Identification Through Interaction

AUTHORS [JSS: *TODO: LIST AUTHORS*]

I. Introduction

Why do we care? Why is it hard? Why can we do it?

Recognizing individuals is critical to forming strong human-agent relationships. A person benefits from being recognized in several ways. An agent can offer a recognized person customized behaviors based on previous encounters to facilitate the interaction. Recognition breeds familiarity with people. Most interestingly, if the agent can recognize the person through a modality that is not readily transferrable to another device, the person may experience the benefits of recognition without the privacy and security drawbacks inherent in a biometric fingerprint.

Modern person identification is accomplished through face or voice recognition, but these modalities are sensitive to environmental conditions and may have stringent set-up requirements. Not every interactive device is able to meet these requirements – for example a vacuum robot is too low to the ground to obtain a clear shot of a person's face, a tabletop robot needs to be responsive to vocal commands even when the person is turned away, and an exercise watch may be used on a loud, chaotic city street.

Even if it was possible to use voice and face recognition in all contexts it would not necessarily be desirable. A facial or vocal fingerprint can be easily transferred between devices that have cameras and microphones. Privacy conscious individuals are aware that the security of their identifying data is entirely in the hands of the companies supplying their technology.

Identification schemes based around behavior can offer a solution to this privacy concern by making the identification contingent on the interaction between the agent and the person. The fingerprint left by an interaction can not be recreated by a passive device as it requires action and reaction between two parties. It may be the case that the *interaction fingerprint* generated by a specific instance of an agent is usable only by an agent of the same type, configured in the same way. This kind of identification would be harder to share and so would protect the privacy concerns of the user without sacrificing the benefits of human-agent identification.

In addition, interaction fingerprinting generalizes across many types of interactive systems. It sits atop the existing infrastructure between the person and the agent, and could be incorporated directly into the user-interface.

In this paper, we test whether modality-agnostic features such as average signal activity and oscillation can be used for person identification. This lays the groundwork for an alternate identification system that can be used with any socially interactive device by utilizing the interaction itself to recognize individuals. For example, we use the embedding of the behavior in time as a way to represent interactions without requiring prior knowledge about the meaning of each signal stream, a change from prior work in open-set recognition which has only used static features of the user. Prior work in open-set recognition has only used temporally static features of the user, but a key feature of an interaction is that it includes behaviors that are embedded in time. In this work we present a novel method for doing person recognition on time-series data with support vector machines, and discuss how this work can be extended to address the problem of recognition through interaction as well as open-set recognition. We evaluate our system on the AMI corpus [8] and discuss how it could be extended in a case where the agent is able to take action and elicit behavioral responses from a user.

II. Background Related Work

All person recognition systems extract features from the target modality to run through a classifier. Face recognition systems extract computable features through classical computer vision [16] or with modern neural networks [11] to learn classification models for known individuals. Similarly, voice recognition systems generate fingerprints for users by extracting features based on different auditory domains like time and frequency[2]. Please see Tolba et al. [15] and Beigi et al. [3] for literature reviews of face and voice recognition, respectively.

Alternate identification schemes, like gait detection, require the user's full body to be recorded for multiple striding frames [9][5]. Gait detection is an effective means of closed set person recognition [14], but infeasible for many pieces of interactive technology because of the perspective on the user required to observe their gait.

In addition to hard biometrics like face, voice, and gait features, soft biometrics can improve identification rates. Jain et al. [7] showed how a combination of soft biometrics like height, weight, and age could be used to augment a primary identification system for a closed set of individuals. Multi-modal systems using soft biometrics have been applied to open set identification problems as well with mixed results. [6][10]

A multi-modal approach can be applied to behavior as well. Perez et al. showed that social media metadata, which can be considered to be a behavioral signature, was effective at accurately identifying individuals in closed groups of up to 10,000 people [12]. Their accuracy was due in part to the large number of raw features available – the meta data features could be used without processing – but this work shows that high degrees of accuracy are possible with out of the box classifiers if an identifying set of behavioral features can be extracted.

