence.

INT-2 **Not Interruptible**. The person is busy, but the robot may interrupt if necessary.

INT-1 **Highly Not Interruptible**. The person is very busy, the robot should not interrupt.

INT-0 **Interruptibility Unknown**. The robot is aware that a person is present, but does not have sufficient sensory input to analyze interruptibility.

Values 1-4 in the scale capture the full range of interruptibility states that can help guide the robot's decision making process. We include the rating of 0 to represent states in which the robot does not yet have sufficient information about the person, such as when the person is too far away or out of view. In this case the robot may choose to approach another person, or take actions to improve sensing quality.

4. PERCEPTION OF INTERRUPTIBILITY

As defined in prior work, interruptibility can be characterized based on two sources of information – *person state* and *interruption context* (Figure 1). Person state, inferred from laser, video, and audio sensor data, has been widely used to estimate level of engagement and human awareness in robotics [15, 3]. Based on our survey of research in this area, we propose to use the following information categories to represent person state in this paper:

- The **position and orientation** of a person within the environment. This includes where they are located as well as how their body is oriented with respect to the robot.
- The **head orientation and gaze direction** of the person.

- The **presence and orientation of sound** within the environment.

Interruption context can be inferred from many factors, including known information about the user, the task, the environment and the type of interruption [28]. In this work, we consider environmental (or scene) context, which we define as visually observable cues that may inform the robot about the interruptibility of a person. In particular, we propose that objects the person is interacting with can serve as useful visual context cues. For example, an individual nursing a coffee mug in a lounge is judged to be more interruptible than someone engaged with a laptop in the same lounge. Although objects cannot capture all of the complexities of interruption context, object recognition is widely available on robotic systems and we hypothesize that, combined with traditionally used cues of person state, object labels can improve the estimate of a person's interruptibility. Thus we add the following information type to our perception model:

- The **labels of objects** that are being used by the person, or those that lie near them.

The contribution of our work does not lie in the realm of object detection and hence we use hand annotated object labels in this paper. In the following section, we describe how this information can be leveraged in multiple computational models.

5. MODELS FOR INTERRUPTIBILITY

Following the example of prior work, we used temporal models to estimate interruptibility given data inputs of the form in Section 4. Mollaret et al. [15] and Chiang et al. [3] both used Hidden Markov Models to address related problems, with promising results, and so we adopt this model as our baseline. In this paper, we explore the use of Conditional Random Fields [12] and derivatives thereof, Hidden Conditional Random Fields [26], and Latent-Dynamic Conditional Random Fields [16], as alternate temporal models to classify interruptibility. We hypothesize that CRFs will outperform HMMs in classifying interruptibility due to their more expressive representation. We also hypothesize that HCRFs and LDCRFs will perform better than the CRFs due to their ability to model intra-class variation within observed data. Below we present an overview of each of these models and how they can be utilized for interruptibility classification.

5.1 Hidden Markov Models

An HMM [17] models two stochastic processes. The first process is a Markov chain through a sequence of discrete hidden states, while the second process produces observable continuous or discrete emissions given a hidden state (see Figure 2a). HMMs have found widespread use in areas such as natural language processing and speech recognition, and in the context of human-robot interaction have been used for tasks such as activity recognition and human engagement detection. Given their use in the works of Mollaret et al. [15] and Chiang et al. [3], we assume HMMs are an appropriate model to classify interruptibility and therefore use them as a baseline for comparison.

The HMM is characterized through five parameters

$$\lambda = (N, M, A, B, \pi)$$

where each of the parameters has the following significance:

N is the number of hidden states in the model. Although it is common for the hidden states to have some physical significance, this need not be the case.

M is the number of distinct observation symbols per state if the observation sequence is discrete valued. In the case of continuous observation sequences, M denotes the number of mixture components that contribute to producing an observed value.

A is an $N \times N$ state transition matrix where each element of the matrix signifies the probability of transitioning from one hidden state to another.

B is the observation symbol probability distribution for all hidden states. In the case of discrete emissions, B is an $N \times M$ matrix; in the case of continuous emissions, B is a parameterized specification of M mixtures (usually Gaussian) for each of the N hidden states.

π is the initial state distribution over the hidden states.

To classify interruptibility, we trained separate ensembles of HMMs for each of the five different interruptibility classes that we have defined. Within each ensemble, we trained a separate HMM for each of the features (see Section 7) in the data sequences that we used. We used a uniform initial distribution over all hidden states, and we modeled the continuous valued features using Gaussian Mixture Models. The HMMs were implemented with the GHMM library[1] and trained using Baum-Welch.

Given a sequence of data, each of the HMMs in each ensemble ran the Forward algorithm to return a log likelihood of the data being generated by the HMM; the log likelihood result for the ensemble was taken to be the sum of log likelihoods within the ensemble. Finally, the interruptibility label derived for the sequence was the maximum log likelihood from each of the different ensembles.

