Learning Visual Models of Social Engagement

Bradley A. Singletary and Thad E. Starner
College Of Computing, Georgia Institute of Technology, Atlanta, GA 30332
{bas,thad}@cc.gatech.edu

November 7, 2001

Abstract

We introduce a face detector for wearable computers that exploits constraints in face scale and orientation imposed by the proximity of participants in near social interactions. Using this method we describe a wearable system that perceives "social engagement," i.e., when the wearer begins to interact with other individuals.

Our experimental system proved $>90\%$ accurate when tested on wearable video data captured at a professional conference. Over 300 individuals were captured during social engagement, and the data was separated into independent training and test sets. A metric for balancing the performance of face detection, localization, and recognition in the context of a wearable interface is discussed.

Recognizing social engagement with a user's wearable computer provides context data that can be useful in determining when the user is interruptible. In addition, social engagement detection may be incorporated into a user interface to improve the quality of mobile face recognition software. For example, the user may cue the face recognition system in a socially graceful way by turning slightly away and then toward a speaker when conditions for recognition are favorable.

1 Introduction

In casual social interaction, it is easy to forget the names and identities of those we meet. The consequences can range from the need to be reintroduced to the "opportunity cost" of a missed business contact. At organized social gatherings, such as professional conferences, name tags are used to assist attendees' memories. Recently, electronic name tags have been used to transfer, index, and remember contact information for attendees [3]. For everyday situations where name tags are inappropriate, a face recognition system may provide face-name associations and aid in recollection of prior interactions with a conversational partner.

One way to implement such a system would be to place cameras in every environment in which a user may meet new acquaintances. This method is untenable from a cost perspective. Since wearables are self-contained, a face recognizer implemented on a wearable could function in environments devoid of specialized infrastructure. In this manner, the face recognition resource would always be available to the wearer. Intelligent interface agents implemented on top of this system can then provide the face-name associations suggested earlier [32, 20, 6]. Rhodes accurately labels such systems just-in-time information retrieval agents [24].

Effective wearable interfaces apply what they understand about the wearer's context (directly or indirectly provided by the wearer) to some problem to be solved for the user [31]. These systems must balance computation against human burden. For example, if the wearable computer interrupts its wearer during a social interaction (e.g. to alert him to a wireless telephone call), the conversation may be disrupted by the intrusion. Detection of social engagement allows for blocking or delaying interruptions appropriately during a conversation.

Hall [8] defines "near social interaction" to be from four to seven feet of separation between the participants. To segment casual social interaction visually, we identify social engagement – the first stage of social intercourse where one or both parties exchange a desire to communicate through verbal or non-verbal behaviors – as the start of conversation. Proxemics, non-verbal communication, and other social interplay are outside the scope of this paper, but we refer the reader to Hall [8] and Harrison [9] for such topics. To identify social engagement visually from the first-person perspective, we wish to use features endemic to engagement. For example, eye fixation, patterns of change in head orientation, social conversational distance, and change in visual spatial content may be relevant [30] [23]. We are as yet uncertain which features are required for recognition, so we induce a set of behaviors that assist the computer with face recognition. Specifically, the wearer aligns x's on a head-up display with the eyes of the subject to be recognized. As we learn more about the applicability of our method from our sample data set, we will extend our recognition algorithms to include non-induced behaviors.

When a conversant is socially engaged with the user, a weak constraint may be exploited for face recognition and detection. Specifically, search over scale and orientation may be limited to that typical of the near social interaction distances. Thus, a method for determining these types of social interactions may be profitable. Example works on visual modeling of human interaction include hidden Markov models (HMMs), Coupled HMMs(CHMMs)[19], and stochastic grammars [12]. These works were primarily conducted from the third-person perspective of surveillance but serve as a model for our work with the first-person. HMMs with stochastic grammars were used by Moore [17] to model complex actions. Furthermore, HMMs have been successful in recognition of American Sign Language gestures [33] and location recovery[34].

A similar problem exists in ethologically inspired robotics. In Breazeal et al.[1, 4], the authors code high level knowledge of social behavior into robots as part of a four level hybrid architecture. The knowledge of social constraints enhances situational awareness. Pre-attentive, visually-attentive, and post-attentive processing of video obtained from the robot's 'eyes' are applied in succesion, each constraining or refining the successor to a smaller, more salient, subset of visual information.