We looked to open-set classification theory for hints on identifying people through their behavioral patterns since we prefer to avoid a standard behavioral fingerprint learning routine of the kind used for face and voice identification. Scheirer et al. uses SVMs augmented with a measure of open-space risk to determine how a new sample should be classified[13]. Based on Scheier et al. and Bendale et al's promising results with SVMs on open-set problems [4][13], we used a multi-class SVM trained on multiple labelled snippets of interaction. In this work, we focus on distinguishing between four known participants in an ongoing interaction, leaving the open-set extension of this to future work.

III. PROBLEM FORMALIZATION

We identify a set of users through their *interaction fingerprint*. A person produces measurable signals when interacting with another human or agent. These signals may come in several forms and are detectable through different sensor modalities – a person might frequently nod in response to ideas from their colleague or be prone to verbosity in their explanations. We can identify individuals by extracting features from their temporal behaviors, which we refer to as their *interaction fingerprint*. A person may persist behaviors between interaction partners, and may exhibit alternate behaviors with specific partners over the course of an interaction.

As interactions are embedded in time, we examine whether an informative interaction fingerprint can be taken from behavior durations. In addition we look at the identification accuracy and precision of feature set describing the activity, variability, and concurrency of a set of behavioral signals. We identify a set of users $u_n \subseteq U$, where the number of users in the set, n, are unknown. Each user exhibits interactive behaviors that are measurable as k time-series signals.

$$b^k \subseteq B_n \tag{1}$$

$$[b_{t=0}^k...b_{t=n}^k] \in b^k \tag{2}$$

Where b^k is the time-series interactive signal, and b_t^k indicates whether the interactive behavior is present at time-step t. In order to convert the time-series interactive behavior into an SVM classifiable signal, a discrete set of i features are extracted from b^k to form d^k.

$$[d_0^k...d_i^k] \in d^k \tag{3}$$

Where each entry in d^k is an extracted feature that describes an aspect of the time-series signal b^k. We can then feed the extracted features into n one-against-all Support Vector Machines which find the hyperplane that most separates the target class from the rest of the data by finding w and b in the following equation. [1]

$$D(x) = w^t x + b \tag{4}$$

Where w is a $k*i$ (the number of time-series signals times the number of features extracted from each signal) dimensional

vector and b is a scalar. These values are found by training on a labelled subset of the data.

IV. METHODOLOGY

We take time-series data from recorded, annotated meetings, cut it into multiple training and test samples, perform feature extraction on each sample, and train a binary SVM classifier for each pair of users. We hypothesized that we would extract interaction fingerprint descriptive enough to accurately classify each person.

A. Data Collection - AMI Corpus

We used the AMI Meeting Corpus [8] to test the efficacy of our classification scheme. The AMI corpus is multi-modal dataset consisting of video and audio recordings of mock and real meetings. The AMI corpus also contains automatic and manual annotations taken over the raw recordings. Participants are assigned roles that determine their dialogue and activity within the meetings.

We evaluated our system on meetings which had been manually annotated for three modalities, leg movement, head movement, and dialogue transcripts. Each datum is associated with a timespan and may be tagged as a type when relevant (e.g. a head movement may be tagged as a 'shake', 'nod', or untagged)

B. Feature Engineering

Interaction data is time-series by its nature, so in order to classify we need to extract features that represent snippets of time as single values.

To build a system that generalizes across interaction types, the signals making up B_n are processed without taking their contextual meaning into account. Each signal in B_n occurs over a variable time period. In a given time-step of the interaction a signal will only be either present or absent. Using this information we can classify with several features:

1) Active Ratio - For what percentage of the snippet was the signal active (on versus off).
2) Average Oscillations - How many times did the signal toggle from on to off (normalized by snippet duration).
3) Signal Concurrency - For what percentage of the snippet are multiple signals active at the same time.

