Modeling Human Need for Attention and Interruptibility in Restaurant Scenarios

Ada Taylor , Roman Kaufman , Siddharth Girdhar , Henny Admoni
Carnegie Mellon University
{adat, rkaufman, sgirdhar, hadmoni}@andrew.cmu.edu

Abstract

The goal of this work is to train a model to quantify mental states such as neediness and interruptibility from human action patterns in restaurant scenarios. Our long-term vision is to develop robot waiters that can intelligently respond to customers. Our key insight is that human behaviors can both actively and passively communicate customers' underlying mental states. To interpret behaviors indicating neediness and interruptibility, we automatically label key moments of human service patterns in restaurants based on waiter location and objects, as well as human behavior patterns in terms of pose, facial expression, and facial action units. Our effort to build a model is complicated by a lack of ground truth information, unreliability in waiter actions, and the effect of distractions and non-service social interactions on customer signals, and we propose solutions to each of these issues. We plan to compare the performance of several machine learning methods in predicting moments when waiters attend to customer needs based on this model.

1 Introduction

For robots to be able to autonomously aid humans in tasks, it is key that they be able to identify when humans need help, and when that help can be best administered. We specifically seek to model and test these skills in a restaurant setting, with the goal of providing information that will enable robots to act as waiters and proactively address customer needs.

In this work we define two human mental states that robots providing aid should act on: neediness and interruptibility. Neediness is the potential impact of receiving aid at certain point in time on customer experience. Notably, kinds of impact include objective aid such as delivering food, intangible aid such as receiving information on what a customer wants to order, as well as emotional reactions such as a growing impatience over time that can provide impact by being resolved. We define interruptibility as a numerical descriptor of the receptiveness of a human to a waiter providing help at a given moment. These models will allow robotic waitstaff to determine which customers need waiter help the most, and then plan the best time to deliver that help.

Overall, needs can be divided into two categories: systematic and reactive. Systematic needs include those that a completely naive robot with no emotional insight could plan on attempting to address. This includes many of the standard elements of a restaurant service outlined in Figure 1 such as asking for orders, delivering food, and cleaning up.

Reactive needs arise over the course of a meal, and are tied to customer emotional state. These include reacting to a customer with questions about the menu, or is growing impatient. These can arise at any time, and are signaled most directly by customer behaviors. Notably, reactive needs are often tied to systematic causes that can be addressed.

The goal of our metric of neediness is to incorporate both of these aspects of need. This will allow our system the benefit of being able to recognize reactive cues that indicate rising systematic problems that can be addressed, while also not neglecting customers that happen to be less expressive.

Interruptibility seeks to describe when that help should be administered, and is a measurement of whether it would be appropriate to interrupt a customer's current task. It is most closely aligned with the prior work of [Banerjee *et al.*, 2018], which classified users into categories of interruptibility.

Being able to assess and predict these metrics of neediness and interruptibility will enable more sophisticated collaborative and managerial behaviors such as deciding which group of people to prioritize, assessing overall group satisfaction over time, and scheduling future interactions so as to minimize overall neediness or wasted effort.

The models we build are based entirely on data on restaurant interactions found in the wild. This gives us a very realistic substrate for modeling these human interactions, with examples of natural customer reactions and interactions, as well as data on the observed decisions of waiters.

Notably, such a system could provide obvious value to instances where a robot is working as waitstaff, and it could also provide useful managerial skills in a human-only or blended human-robot environment. For example, the ability to gauge the levels of human neediness can allow the system to mark understaffed areas of the restaurant, alert staff to individuals who have been overlooked or have an urgent issue, and schedule tasks or food preparation in anticipation of future needs.