5.2 Conditional Random Fields

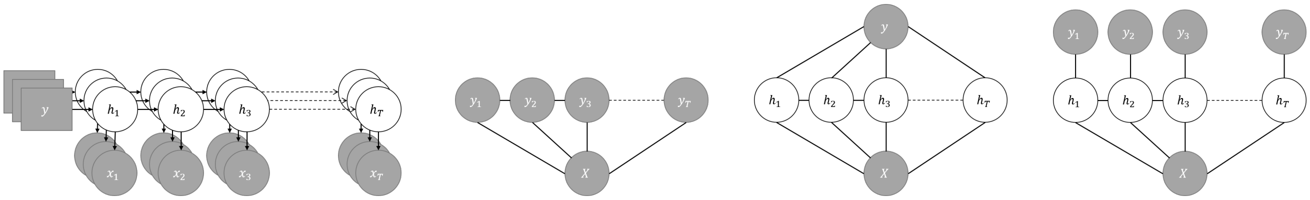
Represented as an undirected graphical model, a CRF [12] models the probability of a label sequence conditioned on the entire observation sequence (see Figure 2b), as opposed to an HMM which models the joint probability of both the hidden state for a label and the observation at any timestep. This allows the CRF to utilize domain knowledge and incorporate information over multiple timesteps within the observation sequence without violating the assumptions of the model. In addition, the CRF is capable of linking state transitions within the model directly to the observations. This capability allows a richer specification of relevant factors within the model using prior domain knowledge. Previous work has successfully demonstrated the superiority of CRFs over HMMs in the realms of Activity Recognition [29] and Natural Language Processing [12], leading us to hypothesize that CRFs hold promise for gauging interruptibility.

Concretely, the CRF model provides

$$P(Y|X) = \frac{1}{Z} \prod_{t=1}^{T} \Psi_t(Y_a, X)$$

$$Z = \sum_{Y_a} \prod_{t=1}^{T} \Psi_t(Y_a, X)$$

[1] http://ghmm.sourceforge.net/index.html



(a) HMM (b) CRF (c) HCRF (d) LDCRF

Figure 2: Graphical representation of each of the computational models in this paper. Gray elements represent observed variables, and white elements represent hidden variables

where $Y = \{y_1, y_2, ..., y_T\}$, each $y_i \in \mathcal{Y}$, is the label sequence, $\mathcal{Y}$ is the set of possible labels, X is the observation sequence, Z is a normalization function, and T is the length of the observation sequence. Y_a is a subset of the label sequence considered for Ψ_t, a local feature function dependent on time that contains the parameters to be trained for the CRF. In our work, $\mathcal{Y} = \{0, 1, 2, 3, 4\}$, the set of possible interruptibility labels, and we used two types of feature functions—*windowed* observation feature functions and *edge* observation feature functions.

Windowed observation feature functions include a window parameter, ω, that defines the number of past and future observations to use when predicting a label at time t. These feature functions are of the form:

$$\Psi_t(Y_a, X) = exp\{\sum_{k=1}^{K} \theta_k f_k(y_t, x_{t-\omega}, x_{t-\omega+1}, ..., x_{t+\omega}\} \quad (1)$$

where y_t is the label at time t, x_i is an observation value at time $t = i$, and K is the number of feature functions, f_k; in our case K was the same as the number of attributes in the data. The parameter θ_k is a parameter that is trained using gradient descent.

Unlike windowed observation feature functions, edge observation feature functions model transitions from one interruptibility class to another. These feature functions have the form:

$$\Psi_t(Y_a, X) = exp\{\sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t)\} \quad (2)$$

where all the variables have the same meaning as they did in Equation 1 and the value of K is the number of possible transitions, 25, from one interruptibility class to another.

The feature functions were specified using the implementation of CRFs in the HCRF library[2], and we trained the parameters θ_k using the BFGS gradient descent method. Unlike with the HMMs, we did not train separate CRFs for each of the interruptibility classes; instead we trained the CRF to perform multiclass classification.

5.3 Hidden Conditional Random Fields

The HCRF [26] extends the CRF by including hidden state variables to more accurately model intra-class variation within observation data. In addition, the HCRF provides a single label for the entire sequence (see Figure 2c) and thus prevents the need for an a-priori segmentation of the observed sequence into substructures. Prior work has successfully used the HCRF for Gesture Recognition [26], and thus we consider it a good candidate for modeling interruptibility.