The Active Ratio and Average Oscillation features produce one value for each signal in B_n and were chosen because of their simplicity and straightforward meaning. Signal Concurrency attempts to capture a relationship between signals based on the intuition that different individuals may coordinate behaviors between modalities, for example a person who consistently vocalizes an "mmhmm" while nodding.

C. Experimental Design

We selected eight different meetings from the AMI corpus that had manually annotated leg movements, head movements, and dialogue for each participant. Each meeting contained four unique individuals, no individual participated in multiple meetings. The annotations were extracted into data streams

for each participants and split into ten, equal length segments. 75% of the data was designated as training data and paired with the corresponding participant label while the remaining 25% was designated as a testing set. A scikit-learn SVM SVC one-vs-one classifier was then trained and used to classify the test set. Each participant was classified within their own meeting, and not between meetings. We performed 20-fold validation in each case. We examined how this standard SVM classifier performed with different feature sets and for different interaction sample durations.

We performed an ablation study for the hand-crafted features in section IV-B. We left one feature out for each trial we ran to determine which feature had the greatest impact. We were curious whether a minimum duration of interaction was needed for accurate classifications, so each pair of features was tested for a range of sample durations between 1 and 100 seconds. The sample duration that gave the highest accuracy was used to produce an average f-score for each feature set.

V. RESULTS

We found low variation between interaction sample durations. There was an increase in accuracy around the 60 second mark, but the change was not significant. We ran the the ablation f-score study using 60 seconds becuase it was nominally the best and because one-minute of interaction is a feasible amount of time to interact with a person before recognizing them.

Fig. 1. Feature set vs. Average F score for all people in eight different meetings. Each person was classified within their meeting of four.
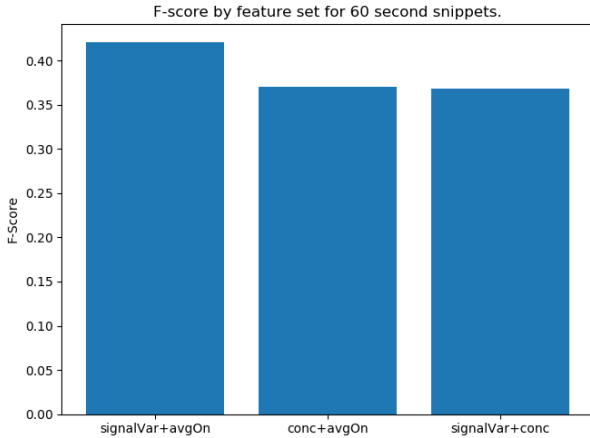

Our ablation study (figure 1) showed that Signal Concurrency was the least helpful of our three features. Our measures of signal activity and oscillation resulted in 10% increase in f-score over a feature set using concurrency.

VI. DISCUSSION

Our results show that the simplest features, activity and oscillation, which were taken from independent streams, performed better than Signal Concurrency, which attempted to

Fig. 2. Average classification accuracy by snippet duration.
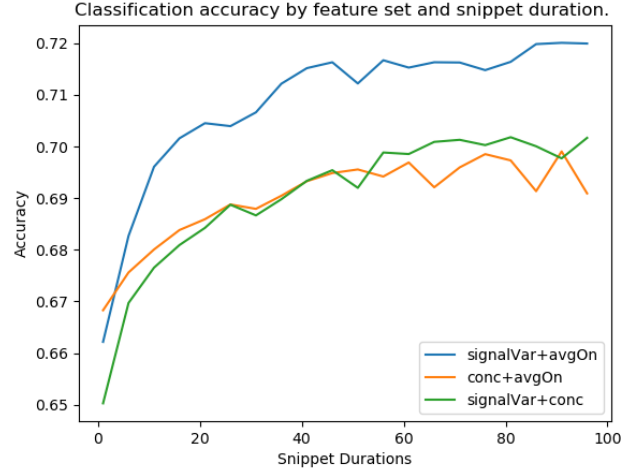
connect activity between signal streams. However, it seems reasonable that better accuracy and precision can be obtained by correlating activity between behavioral signals, but our naive approach to describing this correlation did more harm than good. This may indicate that automatic time-series feature generation, of the sort Recurrent Neural Networks are good for, could result in better f-scores. It is also worth noting that the selected features were intentionally kept generic with no implicit understanding of the behavior they described, it may be reasonable to craft behavior specific features for a more specialized identification system.

