Identification Through Interaction

NO AUTHORS

I. Introduction

[ESS: *you need to include matthias as an author, even if he's not last author, and we need to run the paper by him before submitting*] Why do we care? Why is it hard? Why can we do it?

Recognizing individuals is critical to forming strong relationships. Person identification is primarily accomplished with face or voice recognition, but these modalities are sensitive to environmental conditions that make them unreliable in many industry applications. Some applications, like Amazon's Smart Speaker, can get around this with hardware design and big-data enabled machine learning, but this is not possible for many organizations and may not be a desired feature from the user's perspective. Robust voice and face recognition allow a user to be tracked across multiple contexts, but identification through interaction isn't likely to be reusable in the same way, giving the user the benefits of personalized interaction without further compromising their right to privacy.

Some state of the art identification schemes require resource intensive neural networks for face and voice detection. Others, like gait detection, require the user's full body to be recorded for multiple striding frames [6][5]. It's unreasonable to expect every piece of interactive technology to implement robust person identification systems given their requirements. However, most embodied piece of interactive technology can benefit from remembering previous interactions with individual users, and so a more reliable, scalable person recognition scheme is needed.

This paper implements the groundwork for an alternate identification system that can be used with any socially interactive device by utilizing the interaction itself to better recognize individuals. Prior work in open-set recognition has only used 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 a preexisting dataset 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 [8] and voice recognition [3] put features through neural networks, template matching, and other other techniques to identify known individuals. The same idea can be applied to behavior, 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. 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 also looked to open-set classification theory for hints on identifying people through their behavioral patterns. Scheirer et al. uses SVMs augmented with a measure of open-space risk to determine how a new sample should be classified[7]. Based on Scheier et al. and Bendale et al's promising results with SVMs on open-set problems [4][7], we decided to see how far we could get with a simple multi-class SVM trained on multiple labelled snippets of interaction. We chose to ignore the possibility of unlabelled classes to begin with, and instead explored whether we could distinguish between four known participants in an ongoing interaction.

A. 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 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 classifying 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 a classifiable fingerprint can be taken from different durations of the interactions. In addition we look at the identification accuracy and precision of different extracted temporal features. 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.

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

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

Where b^k is the time-series interactive signal, and b^k_t indicates whether the interactive behavior is present at time-step t. We need to take the time-series interactive behavior and

convert it into an SVM classifiable signal, so 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 Machine 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. [2]

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

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.

III. 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 [1] 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. It's worth noting that since role assignment determines behavior within a meeting, it may serve as a confounding variable.

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. An alternate approach might be to use Recurrent Neural Networks to classify the time-series data without modification, but we leave that as future work.

We're interesting in building a system that generalizes across interaction types, so 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.

[JSS: *Explain the rationale for selecting these features more?*]

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 III-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.

IV. Results

[JSS: *TODO: Fill in specific values once exhaustive snippet duration experiment is done (literally takes forever...figuratively)*] 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 seems like a reasonable amount of time to interact with a person in future experiments.

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. [JSS: *Should I collect stats to get a p-value or is that not necessary here?*]

V. Discussion

The results for our ablation study show why feature engineering is rarely effective. Our simplest features, Activity and Oscillation, which were taken from independent streams, performed better than Signal Concurrency, which attempted to connect activity between signal streams. It seems reasonable

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

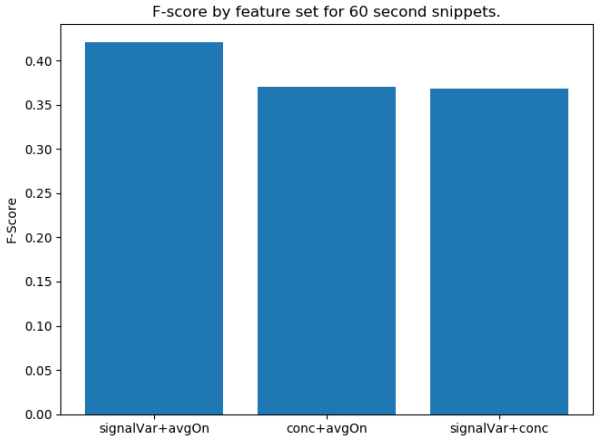
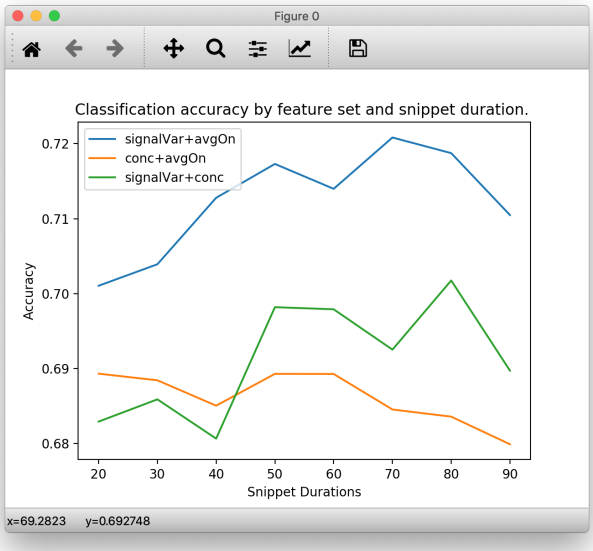


Fig. 2. Average classification accuracy by snippet duration.



that better accuracy and precision can be obtained by correlating activity between behavioral signals, but our simplistic approach to describing this correlation did more harm than good. This may indicate that automatic feature generation, of the sort ANNs are good for, could result in better f-scores.

We found that there was generally no meaningful difference in identification accuracy between different snippet durations above a threshold. As long as a sufficient amount of interaction is observed, an interaction fingerprint can be generated from our engineered features that allows for identification. 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 be the signal that solidifies their identification.

VI. 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 i.e. encryption through interaction. This method may also work with the reasonable person's expectation that only agents which the person has interacted with will remember them. A study should be conducted on interactive human-agent identification.

References

[1] Welcome to the ami corpus.

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

[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] 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.

[7] 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.

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