Unobtrusive, Wearable Social Interaction Detection and Assistance

David S. Hayden, Robert C. Miller, Seth Teller
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
{dshayden,rcm,teller}@mit.edu

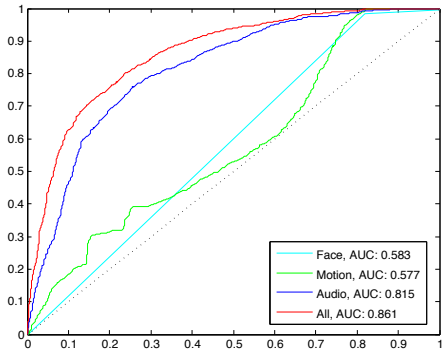

Figure 1: ROC curves of Linear SVMs trained to detect social interactions on egocentric face, motion and audio data.

AA 2014, April 27, 2014, Toronto, Canada.

Abstract
We present an unobtrusive wearable system built to provide private, timely notifications of nearby acquaintances through glanceable or vibration-encoded text and to detect social interactions via SVM classification of features extracted from egocentric image, motion and audio data.

Author Keywords
Wearable, Social Interaction, Vibration, Discreet

ACM Classification Keywords
H.5.2 [User Interfaces]: Prototyping.

General Terms
Algorithms, Experimentation, Human Factors

Introduction
We present a wearable assistant that helps its wearer socialize by providing timely identification of nearby acquiantances obtained via facial recognition. Important to our work is the detection of social interactions, which enables the wearable assistant to reason about new acquaintances it should learn. This work builds on the wearable detection of social interactions in [2] by looking beyond egocentric vision to also incorporate egocentric accelerometer and audio

Figure 2: Discreet wearable assistant as a jacket with an internal pocket out from which a camera observes, and a watch, through which private information is communicated to the wearer through glanceable or vibration-encoded text.

data, which prove helpful due to design decisions made for accomplishing a discreet wearable design with the incorporation of a private information channel in the style of [1].

Our system consists of two components. A smartphone sensor is discreetly worn within the interior lining of a modified jacket or within a pouch as a lanyard around the neck (see Figure 2. In either case, it records egocentric images, audio and motion data and communicates notifications to a wristwatch which can convey them through glanceable or vibration-encoded text. These work together to inform the wearer of nearby acquaintances and to catalog the social interactions that the user engages in.

Wearable Detection of Social Interactions

In this work, we consider a social interaction as the minimum window of time in which the wearer and another person acknowledge each other through visual and verbal means at least once. Detecting such interactions might be considered to be nearly the same as detecting faces or, alternatively, detecting voice activity. Our experiments (Figure 1) show, however, that neither signal is by itself sufficient in practice. People do not always look at each other while interacting (as when they are walking together), nor do they always speak (as when they are simultaneously observing some other event). Furthermore, the wearable assistant is likely to pick up faces and voice activity from people that are nearby, but with whom the wearer is not interacting .

We thus seek to detect social interactions with a combination of egocentric vision, accelerometer and audio data. To this end, we have built a wearable dataset consisting of images recorded at $1/3$ Hz, motion data recorded at 50 Hz, and audio data recorded at 16000 Hz. Images are taken sparsely to save battery life. The dataset consists of 2.5 hours of manually-labeled recordings spread across four separate intervals throughout a single workday. We did not record continuously because a typical day includes long stretches of time where rather little is happening (i.e. the wearer is working at their computer). Features are quantized into one-second intervals, of which 76% contain social interactions. Data are split into 70% training and 30% testing.

Social interaction detection is accomplished by SVM classification of facial features (presence, size) and a subset of the basic energy statistics used in [3], taken on both accelerometer data and the envelope of audio data). Our results can be seen in Figure 1, in which we show ROC curves of classifiers built from facial, audio and motion features. Notably, no single signal performs so well as their combination. We see that a Linear SVM on all features yields AUC of 0.861 and that audio is actually more informative than facial information for detecting social interactions, owing in part due to a low sampling rate and non-ideal image field of view of $46°$.

References

[1] Clawson, J., Patel, N., and Starner, T. Digital kick in the shin. *MobileHCI 20* (2010).

[2] Fathi, A., Hodgins, J. K., and Rehg, J. M. Social interactions: A first-person perspective. In *CVPR*, IEEE (2012).

[3] Kwapisz, J. R., Weiss, G. M., and Moore, S. A. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter 12*, 2 (2011).