Mathematically, the HCRF is formulated in a similar manner to the CRF:

$$P(y|X) = \sum_H P(y, H|X) = \frac{1}{Z} \sum_H \prod_{t=1}^{T} \Psi_t(y, H, X)$$

$$Z = \sum_y \sum_H \prod_{t=1}^{T} \Psi_t(y, H, X)$$

where $H = \{h_1, h_2, ..., h_T\}$ each $h_i \in \mathcal{H}$, is a sequence of hidden states that capture the underlying structure of class y, and $\mathcal{H}$ is the set of possible hidden states. Correspondingly, $|\mathcal{H}|$ is the number of hidden states that the HCRF can use; this parameter is optimized during training.

In our work, the feature functions in Equations 1 and 2 were modified so that y_t and y_{t-1} were replaced with h_t and h_{t-1}, where h_t and h_{t-1} were the hidden states at time t and $t-1$ respectively. We also created an additional feature function to model the association of a hidden state to the interruptibility class label for a sequence. This feature function was of the form:

$$\Psi_t(y, H, X) = exp\{\sum_{k=1}^{K} \theta_k f_k(y, h_t)\} \quad (3)$$

where all the variables have the same meaning as they did in Equation 1. The value of K equals $|\mathcal{H}| \times |\mathcal{Y}|$, which is the number of hidden states per interruptibility class.

The feature functions were implemented using the HCRF library[2] and training was performed using BFGS. As with the CRF, we trained the HCRF to perform multiclass classification.

5.4 Latent-Dynamic CRFs

The LDCRF [16] offers several advantages over CRFs and HCRFs by modeling both extrinsic dynamics between interruptibility classes as well as the intrinsic substructure within an interruptibility class. It does so by using hidden states, as the HCRF, and at the same time by removing the need to label an entire sequence with a single interruptibility class label (see Figure 2d). In prior work, the LDCRF has been shown to outperform both the CRF and HCRF in Gesture Recognition [16], and therefore we consider it a good candidate for classifying interruptibility.

Mathematically, the LDCRF assumes that each sequence label y contains a corresponding set $\mathcal{H}_y$ of hidden states to capture intra-class substructures. Therefore, the LDCRF

[2] https://sourceforge.net/projects/hcrf/

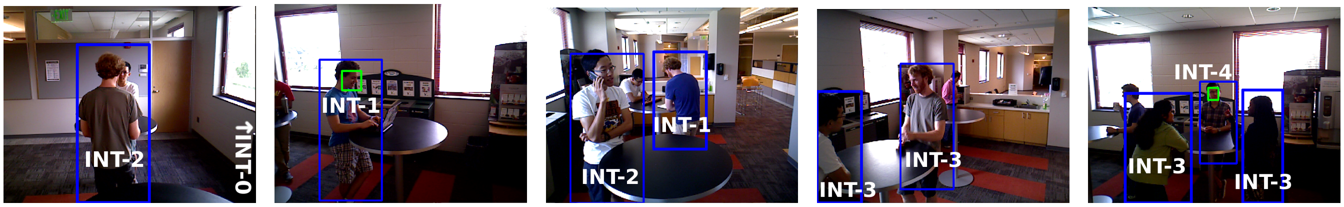


Figure 3: Example scenes from each of the five data collection runs. The green bounding box denotes a face identified by the face recognition component and the interruptibility label of individuals within the blue bounding box is shown.

evaluates the following conditional model

$$P(Y|X) = \sum_{H} P(Y|H,X)P(H|X)$$

where $H = \{h_1, h_2, ..., h_T\}$ is a sequence of hidden states and each h_i belongs to the hidden state set $\mathcal{H}_{y_i}$ of its corresponding label y_i. To keep training and inference tractable, these sets are assumed to be disjoint for each class label. With the disjoint assumption, the conditional probability evaluated by the LDCRF reduces to

$$P(Y|X) = \sum_{H:\{h_1,\ldots,h_T\}, h_i \in \mathcal{H}_{y_i}} P(H|X)$$

where $P(H|X)$ can be derived using the CRF formulation:

$$P(H|X) = \frac{1}{Z} \prod_{t=1}^{T} \Psi_t(H_a, X)$$

$$Z = \sum_{H_a} \prod_{t=1}^{T} \Psi_t(H_a, X)$$

In our work, we used the same feature functions that we had for the CRF (Equations 1 and 2), with suitable updates to the variables. The feature functions were again implemented using the HCRF library[2] and training was performed with BFGS. As with the HCRF and CRF, the LDCRF was trained to perform multiclass classification.

6. EXPERIMENTAL SETUP

To evaluate the performance of the above models on classifying interruptibility, we performed data collection to obtain videos of small groups of people in a public space. In this section, we describe the experimental setup, including data collection and annotation.