Mann [14] and Starner [32] describe manual alignment of target faces with calibration marks overlaid by a head mount display. Users must explicitly request recognition after aligning the face. A likelihood sorted list of candidates is presented to the wearer for human selection. Neither paper quantifies the performance of either human or computer at detection or recognition. If detection can be done automatically, there would be less load on the human. Conversely, if the processor or algorithm for detection are weak, hand-aligned recognition may be desirable. In Brzezowski, Dunn, and Vetter [5], a mobile system for military and police use in identifying criminals is presented. This system was a combination of commercial face recognition software and mobile-wireless variable bandwidth infrastructure. Its limitations are numerous, but it notably failed in varied lighting conditions common to mobile usage. Finally, in Iordanoglou et al. [11], a method for wearable face recognition is described but was not prototyped or tested on data acquired from a wearable computer. While the system does not address detection, it discusses algorithm performance under bandwidth limitations, a problem central to mobile computing.

While there are many face detection, localization, and recognition algorithms in the literature that were considered as potential solutions to our problem, our task is to recognize social engagement in context of human behavior and the environment. Face presence may be one of the most important features, but it is not the only feature useful for segmenting engagement. Generally, face detection consists of a search across scale and within some tolerance of in-plane or out-of plane rotations [7, 25, 29, 35, 13]. Prior work with HMMs on face detection by Nefian [18] modeled both face recognition and face detection using embedded HMMs. This work demonstrates the feasibility of HMMs for face recognition and detection. However,search at scale was performed and no background or noise models were used. Unfortunately, classic detection is usually under-constrained and over-optimistic about background content. In examination of 10 standard face databases ($> 19,000$ images)[21, 2, 29, 28, 25, 16, 10, 35, 27, 15, 22], we found that background contents had little variation. By comparison, scenes obtained from a body-worn camera in everyday life contained highly varied scene backgrounds. Furthermore, current general purpose detectors are very compute-intensive due to searching at scale. Though a low false positive rate is relatively important for interface reasons, the computational price without specialized hardware is generally unacceptable. Expensive algorithms should only be computed if there is a reasonable chance a face exists or if fine grain localization is required. We detail comparison metrics below for deciding which methods better satisfy real-time interface requirements.

2 Engagement Dataset

Many face detection and face recognition datasets are constructed with the goal of understanding how to separate faces from their background and how to separate identity across face images respectively. Most face databases are collected under controlled camera parameters, lighting, and orientation. It is not clear that results derived from these previous experiments will be applicable across the environments where everyday human interaction occurs. Therefore, we collected video data from a wearable camera at an academic conference, a setting representative of social interaction of the wearer and new acquaintances. The capture environment was highly unconstrained and ranged from direct sunlight to darkened conference hall. Approxi-

Figure 1: Representative data set

Figure 2: Marks for user alignment and face capture apparatus

mately 300 subjects were captured one or more times over 10 hours of captured video. The images in Figure 1 are locations in the video annotated by the wearer to be faces.

We assembled a prototype wearable camera system to acquire necessary preliminary test data. (see Figure 2)The apparatus consists of: a color camera, an infrared(IR) sensitive black and white camera, a low-power IR illuminator, two digital video(DV) recorder decks, one video character generator, one audio tone generator, and four lithium ion cam-corder batteries. The DV deck, character generator, tone generator, camera DSP unit, and battery/power system are housed in a camera vest. The cameras, head mount display, and infrared illuminator are mounted on a plastic helmet for increased stability and precision of capture. Using infrared sensitive cameras with infrared emitting illuminators allows for night time capture or semi-covert operation. Unfortunately, data captured at the conference did not make use of the illuminator as sunlight and incandesant lighting provided more than enough IR radiation for capture.

The output of one camera is split with a low-power video distribution amplifier and displayed in one eye of the head mount display. The signal is annotated with two 'x' characters spaced and centered horizontally then placed one third of the way from the top of the video frames (Figure 2). The other copy of the signal is saved to DV tape. To capture face data, the wearer of the vest approaches a subject and aligns the person's eyes with the two 'x' characters. The 'x' characters represent known locations for a subject's eyes to appear in the video feed. The marks and lens focus are ideally calibrated to be appropriate for footage taken at normal conversational distances from the subject. Once the marks are aligned, the wearer pushes a button that injects an easily detected tone into the DV deck's audio channel for later recovery. The audio tones serve as ground-truth markers for training purposes.

3 Method and Results

The video data was automatically extracted into 2 second partitions and divided into two classes using frames annotated by the wearer. The two classes were 'engagement' and 'other'. Due to the fact that the wearer annotation was highly incomplete, we had to filter the remaining facial interactions that were misclassified. This editing was also used to protect the privacy of non-participants in the experiment. As may be expected, the number of engagement gestures per hour of interaction was much smaller than the number of examples in the garbage class. Finally, the time window was selected based on our prior belief that engagement happens in a time window of approximately half a second to two seconds.

