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 translatable in the same way, giving the user the benefits of personalized interaction without restricting their right to privacy. [JSS: *Putting this argument here makes sense given the context, but it seems weird to do it before the topic of the paper is introduced. Should I just make the footnote longer?*]

Other identification schemes, like gait detection, require the user's full body to be recorded for multiple striding frames [5][4]. It's unreasonable to expect every piece of interactive technology to implement robust person identification systems with these restrictions in mind. 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 attempts to implement the groundwork for an alternate identification system that can be used with any socially interactive device by using 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 open-set 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.

TODO: add results and explanation

II. Background Related Work

All person recognition systems extract features from the target modality to run through a classifier. Face recognition [7] and voice recognition [2] use neural networks, spectral analysis, and template matching, as well as other techniques to identify known individuals. Perez et al. showed that social media metadata, which can be considered to be a behavioral signature [JSS: *Can I say this or do I need to define what this means somewhere (or am I not allowed to define things like this)*], 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[6]. Based on Scheier et al. and Bendale et al's promising results with SVMs on open-set problems [3][6]. 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 aim to identify a set of users $u_n \subseteq U$, where the number of users in the set, and so 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.

$$[d^k_0 ... d^k_i] \in d^k \tag{3}$$

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. [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.

III. Technical Approach / Methodology / Theoretical Framework

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 hypothesizes that we would extract features description enough to accurately classify each person in the dataset.

A. Data Collection - AMI Corpus

We used the AMI Meeting Corpus [JSS: *citation?*] 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 in the meetings are assigned roles that determine their dialogue and activity within the meetings. Due to this role assignment, the reader should be aware that we may be detecting role assignments, since they may serve as a confounding variable.

We evaluated our system on meetings which had manual annotations 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 by its nature time-series data, so in order to classify we need to extract features that represent snippets of time in single values. An alternate approach might be to use RNNs to classify the time-series data without modification, but we leave that as future work.

[JSS: *shorthand brainstorming, needs editing*] Signals making up B_n are handled without taking the meaning of the signal into account. We want to build a system that generalizes across interaction types. Each signal in B_n occurs over a time period. In a given segment of the interaction each signal will be present or absent at each timestep. Using only this information we have can produce a range of potentially useful features.

1) Active Ratio - For what percentage of the snippet was the signal active.
2) Total oscillations - How many times did the signal toggle from on to off (normalized by snippet length).
3) NEED a measure that connects different b^ks and/or catches coordinated transitions between signals. Encodable by an SVM? You could do a transition matrix (e.g. b^0 turned off within $epsilon$ seconds of b^1 turning on x times in this snippet). But that's a k^2 on its own.

IV. Experimental Design

We selected three different meetings from the AMI corpus that had manual annotated leg movements, head movements, and dialogue for each participant. The annotations were extracted into data streams for each participants and split into ten, equal length segments. 80% of the data was designated as training data and paired with the corresponding participant label while the remaining 20% as testing. A scikit-learn SVM SVC one-vs-one classifier was then trained and used to classify the test set.

We explored the effect of segment length on classification accuracy by producing the ROC curve in figure doesn't-exist-yet.

We also evaluated the effectiveness of different feature sets. [JSS: *SLAUNCHWISE*]

[JSS: *Frequency domain - connecting different signals. Spectral analysis.*]

[JSS: *Covariance matrix - Eigenvectors / Eigenvalues and how they relate different variables to each other.*] [JSS: *ROC curves for each combinations*]

V. Results

VI. Discussion

References

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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