The ability for robots to initiate interactions to administer help is particularly valuable in high-distraction scenarios with multiple groups of humans who have competing needs. These

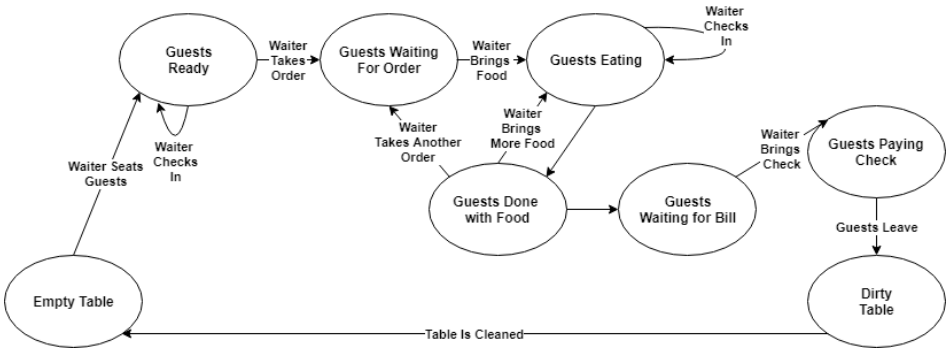


Figure 1: Flowchart describing the systematic needs of a table during the course of a restaurant experience

situations also inherently contain many additional patterns of human behavior that make isolating only signals of neediness or interruptibility difficult. However, due to the structured nature of competing group conversations or background distractions, it may be possible to identify and extract recurring configurations of distraction. In order to meet these goals, we seek to address several challenges:

1. Finding the Ground Truth: Translating video streams of real-world restaurant interactions into feature and reward signals we can use for training.
2. Noisy Reward Signal in Waiter Actions: Accounting for the imperfect nature of using waiters as training signals.
3. Structured Noise in Customer Signals: Identifying recurring patterns in customer attention and behavior that obscure signals that contribute to our target metrics.

2 Background

In commercial restaurant settings, robots have primarily been deployed to assist waiters by carrying food on a static tray [Pieska *et al.*, 2013]. These robots do not yet independently perform customer interaction tasks except in very specialized scenarios. Many are even limited to remaining on tracks.

However, robots with a focus on social awareness have been deployed in a bartending context. JAMES, created by the University of Edinburgh, is a robot built to demonstrate the value of being able to interpret and respond to human social cues in a bar context [Foster *et al.*, 2012]. This setting emphasized the value of this genre of tasks, however, it required a fixed-location robot and depends on humans directly approaching and soliciting the bartender. Our system aims to detect growing problems before a customer feels the need to actively ask for help.

Our work also builds on the work of [Banerjee *et al.*, 2018]. This work highlighted the usefulness of object labels as visual indicators of human interruptibility, and inspired our focus on examining whether this kind of feature can be used for automatically labeling key moments or states in human interactions, or as features for classifying both neediness and interruptibility. Our model seeks to add additional meta-object features by combining information about multiple objects in the scene, as well as incorporating human features such as pose with object information to determine which objects a human is focused on within a multi-object scene.

3 Challenges

3.1 Finding the Ground Truth

A strength of our approach is the use of data taken from "in the wild" because it is captured in a natural, non-lab environment. Our data comes from publicly available video streams from inside restaurants, published by those establishments in order to increase tourism and help potential visitors assess traffic. Specifically, our data so far as been collectd from Myrtle Beach's Ridtydz beachside bar [Bar,] public-access webstream. These streams consist of 2D RGB video from a single high viewpoint, with no audio. While the presence of cameras often does influence human behavior [Jansen *et al.*, 2018], there should not be changes in behavior due to the novelty of the camera, because the web stream we are working with is longstanding.

The choice to use unannotated "in-the-wild" streams with no method for interacting with the scene or asking people in the scene for additional information complicates the process of directly determining a "ground truth" for the neediness and interruptibility of a given customer or table. Manual labeling will be challenging, due to the fact that it is difficult for the labeler to understand the complete underlying context. Instead we plan to gather sufficiently large sets of real-world data to learn the strategies being used by waiters from scratch, and break down relevant customer behaviors granularly enough to capture the relevant signals. Both of these constraints lead to our decision to focus on automated methods of extracting and labeling features.