As figure 2 shows, we found that there was generally no meaningful difference in identification accuracy between different snippet durations above a threshold. A useful interaction fingerprint can be generated from our engineered features as long as a sufficient amount of the interaction is observed. This result does not necessarily hold for every kind of feature, its possible that individuals demonstrate identifying behaviors on the scale of minutes or entire interactions, e.g. the way a person leaves an interaction could produce the fingerprint that solidifies their identification.

It's important to point out that our relatively high accuracy could be due to the AMI corpus' use of role assignments within meetings. Each person in a meeting was given a role that in part determined their behavior. We did not control for these role assignments in our study, and the role assignments may serve as a confounding variable.

VII. FUTURE WORK

The generic features we chose to classify upon were insufficient to give consistently accurate identifications. However, our negative results hint at promising research directions to pursue towards HRI identification schemes. The passive signals we used may still be useful, but we can gain additional information by using the agency inherent in our HRI systems. An agent can prompt a user for behavioral signatures rather

than wait for the user to exhibit identifying behavior – even negative responses to prompts may be distinguishing.

In addition, identifying users through elicited behaviors promotes a unique one-agent-one-human relationship. An agent's ability to trigger a particular behavior may be dependent on the parameterization or even specific hardware of the agent, making it difficult to transfer the learned identification to another agent and promoting the privacy of the user. This method may also work with the reasonable person's expectation that only agents which the person has interacted with will remember them. Finally, we plan to run a study on this human-agent identification.

References

[1] Shigeo Abe. Analysis of multiclass support vector machines. *Thyroid*, 21(3):3772, 2003.

[2] Fernando Alonso-Martín, Arnaud Ramey, and Miguel Angel Salichs. Speaker identification using three signal voice domains during human-robot interaction:[late breaking reports]. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 114–115. IEEE, 2014.

[3] Homayoon Beigi. Speaker recognition. In *Fundamentals of Speaker Recognition*, pages 543–559. Springer, 2011.

[4] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015.

[5] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, February 2006.

[6] Bahar Irfan, Natalia Lyubova, Michael Garcia Ortiz, and Tony Belpaeme. Multi-modal open-set person identification in hri. In *2018 HRI Workshop*, 2018.

[7] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. Soft biometric traits for personal recognition systems. In *International conference on biometric authentication*, pages 731–738. Springer, 2004.

[8] Jonathan Kilgour. The AMI Meeting Corpus. Publication Title: AMI Corpus.

[9] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, December 2003.

[10] Eric Martinson, Wallace Lawson, and J Gregory Trafton. Identifying people with soft-biometrics at fleet week. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 49–56. IEEE, 2013.

[11] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[12] Beatrice Perez, Mirco Musolesi, and Gianluca Stringhini. You are your metadata: Identification and obfuscation of social media users using metadata information. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[13] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, July 2013.

[14] Yonghong Tian, Lan Wei, Shijian Lu, and Tiejun Huang. Free-view gait recognition. *PloS one*, 14(4), 2019.

[15] AS Tolba, AH El-Baz, and AA El-Harby. Face recognition: A literature review. *International Journal of Signal Processing*, 2(2):88–103, 2006.

[16] Matthew Turk and Alex Pentland. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition*, pages 586–587, 1991.