Since the wearer lined up two x's with the eyes of a viewed subject, the presence of a face could safely be guaranteed to be framed by a 360x360 subregion of the 720x480 DV frame at the annotated locations in the video. Faces present at engagement were large with respect to the subregion. We first convert to greyscale, deinterlace, and correct non-squareness of the image pixels in the subregion. We then used Gaussian sub-sampling to reduce the size of the images to 22x22 pixels. Therefore, each feature vector consists of 484 elements. We model the face class by a 3 state left-right HMM as shown in Figure 3. The other class was much more complex to model and required a 6 state ergodic model to capture the interplay of garbage types of scenes as shown in Figure 4. We plot the mean values of the state output probabilities. The presence of a face seems important for acceptance by the engagement model. The first state contains a rough face-like blob and is followed by a confused state that likely represents the alignment portion of our gesture. The final state is clearly face-like, with much sharper features than the first state and would be consistent with conversational engagement. Looking at the other class model, we see images that look like horizons and very dark or light scenes. The complexity of the model allowed wider variations in scene without loss in accuracy. It is clear that different environments and viewpoints would derive different model structures. Thus, user and location specific models can likely be derived or adapted to improve the general detection strategy. For single user wearables, only learning location-dependent models may be sufficient. For a reference on visual modeling of location see Rungsarityotin [26].

Accuracy results are shown in Table 1. Confusion matrices are given in Table 2 and Table 3.

Table 1: Accuracy of engagement detection

experiment	*training set*	*independent test*
22x22 video stream	89.71%	90.10%

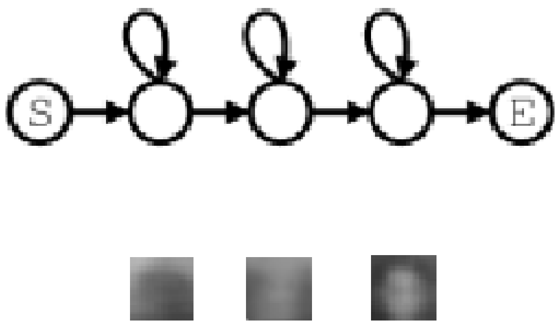

Figure 3: Engagement class

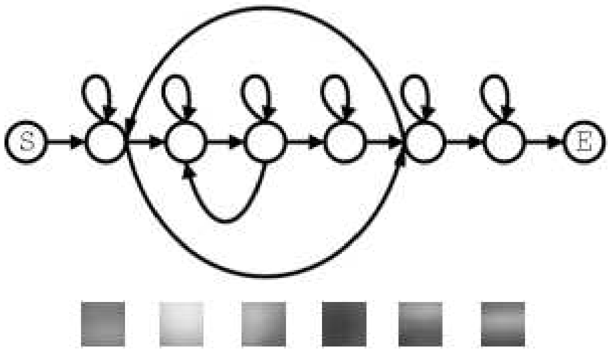

Figure 4: Other class

Table 2: Confusion matrix for engagement detection on training data

train confusion, N=843	*engagement*	*other*
engagement	82.1%(128)	17.9%(28)
other	8.6%(63)	91.3%(665)

Table 3: Confusion matrix for engagement detection on independent test data

test confusion, N=411	*engagement*	*other*
engagement	83.3%(50)	16.7%(10)
other	8.7%(30)	91.3%(314)

4 Evaluation Metrics for Separating Face Detection, Localization, and Recognition

In wearable computing, battery life and processor speed are at a premium, resulting in a very specific evaluation criteria for our system. How effective is leveraging detection of social engagement as compared to continuously running face recognition? If we were to construct a wearable face recognition system using our engagement detector, we would combine the social engagement detector with a scale-tuned localizer and a face recognizer. The cost of the social engagement detector must be sufficiently small to allow for the larger costs of localization and recognition. This is described by the inequality

$$z - R_a * a \geq R_b * b$$

where $z := 1$ is the total resources for detection,localization, and recognition, a is the fixed cost of running engagement detection once in sec/frames, b is the fixed cost of running localization, and recognition methods once in sec/frames, R_a and R_b are the rate at which we can supply the respective detectors with frames in frames/sec. However, R_b has a maximum value determined by either the fraction of false positives U_{fp} multiplied by the maximum input frame rate or the rate at which the user wants to be advised of the identity of a conversant R_{ui}. Thus, the above equation can be re-written

$$R_b * b \geq max\{R_a * U_{fp}, R_{ui}\} * b$$

Note that fixating the camera on a true face could cause up to R_a frames per second to be delivered to the face recognizer. However, we assume that the user does not want to be updated this quickly or repeatedly (i.e. $R_{ui} << R_a$). We also assume that our rate of false positives will almost always be greater than the rate the user wants to be informed, leaving us with

$$1 - R_a * a \geq R_a * U_{fp} * b$$

For comparison purposes, we will assume that the average time per frame of processing for the localization and recognition process can be represented by some multiple of the average detection time (i.e. $b = c * a$). Thus, for a given multiplier c, we can determine the maximum rate of false positives allowable by the face detection process.