To gather features from these 2D RGB video streams, we are using OpenFace [Baltrusaitis *et al.*, 2018] and OpenPose [Cao *et al.*, 2018] to collect information on each customer's head position, torso position, and facial action units to look for indicators of reactive neediness, as well as interruptibility. An example of this kind of feature can be found in Figure 2.

We also use TensorFlow's Object Detection API [Huang *et al.*, 2017] trained on the Common Objects in Context (COCO) dataset [Lin *et al.*, 2014] to identify the type and location of key objects in dining scenes. Object and person tracking information can be used as features for classification, as seen in Figures 3 and 4, as well as to provide automatic labels for our data. By tracking the location of the waiter in the scene as compared to tables in the scene, as well as the appearance or disappearance of objects from the table, we la-

Figure 2: OpenPose on restaurant footage

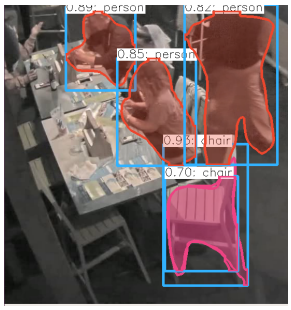


Figure 3: TensorFlow and COCO object recognition on restaurant footage

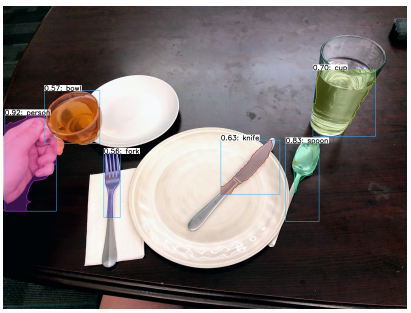


Figure 4: Object detection on restaurant objects

bel systematic needs that are being addressed by waiters over the course of service. For example, if menus disappear from the table after a waiter interaction, we assume that the waiter took order information from customers.

By automating the generation of these labels, we can collect and model a larger corpus of data, and it may be able to better capture overall service patterns than human labeling of short segments could.

3.2 Noisy Reward Signal in Waiter Actions

In an ideal world, the waiter would never miss indicators of need, and would address them immediately. Unfortunately, real world restaurants are more complex than this. Waiters can be distracted, miss cues due to occlusion, or have other tasks that consume their time and attention.

However, the waiter does remain a valuable signal for customer neediness and interruptibility. To make this problem tractable, we make the following assumptions about the moment a waiter visits a particular table:

1. The neediness of the table was monotonically increasing before the waiter's arrival.
2. The neediness of the table drops after the waiter visits it.
3. The table the waiter is visiting has the highest neediness of all tables within the scene.
4. The table was sufficiently interruptible for a waiter visit.

Using these assumptions, we can still glean information from moments that the waiter approaches the table that can be used as a reward signal for training a model.

Furthermore, as noted in Section 3.1, by incorporating object recognition we can more granularly classify the category of task that the waiter might have performed at a table by identifying objects they brought or removed it. This can enable us to identify systematic needs being addressed by the delivery of objects or taking of orders, as well as potential addressing of reactive needs in the scenario that a waiter makes no object changes to the table.

3.3 Structured Noise in Customer Signals

While food is a major reason that customers visit a restaurant, distractions such as company and entertainment are also crucial aspects of the experience that affect their behavior. This "structured noise" has patterns that we can both interpret and extract from our signals of neediness and interruptibility. For example, inanimate elements such as televisions, windows, and cell phones consume attention and can obscure passive signals. Group conversations are more complex to recognize and generalize over than interactions with static objects, but they are particularly relevant to determining interruptibility.