6.1 Robot Hardware and Software

The robot used in this project is a mobile platform with a holonomic base, a 6-DOF Jaco-2 arm, and is outfitted with a Hokuyo laser scanner, a Kinect One RGB-D camera, and an ASUS Xtion Pro Live RGB-D camera. The Kinect directional microphone array was used to collect audio data. Additionally, we used the STRANDS perception pipeline [4] for people tracking at approximately 10 Hz and the Sighthound Cloud API[3] for face detection and tagging at 3–4 Hz.

6.2 Data Collection and Processing

During the data collection process, five people (not co-authors on the paper) were asked to take part in everyday activities in a common area of the building. Five data collection runs were conducted, each with 3–5 participants in the scene engaged in activities such as drinking coffee, having a conversation, or working on their laptops (see Figure 3). The common area and activities were chosen because they allowed for a wide range of likely activities and a variety of visual scenes with different numbers of people and varying levels of occlusion. During each run, the robot was teleoperated through a preset series of waypoints that enabled it to observe the group from different perspectives; each run lasted an average of 108 seconds.

Following recording, the data was processed into segments that could be annotated with a person's interruptibility. Due to motion blur during navigation, only data from stationary robot observations was used. First, data from all sensor streams was segmented into 250 ms non-overlapping windows. For each sensor stream, the window of data was condensed into a single value consisting of the last recorded value for that sensor stream (if available). A Euclidean distance heuristic was then used to merge data for each detected person across all sensor streams. The result of this process was the creation of 1516 data segments, each of duration 250 ms, and each of which contained all the information available about a single person detected within the environment. Each segment was then annotated with ground truth interruptibility labels (see Section 6.3).

Post-annotation, consecutive data segments were concatenated into sequences of minimum length 4 (1 second) and maximum length 8 (2 seconds). In the event of missing data (e.g., face recognition failure), missing values were filled in through linear interpolation (continuous) or by propagating the last known value (boolean). If neither approach was available, such as in the case where the beginning segment of the sequence was missing required data, attributes were assigned a value of NaN to distinguish them from other valid values in the domain. All models were modified to ignore NaN values. This process resulted in the creation of 671 sequences that were used for model evaluation in Section 8. During evaluation, and for training HCRFs, we defined the interruptibility label of a sequence to be the interruptibility label of the last segment in the sequence.[4]

6.3 Annotation

The authors used the interruptibility scale from Section 4 to annotate each of the 250 ms data segments. Additionally, two independent coders were each asked to annotate a random subset consisting of approximately 40% of the data. To verify label consistency we calculated the Cronbach's Alpha measure of inter-rater reliability between the authors' anno-

[3] https://www.sighthound.com/products/cloud

[4] No significant difference was observed in using alternate sequence labeling methods, such as mode of all segment labels.

Feature	**Min**	**Std**	**Ext**
Body Position	×	×	×
Face Gaze	×	×	×
Body Orientation*		×	×
Audio Angle		×	×
Audio Confidence		×	×
Audio Angle Near Position*			×
Within Camera Field-of-View			×
Body Distance Thresholds			×
Linear Velocity			×
Quaternion Rate of Change*			×
Face Bounding Box*			×
Body Bounding Box*			×
Body Bounding Box Area*			×

Table 1: Membership of each feature to the different feature sets—Minimal (Min), Standard (Std), and Extended (Ext). *These features provided unreliable data either due to sensor noise or sensor unreliability.

tations and those of the other annotators, resulting in scores of 0.81 and 0.96. The level of agreement highlights not only label reliability, but also the fact that humans are generally very consistent in judging the interruptibility of others.

In addition to annotating interruptibility, the authors annotated the data with the labels of objects in the scene. The labels included *unknown, none, laptop, bottle, book, headphones, mug, phone_talk,* and *phone_text*. The label *unknown* was frequently used in conjunction with the interruptibility label 0, which was used in situations when the person-of-interest was outside the camera field-of-view but detected by the laser and audio (see leftmost example in Figure 3). Separate labels were assigned to phone use for speaking or texting (*phone_talk* and *phone_text*) because the activities correspond to different visual features and because the associated interruptibility of the person would likely also be different. In the future, this component will be replaced with automated object or activity recognition.

7. TRAINING FEATURES

Given the training data described in the previous section, we evaluated the performance of the proposed computational models using different subsets of features. As discussed in Section 4, we considered *person state* features (Section 7.1) and *interruption context* features (Section 7.2).

7.1 Person State Features

The primary interruptibility cues about a person include head orientation, body position, and audible signals. We note that the recognition of these cues by a mobile robot in a public space can be quite noisy. As a result, we structure our evaluation to consider three subsets of features—*Minimal* (Min), *Standard* (Std), and *Extended* (Ext)—which are summarized in Table 1. Our goal here is to explore the robustness of the temporal models to additional data and noise; we do not propose that this is the best set of features for *person state* in general.

