Human-Centered Machine Learning in a Social Interaction Assistant for Individuals with Visual Impairments

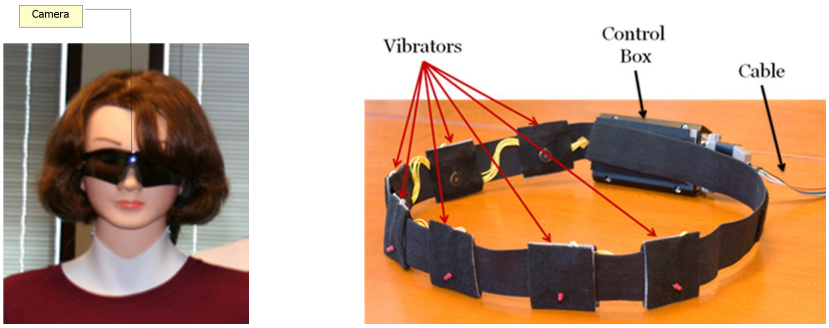
Vineeth Balasubramanian, Shayok Chakraborty, Sreekar Krishna, Sethuraman Panchanathan
Center for Cognitive Ubiquitous Computing (CUbiC)
Arizona State University
Tempe, AZ 85287
{vineeth.nb,schakr10,sreekar.krishna,panch}@asu.edu

1 Introduction

Over the last couple of decades, the increasing focus on accessibility has resulted in the design and development of several assistive technologies to aid people with visual impairments in their daily activities. Most of these devices have been centered on enhancing the interaction of a user who is blind or visually impaired with objects and environments, such as a computer monitor, personal digital assistant, cellphone, road traffic, or a grocery store. Although these efforts are very essential for the quality of life of these individuals, there is also a need (which has so far not been seriously considered) to enrich the interactions of individuals who are blind, with other individuals.

Non-verbal cues (including prosody, elements of the physical environment, the appearance of communicators and physical movements) account for as much as 65% of the information communicated during social interactions [1]. However, more than 1.1 million individuals in the US who are legally blind (and 37 million worldwide) have a limited experience of this fundamental privilege of social interactions. These individuals continue to be faced with fundamental challenges in coping with everyday interactions in their social lives. The work described in this paper is based on the design and development of a Social Interaction Assistant that is intended to enrich the experience of social interactions for individuals who are blind, by providing real-time access to information about individuals and their surrounds. The realization of a Social Interaction Assistant device involves solving several challenging problems in pattern analysis and machine intelligence such as person recognition/tracking, head/body pose estimation, gesture recognition, expression recognition, etc on a wearable real-time platform. A list of eight significant daily challenges faced by these individuals was identified in our initial focus group studies conducted with 27 individuals who are blind or visually impaired [1]. Each of these problems raises unique machine learning challenges that need to be addressed.

While the problems discussed above are typically encountered in many other fields including robotics, this application presents a unique perspective to the design of machine learning (ML) algorithms for recognition and learning: the presence of the 'human in the loop'. In addition to the challenges of implementing such systems on wearable real-time platforms, we note that the end users in this context have their cognitive capabilities intact (as we have been often reminded by our target user population). The intelligence of the human user can be judiciously used to design systems that demonstrate improved reliability and performance. More importantly, these users often would like to use their cognitive and decision-making capabilities at every possible opportunity to compensate for the sensory deficit. To illustrate with an example, if a system was built to recognize individuals standing in front of a user who is blind, the user may not like to receive a singular answer with the identity of the individual. Rather, the user would prefer to receive a set of possible identities with their confidence levels, and make the decision himself/herself. This necessitates the design of ML algorithms and systems that are user-centric by design, utilize the cognitive capabilities of the user



(a) A pair of glasses with a small camera mounted on the nosebridge

(b) A first prototype of the haptic belt. A wireless version of this belt has already been built in the second phase

Figure 1: A first prototype of the Social Interaction Assistant

to solve the problem at hand, and support the user actively in decision-making, rather than provide decisions passively. We term such algorithms as '*human-centered*' ML algorithms.

In this paper, we present a few of our efforts in machine learning that address the challenges of the Social Interaction Assistant. We first present a brief introduction to our current prototype, followed by two examples that illustrate our human-centered approach and then briefly review other ML contributions that have been made as part of this effort.