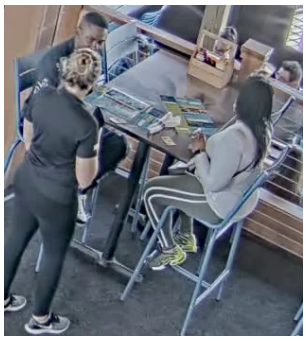
This is a unique challenge to restaurant scenarios compared to a one-to-one interaction with a shorter timespan, such as ordering at a bar. The static position of chairs at the table means that techniques used for studies of humans interacting in standing F-formations [Cristani *et al.*, 2011] are less directly applicable. Humans are distractable [Stewart and Chase, 1999], and it is unreasonable to assume that their only focus is generating signals to cue their waiters. In fact, this would defeat the goal of requiring less manual demand for service from customers.

There are two kinds of distractions we hope to capture: social and non-social. Non-social distractions can be detected with the feature-detection methods outlined in Section 3.1, such as checking for common distraction objects such as a cell phone, an approach similar to [Banerjee *et al.*, 2018]. Other objects of focus can potentially be detected using a combination of ray-tracing customer gaze based on head pose and checking for intersection with detected objects, as in the case of a wall-mounted television or a window view.

Recurring patterns of social interaction are more difficult to characterize, in part because of the 3D nature of human interactions. Even given the locked positions of table chairs, humans can sit in several positions that are symmetrical from the perspective of their social interaction with other table members, but not from the point of view of our single-angle camera. For example, there are 4 homographic seatings of 3 customers at a four-person table with respect to their intra-table interactions. Therefore, our feature space for group interactions needs to account for recurring configurations in 3D space, which do by translating 2D OpenPose data into 3D pose information using the work of [Martinez *et al.*, 2017].

4 Methods

In addition to defining our problem, we have so far successfully extracted features from our livestreams in the wild, and developed mechanisms for detecting when customers visit a table, and when a waiter interacts with a given table. We have



(a) Waiter visiting a table

(b) Distraction: person on phone

Figure 5: Challenges in interpreting in-the-wild footage

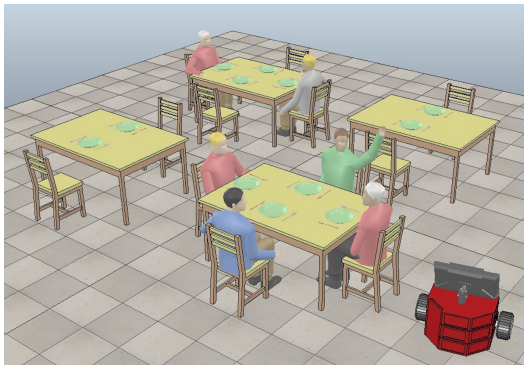


Figure 6: Dining scene in simulation environment

also created a simulation and data synthesis environment for creating and displaying example scenarios to test future models. Our next steps are to consult with waiters who can provide expert insight into our problem, as well as to train and test several machine learning models of neediness and interruptibility.

4.1 Feature Extraction

We have built a pipeline for downloading livestreams of restaurant 2D RGB video and extracting the features mentioned in Section 3.1. We automatically labeling the window that customers are at a table by tracking the number of moving objects within the pixel bounds of a table space, and track the waiter location and time of visits by noting recognized people who do not remain at a particular chair. We are in the process of combining these features to label systematic needs from more granular object analysis, as well as attempting to label non-animate objects of focus from those features.

4.2 Waiter Interviews

To attempt to shed additional light on the mechanisms of human assessments of neediness and interruptibility as related to service in restaurants, we plan to interview several waiters to learn how they perform these tasks and incorporate them into planning. We are specifically most interested in understanding the inherent model that human waiters already apply to this task, as well as key features that they rely on.

4.3 Data Synthesis and Simulation Environment

For testing purposes, we have developed a simulation environment that generates 3D data and animations using V-REP [Rohmer *et al.*, 2013]. This systems accepts either restaurant footage or short descriptions of sequences of common events, and generates visualizations as demonstrated in Figure 6. It is also able to generate 3D representations of customer poses based on the 2D RGB footage as based on the work of [Martinez *et al.*, 2017].