Minimal Feature Set. We hypothesize that the position of a person and an indication of whether they are looking at the robot or not, are the most decisive factors when gauging interruptibility. Therefore, we use *Min* to test our model performance when rich data from other sensors (such as microphones), or from additional visual detectors (such as an upper body detectors), are unavailable. This set contains:

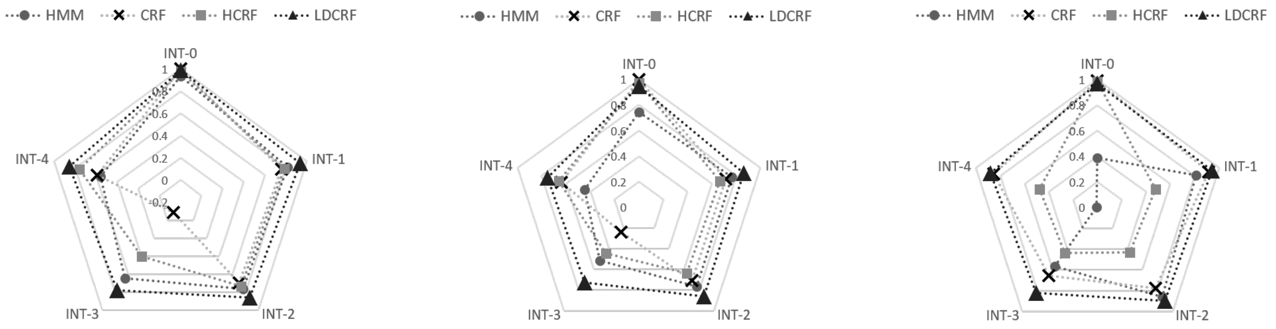
- **Body Position** Tuple, (x, y), denoting the position of the body in the environment relative to the robot base.
- **Face Gaze** Boolean, *True* when a face is detected and the head is oriented towards the robot, *False* when a face is detected but the head is not oriented towards the robot or if the eyes are shut, and NaN when no face is detected.

Standard Feature Set. This set of features represents the full breadth of information enumerated in Section 4 and is most similar to the features used in [15, 3]. In addition to *Min*, the set contains:

- **Body Orientation** Tuple, (z, w), of the quaternion, (x, y, z, w), denoting the rotation of a person's upper body relative to the robot's base frame. The (z, w) values specify rotation estimates about the upright axis and are thus the only meaningful values in the quaternion.
- **Audio Angle** Angle, in radians, to the dominant source of detected sound, calculated by the Kinect.
- **Audio Confidence** A $[0, 1]$ confidence measure for the *Audio Angle* estimate.

Extended Features Set. In the final feature set we add additional features, some of which are noisy, to study the effects of extra data on model performance. The features are either obtained from the outputs of intermediate processing steps, such as the body bounding box, which is a supplementary output of the upper body detector, or are obtained through additional post-processing of *Std*, such as the field-of-view boolean, which maps a point in (x, y) to a boolean value indicating whether the point is in the field-of-view of the camera. These features have not been used in prior works but are added with the aim of making explicit some of the decision variables that we think might be useful for interruptibility. We hypothesize that the presence of the explicit decision variables will help the models, regardless of the effects of the noise. The variables include:

- **Audio Angle Near Position** Boolean, *True* when the *Audio Angle* estimate equals the angle from the camera to a detected person (within some tolerance), *False* when this is not the case.
- **Within Camera Field-of-View** Boolean, *True* when a detected person is within the field-of-view of the camera and *False* otherwise.
- **Body Distance Thresholds** Three booleans, each *True* if a detected person is beyond the boundaries of Hall's proxemic distances [7], and *False* if not. The boundaries considered are those of Personal Distance (0.46 m), Social Distance (1.22 m), and Public Distance (3.66 m).
- **Linear Velocity** Tuple, (v_x, v_y), obtained from the rate of change in *Body Position* between data segments.
- **Quaternion Rate of Change** Tuple, (v_z, v_w), obtained from the rate of change in *Body Orientation* between data segments.
- **Face Bounding Box** Four continuous values—x, y, *width*, and *height*—for the bounding box around a detected face.
- **Body Bounding Box** Four continuous values—x, y, *width*, and *height*—for the bounding box around a detected body.



(a) Minimal Feature Set

(b) Standard Feature Set

(c) Extended Feature Set

Figure 4: Radar plots reporting MCC performance of each model as a function of the interruptibility class.

- **Body Bounding Box Area** Area of the *Body Bounding Box.*

In the CRF, HCRF, and LDCRF, continuous multivariate features, such as the *Body Position* tuple, are treated as separate vectors of univariate features. In the HMM, the features are left as multivariate because doing so provides us with the largest log likelihood values post-training. Similarly, combining the *Within Camera Field-of-View* boolean feature with the *Body Distance Thresholds* boolean features, and combining the *Audio Angle* feature with the *Audio Angle Confidence* feature, provides us with the highest log likelihood values for the HMM, and therefore these combinations are used in that model.