2 The Social Interaction Assistant: A Brief Overview

In our current prototype, the Interaction Assistant device consists of a pair of glasses with a small camera mounted on the nosebridge (as shown in Figure 1(a)). The incoming video stream is analyzed using computer vision algorithms to extract information about the surroundings. When a person comes in the field of view of the camera, his/her face is captured using face detection algorithms. The identity of the person is then determined by a recognition engine through a similarity match with a database of stored images [2]. Current work is focused on further analysis of the video stream in highly controlled settings (such as meetings and office rooms) to provide access to other visual nuances like gaze direction, expressions, hand gestures and attire of all the subjects in the scene, which allow the blind individual to interact socially. Most assistive devices provide only audio outputs, which is not a practical solution for visually impaired individuals in the context of social interactions, as they use their ears as their 'eyes' to perceive the environment. To overcome this fundamental limitation, we have designed a vibrotactile haptic belt, which consists of a set of 7 vibrators, to be worn by the user around his waist (Figure 1(b)). Information about the direction of an approaching individual is conveyed through the location of vibration, and the distance to the user is encoded in the duration of vibration [3]. Current work is focused on studying fundamental methods of communicating facial expressions through vibrotactile actuators.

3 Human-Centered Machine Learning: Illustrative Examples

In this section, we present two examples from our implementations that illustrate our perspectives towards human-centered machine learning:

3.1 Integrated Face Localization/Recognition

Face recognition has been a significant challenge that has been attempted by several researchers in the field for over two decades. However, the variations in facial appearance due to changes in pose, illumination and expression have made this task extremely difficult. In our prototype, we have designed an intuitive system using the wearable camera and the haptic belt to address a few challenges in this context. Our system takes advantage of the fact that face detection (we use the Viola-Jones Adaboost algorithm) has been solved to a reasonable extent, and works reliably under various real-world conditions.

When the camera on the user's glasses detects a face in the vicinity of the user, our system localizes the individual into one of 7 regions around the user (such as far left, left, right, far right, etc) (as shown in Figure 2). This is conveyed to the user through a vibrotactile cue in the haptic belt, where the location of the vibration indicates the direction of the person and the duration of vibration indicates the distance between the person and the user [4]. In our experiments, we inferred that the users found the localization vibrations extremely intuitive and automatically turned their head in the direction of the detected face, thereby leading to eye contact between the user and the individual. We gathered from our users that this, by itself, was extremely useful in how they carried themselves in social interactions. Further, this process, in turn, facilitates frontal face recognition (which is reasonably achieved in controlled scenarios) and simplifies the problem at hand, rather than tackling the challenges of pose-invariant face recognition.

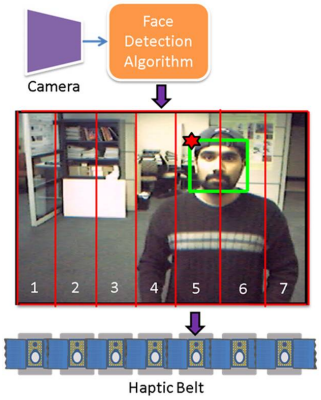


Figure 2: Integrated Face Localization/Recognition

3.2 User-conformal Confidence Measures

An essential component of any ML algorithm in a human-centered context is the computation of reliable confidence measures. Recently, a new framework called Conformal Predictions [5] has been developed to compute calibrated confidence measures that can be built on top of any existing classification/regression algorithm. We have adopted this game-theoretic approach in our work. The advantage of these confidence measures is that the obtained results (class labels in classification and output values in regression) are well-calibrated in an online setting, i.e. the frequency of errors, ϵ, made by the system is exactly bounded according to the confidence level, $1-\epsilon$, defined by the user. The results of this framework are illustrated in Figure 3. Depending on the context, the user can set a confidence threshold, and the system ensures that the number of errors made are exactly bounded by $1-\epsilon$ fraction of the total number of data samples, as theoretically guaranteed. For instance, in a party setting, the user can define a moderate level of confidence as it does not cause much harm if the prediction is incorrect. On the contrary, in an official meeting, it is of paramount importance to accurately identify all the members, and hence, a very high level of confidence may be set to control the number of errors.

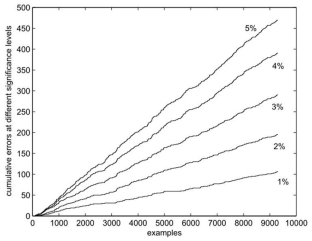


Figure 3: Results from the conformal predictions framework [5]. Note that the errors are calibrated in each of the specified significance levels.

4 Related ML Contributions

Other machine learning techniques such as online learning, active learning and learning from multiple sources form a significant component of the Interaction Assistant. These techniques are extremely essential in such an assistive orthotic device, where the system needs to learn and evolve along with the user over time. We briefly describe a few of our efforts in this direction below.