Future work will expand the palette of script options, and add the ability to generate generalized versions of data that vary parameters such as exact customer dimensions and group placement at a table to investigate the importance of these features and attempt to prevent over-fitting based on these features. This environment will also be used for future testing of planning and scheduling algorithms that depend on integration with an actual robot waiter.

4.4 Model Creation and Validation

Given our inputs of automatically labeled interactions, indicators of systematic and reactive needs, and our assumptions of waiter interaction, we plan to deploy several machine learning models to attempt to create metrics of neediness and interruptibility, then assess their effectiveness predicting future waiter visits. Proposed models include:

1. Hidden Markov Models (HMMs), which are useful for modeling hidden states that generate observable signals, but have low interpretability and might provide insights too coarse for our purposes. One way to address this is to develop a different HMM for each individual need.
2. Long Short-Term Memory networks, which excel at modeling features that include time delays, but require large quantities of data to be effective, which may be beyond even our automatedly labeled corpus.
3. Conditional Random Fields, which excel at predicting structured data, therefore could account for the underlying flow of systematic needs outlined in Figure 1. This technique has also seen previous success in predicting interruptibility in the work of [Banerjee *et al.*, 2018].

5 Contributions

We have outlined an interesting and valuable problem required for robots to effectively take initiative in providing aid in a restaurant scenario, and defined the metrics of neediness and interruptibility in this space. We have divided this problem into the subsets of systematic and reactive neediness, and identified ways to link this definition to real world video of restaurant interactions, as well as surmounting issues of a weak reward signal and how to handle obfuscating distractions in this data. We plan to deploy several models on the features we have curated from our real-world video streams to attempt to predict future waiter actions, with the goal of being able use these models to inform planning and scheduling for a robot waiter.

References

[Baltrusaitis *et al.*, 2018] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th Institute of Electrical and Electronics Engineers International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. Institute of Electrical and Electronics Engineers, 2018.

[Banerjee *et al.*, 2018] Siddhartha Banerjee, Andrew Silva, and Sonia Chernova. Robot classification of human interruptibility and a study of its effects. *Association for Computing Machinery Transactions on Human-Robot Interaction (THRI)*, 7(2):14, 2018.

[Bar,] Riptydz Bar. Riptydz bar webcam.

[Cao *et al.*, 2018] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[Cristani *et al.*, 2011] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of f-formations. In *British Machine Vision Conference*, volume 2, page 4, 2011.

[Foster *et al.*, 2012] Mary Ellen Foster, Andre Gaschler, Manuel Giuliani, Amy Isard, Maria Pateraki, and Ronald Petrick. Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 3–10. Association for Computing Machinery, 2012.

[Huang *et al.*, 2017] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017.

[Jansen *et al.*, 2018] Anja Martine Jansen, Ellen Giebels, Thomas JL van Rompay, and Marianne Junger. The influence of the presentation of camera surveillance on undesired and pro-social behavior. *Frontiers in psychology*, 9:1937, 2018.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Martinez *et al.*, 2017] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *International Conference on Computer Vision*, 2017.

[Pieska *et al.*, 2013] Sakari Pieska, Mika Luimula, Juhana Jauhiainen, and Van Spiz. Social service robots in wellness and restaurant applications. *Journal of Communication and Computer*, 10(1):116–123, 2013.

[Rohmer *et al.*, 2013] Eric Rohmer, Surya PN Singh, and Marc Freese. V-rep: A versatile and scalable robot simulation framework. In *2013 Institute of Electrical and Electronics/Robotics Society of Japan International Conference on Intelligent Robots and Systems*, pages 1321–1326. Institute of Electrical and Electronics, 2013.

[Stewart and Chase, 1999] Douglas M Stewart and Richard B Chase. The impact of human error on delivering service quality. *Production and Operations Management*, 8(3):240–263, 1999.