7.2 Interruption Context Features

In order to evaluate the use of object recognition as a means of conveying the context of a scene, we additionally define an object label feature which can be added to any of the above feature sets. The object feature is defined as a set of boolean values, each of which is *True* or *False* if the corresponding object is present or absent within the scene. Due to perfect object labels, we simulate the noise expected from automated object recognition by corrupting the boolean values in approximately 10% of the data segments.

8. RESULTS AND DISCUSSION

In this section, we present a comparison of the four temporal models in classification of interruptibility based on different feature sets (Section 8.1), and then show the impact of adding contextual data in the form of object labels (Section 8.2). In order to train the parameters for our models, we performed 10 fold cross-validation with 80% of the data in a fold used for training and 20% for testing. Results with the best performing parameters for each model (number of hidden states and number of mixtures for HMMs, and number of hidden states and the time window for *windowed* observation feature functions in the CRF, HCRF, and LDCRF) are reported using a Matthew's Correlation Coefficient (MCC) score for each interruptibility label. The score reflects a model's predictive power in a binary classification task for a given interruptibility label. Significance results are presented using a Wilcoxon signed-rank test with a one-tailed hypothesis using the MCC scores across the different folds of testing and training.

8.1 Temporal Model Analysis

Figure 4 compares the performance of the HMM, CRF, HCRF and LDCRF models across the three feature sets without the inclusion of object context data. The results for each set of features are analyzed below.

Minimal Feature Set. Figure 4a presents the classification results using *Min*. Two notable patterns can be observed in the data. First, all models perform relatively well across INT-0, INT-1, INT-2 and INT-4, with MCC scores in the range 0.55-0.99. A significant variation from this is observed in INT-3, representing the *Interruptible* data class, for which HCRFs obtain a score of 0.4 and CRFs -0.09. The low score in this category, particularly for the CRF, is due to the model's failure to distinguish INT-3 from other classes (65% of INT-3 instances were misclassified as INT-2).

Second, LDCRF consistently outperforms other models ($p < 0.002$ vs. HMM and $p < 0.0001$ vs. CRF and HCRF), with scores of $0.99, 0.93, 0.86, 0.78, 0.85$ for INT-0 through INT-4, respectively. The CRF performance, on the other hand is consistently poor by comparison, especially in the case of INT-3, due to a lack of hidden variables within the model. In particular, for both the LDCRF and the HCRF, the best performance across all the interruptibility classes is obtained with 4 hidden states, whereas the best performance with HMMs is achieved with 2–3 hidden states. These values indicate that there are subclass dynamics that the CRF failed to model because of the lack of hidden states. The HMM achieves average performance among the models, but often outperforms the HCRF, which shows an inability to accurately model interclass dynamics in the data. We suspect that this is due to a weakness of the HCRF, which applies the same interruptibility label to all segments of a sequence even if the ground truth interruptibility label for each of the segments in the sequence might be different.

Standard Feature Set. Figure 4b presents the classification results using *Std*, which includes three additional features beyond the minimal set (Body Orientation, Audio Angle and Audio Confidence). Overall, we observe the same performance pattern, with LDCRF again outperforming all other methods ($p < 0.0003$ vs. HCRF and $p < 0.0001$ vs. CRF and HMM) and with lower performance for all models in INT-3 than other labels. Notably, however, we also observe that classification performance is lower across many

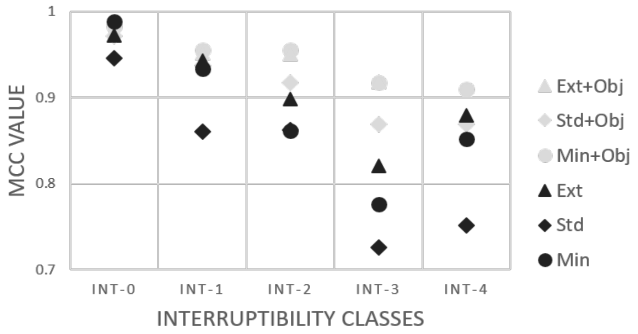


Figure 5: Effect of adding object labels to LDCRF.

interruptibility labels in comparison to *Min.* This drop in performance is due to the nature of the data encoded in the additional features used. The Body Orientation feature is particularly noisy, with orientation values in some segments deviating by 90° or more from the ground truth. The HMM model proves to be sensitive to the variability in the data, resulting in an average reduction of 0.1 in its MCC score. The HCRF also shows some sensitivity to the noise.