4.1 Online Active Learning for Person Recognition

The data used to train the ML algorithms in such an assistive device usually arrives sequentially over a period of time - a blind user, for example, may encounter a new subject on a particular day. To ensure that this subject is correctly recognized in future, an online learning algorithm is necessary to constantly update the classifiers with the new set of data points. Moreover, the data is available in the form of images from video capture sessions and therefore has a high redundancy. Active learning strategies iteratively select the most promising set of examples to update the current model. Online active learning is hence an essential requirement in the Social Interaction Assistant to query the most informative data examples as they arrive over time. In our current work, we have approached this problem using the theory of conformal predictions. A

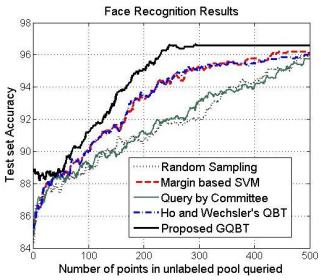


Figure 4: Results on the VidTIMIT dataset

measure of disagreement has been formulated using this theory and we query those points which present a high level of uncertainty regarding the class labels. A sample result of our work from the VidTIMIT face dataset is shown in Figure 4 [6]. It shows that using our Generalized Query by Transduction (GQBT) approach, the system queries the most salient points and thus uses fewer face images to learn (as compared to other related techniques); at the same time, it achieves a high accuracy on a predetermined test set with the minimum number of label queries.

4.2 Context-based Batch Mode Active Learning

As mentioned earlier, the goal in active learning is to select a set of promising data points to update the current hypothesis. However, often in interaction settings, an entire video is recorded and it may be necessary to choose the salient face images in the entire video consisting of thousands of frames. This results in the need for a batch mode active learning algorithm, and we have adopted an optimization framework to address this problem. Further, in our ongoing work, our experiments have shown that the use of context-based priors (whether the user is at work or at home, for example) provides opportunities to significantly improve the performance and reliability of the ML algorithms, thus leading to an integrated approach that combines the knowledge about the user with the algorithm itself.

4.3 Learning from Multiple Sources

To prevent the visually challenged user from being in an embarrassing position (by greeting the wrong person, for example), it is indispensable for ML algorithms to be reliable in the predictions. Merely relying on a single source of information may be misleading in situations where the data is difficult to analyze or is of poor quality. This problem can be alleviated by consolidating evidence from multiple sources leading to more accurate and reliable systems. Current efforts in this project are focused towards the integration of information from the face and speech modalities in a manner such that one modality can supplement/complement the predictions from the other modality. For example, if the name of a person is heard in a conversation context, this may be used to automatically label the face images obtained during the conversation, and train the classifiers accordingly.

In conclusion, we note that the Social Interaction Assistant (or any such assistive orthotic device for individuals with disabilities, at large) serves as a breeding ground for the design and development of *human-centered machine learning* approaches that provide efficient solutions by integrating the cognitive capabilities of the user in the process. We believe that this approach provides a promising direction of research in machine learning (and related fields) that is pragmatic, goal-oriented and user-centric.

References

[1] S Krishna, D Colbry, J Black, V N Balasubramanian, S Panchanathan, A Systematic Requirements Analysis and Development of an Assistive Device to Enhance the Social Interaction of People Who are Blind or Visually Impaired, ECCV Workshop on Computer Vision Applications for the Visually Impaired (CVAVI'08), Marseille, France, Oct 2008.

[2] S Krishna, G Little, J Black, S Panchanathan, A wearable face recognition system for individuals with visual impairments, Proceedings of the 7th international ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2005), Baltimore, MD, USA, Oct 2005.

[3] T McDaniel, S Krishna, V Balasubramanian, D Colbry, S Panchanathan, Using a Haptic Belt to Convey Non-Verbal Communication Cues during Social Interactions to Individuals who are Blind, IEEE International Workshop on Haptic Audio Visual Environments and Games, Ottawa, Canada, Oct 2008.

[4] N Edwards, J Rosenthal, D Moberly, J Lindsay, K Blair, S Krishna, T McDaniel, S Panchanathan, A Pragmatic Approach to the Design and Implementation of a Vibrotactile Belt and its Applications, IEEE International Workshop on Haptic Audio-Visual Environments and Games (HAVE 2009), Lecco, Italy, Nov 2009.

[5] G Shafer, V Vovk, A Tutorial on Conformal Prediction. J. Mach. Learn. Res. 9, Jun 2008, 371-421.

[6] V Balasubramanian, S Chakraborty, S Panchanathan, Generalized Query by Transduction for Online Active Learning, IEEE ICCV 2009 International Workshop on Online Learning for Computer Vision, Kyoto, Japan, Oct 2009.