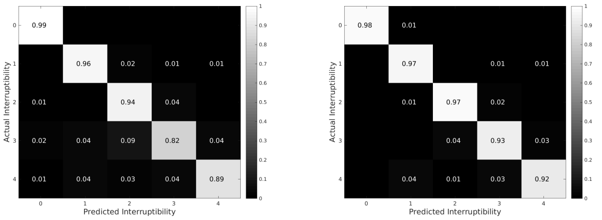
Extended Feature Set. Figure 4c presents the classification results using *Ext*, which includes eight features beyond *Std.* Several of these features contain additional information, such as the *Body Distance Threshold* booleans and the *Linear Velocity* tuple, but significant noise (see Table 1). As can be seen in the radar plot, the performance of the HMM and HCRF drops significantly, whereas performance of CRF and LDCRF is higher on average than with both *Min* ($p < 0.03$ for the LDCRF) and *Std* ($p < 0.0001$ for the LDCRF) features. This continues the trend that we had begun to observe with *Std*; the addition of more information improves the decision making ability of both the LDCRF and the CRF, indicating a higher noise tolerance in these models in comparison to the HMM and HCRF.

Summary. In summary, we found that the LDCRF model consistently outperforms all other methods across all feature sets, with the best performance achieved with *Ext.* This result is also indicative of LDCRF's robustness to noisy data, which is valuable given the expected variability in the quality of data available to a mobile robot in public spaces. Figure 6a shows the confusion matrix of the LDCRF with *Ext*, showing consistently high performance on the diagonal, with most misclassifications occurring in neighboring classes.

8.2 Object Context

In this section, we evaluate the effect of adding object recognition features to each of the three feature sets (*Minimal*, *Standard* and *Extended*) on classification performance. Given the dominant performance of LDCRFs in the previous section, we report analysis of only the LDCRF model on these datasets, although, we observe similar effects across the other three temporal models.

Figure 5 presents a comparison of LDCRF performance on the original feature sets (black) and with the addition of object labels (gray). As can be seen, the addition of object labels consistently increases the classification performance of the model across nearly all conditions ($p < 0.0001$ with *Min* and *Std* features, and $p < 0.002$ with *Ext* features). The only conditions where MCC score did not increase are *Ext+Obj* in INT-4 and *Min+Obj* in INT-0, in which there is a negligible loss of 0.001 and 0.005 in MCC scores respectively. In all other conditions we observe an increase in performance, particularly for INT-3 where the MCC score improves by as much as 0.16 points. The best overall performance is achieved by *Min+Obj* condition. A confusion matrix of the LDCRF trained on *Min* with objects is presented in Figure 6b, showing improved ($p < 0.0003$) performance over the leading LDCRF method without objects.



(a) Extended without Objects (b) Minimal with Objects

Figure 6: Confusion matrices for the LDCRF under different sets of features.

The above results support our hypothesis that contextual information derived from object labels is highly informative to interruptibility classification. Furthermore, we observe a tradeoff between the the use of a larger set of, somewhat noisy, features (*Ext*) and the use of a smaller number of more precise features[5]. Specifically, in the absence of object labels, LDCRF performance is highest with *Ext*, making the best use of the additional information, even when it is noisy. With the introduction of object labels, the extended features serve as a distraction and the best performance is achieved with *Min.* This finding is significant in guiding future development efforts in this area. Specifically, we observe that all domains, but especially ones in which it is relatively difficult to obtain reliable person tracking information, benefit from the incorporation of contextual signals. In future work we will explore the use of automated object recognition, as well as additional contextual information beyond object labels.

9. CONCLUSION

In this paper, we introduced a rating scale for characterizing interruptibility, and compared four temporal models – HMMs, CRFs, HCRFs and LDCRFs – in classifying the interruptibility of multiple people in a scene based on laser, visual and audio data collected by a mobile robot. Our findings show that LDCRFs consistently outperform other models across all conditions. Additionally, our work is the first to introduce contextual scene information beyond the person-of-interest, in this case object labels, to models of interruptibility. Our findings show that adding object labels significantly improves interruptibility classification performance, particularly when combined with reliable person descriptive features. Our approach successfully handles multiple people in a single scene, and in future work we will explore how the presented interruptibility ratings can be used by the robot to decide who to interrupt, and how.

10. ACKNOWLEDGMENTS

This work was supported by an Early Career Faculty grant from NASA's Space Technology Research Grants Program.

[5] Model performance further improved when artificial corruption of object labels was absent

REFERENCES

[1] P. D. Adamczyk and B. P. Bailey. If not now, when? In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, pages 271–278, New York, New York, USA, 2004. ACM Press.

[2] D. Bohus and E. Horvitz. Dialog in the open world. In *Proceedings of the 2009 international conference on Multimodal interfaces - ICMI-MLMI '09*, page 31, New York, New York, USA, 2009. ACM Press.

[3] Y.-S. Chiang, T.-S. Chu, C. D. Lim, T.-Y. Wu, S.-H. Tseng, and L.-C. Fu. Personalizing robot behavior for interruption in social human-robot interaction. In *2014 IEEE International Workshop on Advanced Robotics and its Social Impacts*, pages 44–49. IEEE, sep 2014.

[4] C. Dondrup, N. Bellotto, F. Jovan, and M. Hanheide. Real-time multisensor people tracking for human-robot spatial interaction. In *Workshop on Machine Learning for Social Robotics at International Conference on Robotics and Automation (ICRA)*. ICRA/IEEE, 2015.

[5] J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang. Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction*, 12(1):119–146, mar 2005.

[6] M. E. Foster, A. Gaschler, and M. Giuliani. How Can I Help You? Comparing Engagement Classification Strategies for a Robot Bartender. In *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13*, pages 255–262, New York, New York, USA, 2013. ACM Press.

[7] E. T. Hall. *The Hidden Dimension.* Anchor Books, 1969.

[8] C. E. Harriott and J. A. Adams. Modeling Human Performance for Human-Robot Systems. *Reviews of Human Factors and Ergonomics*, 9(1):94–130, nov 2013.

[9] E. Horvitz and J. Apacible. Learning and reasoning about interruption. In *Proceedings of the 5th international conference on Multimodal interfaces - ICMI '03*, page 20, New York, New York, USA, 2003. ACM Press.

[10] S. T. Iqbal and B. P. Bailey. Leveraging characteristics of task structure to predict the cost of interruption. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, page 741, New York, New York, USA, 2006. ACM Press.

[11] Y. Kato, T. Kanda, and H. Ishiguro. May I help you? In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*, pages 35–42, New York, New York, USA, 2015. ACM Press.

[12] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289, 2001.

[13] D. McFarlane and K. Latorella. The Scope and Importance of Human Interruption in Human-Computer Interaction Design. *Human-Computer Interaction*, 17(1):1–61, mar 2002.

[14] Y. Miyata and D. A. Norman. Psychological issues in support of multiple activities. *User centered system design: New perspectives on human-computer interaction*, pages 265–284, 1986.

[15] C. Mollaret, A. Mekonnen, F. Lerasle, I. Ferrané, J. Pinquier, B. Boudet, and P. Rumeau. A multi-modal perception based assistive robotic system for the elderly. *Computer Vision and Image Understanding*, mar 2016.

[16] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-Dynamic Discriminative Models for Continuous Gesture Recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, jun 2007.

[17] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[18] A. J. Rivera. A socio-technical systems approach to studying interruptions: Understanding the interrupter's perspective. *Applied Ergonomics*, 45(3):747–756, may 2014.

[19] S. Rosenthal, M. M. Veloso, and A. K. Dey. Is Someone in this Office Available to Help Me? *Journal of Intelligent & Robotic Systems*, 66(1-2):205–221, apr 2012.

[20] N. Sarter. Multimodal Support for Interruption Management: Models, Empirical Findings, and Design Recommendations. *Proceedings of the IEEE*, 101(9):2105–2112, sep 2013.

[21] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita. How to approach humans? In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction - HRI '09*, page 109, New York, New York, USA, 2009. ACM Press.

[22] P. Saulnier, E. Sharlin, and S. Greenberg. Exploring minimal nonverbal interruption in HRI. In *2011 RO-MAN*, pages 79–86. IEEE, jul 2011.

[23] C. Shi, M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita. Measuring Communication Participation to Initiate Conversation in HumanâĂŞRobot Interaction. *International Journal of Social Robotics*, 7(5):889–910, nov 2015.

[24] C. Speier, J. S. Valacich, and I. Vessey. The effects of task interruption and information presentation on individual decision making. In *Proceedings of the eighteenth international conference on Information systems*, pages 21–36. Association for Information Systems, 1997.

[25] H. Stern, V. Pammer, and S. N. Lindstaedt. A preliminary study on interruptibility detection based on location and calendar information. *Proc. CoSDEO*, 11, 2011.

[26] Sy Bor Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell. Hidden Conditional Random Fields for Gesture Recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, volume 2, pages 1521–1527. IEEE, 2006.

[27] E. R. Sykes. A Cloud-based Interaction Management System Architecture for Mobile Devices. *Procedia Computer Science*, 34:625–632, 2014.

[28] L. D. Turner, S. M. Allen, and R. M. Whitaker. Interruptibility prediction for ubiquitous systems. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*, pages 801–812, New York, New York, USA, 2015. ACM Press.

[29] D. L. Vail, M. M. Veloso, and J. D. Lafferty. Conditional random fields for activity recognition. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems - AAMAS '07*, page 1, New York, New York, USA, 2007. ACM Press.