$$U_{fp} \leq \frac{1}{R_a * a * c} - \frac{1}{c}$$

Note that if $c \leq 1$, then the localization and recognition process runs faster than the face detection process. This situation would imply that performing face detection separately from face localization and recognition would not save processing time (i.e. localization and recognition should run continually - again, if real-time face recognition is the primary goal).

Given a false positive rate U_{fp}, we can solve the equation to determine the maximum allowable time for the localization and recognition process as compared to the detection process.

$$c \leq \frac{1}{R_a * a * U_{fp}} - \frac{1}{U_{fp}}$$

Thus, we have a set of heuristics for determining when the separation of face detection and face localization and recognition is profitable.

5 Discussion and Applications

Applying the metric from the previous section to our experimental results, we let $U_{fp} = .13$, $R_a = 30$, $a = \frac{1}{60}$ and solving for c we get $c \leq 7.69$. Thus any recognition method used may be up to 7.69 times slower than the engagement detection method and will have a limiting frame rate of about four frames per second. Given that our detection algorithm runs at 30fps, and our knowledge that principal component analysis based face recognition and alignment can run faster than roughly four times a second, we feel that engagement detection can be a successful foundation for wearable face recognition. Post-filtering outputs of detection may help eliminate false positives before recognition [7]. Due to the face-like appearance of the final state of the HMM, it is likely that the output of our method could provide a reasonable first estimate of location to fine grain localization.

We are beginning to model other modalities of engagement behavior. Engagement detection failure in one modality may be discounted by addition of further sensors on the user. For example, Selker [30] discusses an eye fixation detector; eye fixation may help indicate social engagement. Two parties meeting for the first time will usually look to see whom they are meeting. Sound may provide another modality with which to detect social engagement. For instance, personal utterances like "hello, my name is ..." are common during social engagement. A simple range sensor using sonar or pulsed IR could be mounted on the camera to determine presence of objects within near social interaction distances. This could be used as a trigger for activating body-worn cameras. Finally, we have constructed, but not yet integrated, a vision-based walking/not-walking classifier. Detection of head stillness and other interest indicators will likely reduce false positives in our system[23].

We are considering several applications for this technology. Face recognition on a wearable platform aids and protects military and law enforcement officers in the field by providing personnel the ability to conduct comparisons of viewed subjects to records of wanted criminals. Such a tool would ideally augment wanted posters and visual comparison, reduce human error, reduce time to capture, and reduce legal costs. To directly aid border guards, sentries, and patrol officers, such a system should be configured to function in daylight or at night. Medical benefit can be realized by people that suffer from prosopagnosia (face blindness) by restoring their ability to learn and recognize faces directly. More generally, such a system may be useful to anyone who needs to associate large numbers of names with faces. For example, salespersons and politicians could suddenly recall a person's name and any previous salient interactions. As a final application, we are considering the creation of an attention manager to protect the wearer from stimulus overload. Detecting conversational context is key to handling distractors, such as cellular phone calls, in an intelligent fashion.

6 Conclusion

From examining publicly available face databases, face detection and recognition on a wearable computer appear significantly different from off-the-body scenarios. Large variations in lighting, scene, and camera position occur in everyday, human-centered situations. In addition, traditional systems do not exploit constraints in the user's environment that improve the efficiency and accuracy of detection. We described a platform built to capture data from a wearable user's perspective and detailed a method for efficient engagement detection on a wearable computer. Furthermore, we present a metric to determine when separating face recognition into detection and recognition components is profitable from an architectural standpoint. We are currently constructing a prototype face recognition wearable that detects engagement via the described method in order to validate our initial experimental results derived from wearable data.

References

[1] B. Adams, C. Breazeal, R. Brooks, and B. Scassellati. Humanoid robots: A new kind of tool. *IEEE Intelligent Systems*, August 2000. to appear.

[2] Y. Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *Pattern Analysis and Machine Intelligence*, 19(7):721–732, July 1997.

[3] R. Borovoy, M. McDonald, F. Martin, and M. Resnick. Things that blink: A computationally augmented name tag. *IBM Systems Journal*, 35(3), 1996.

[4] C. Breazeal, A. Edsinger, P. Fitzpatrick, B. Scassellati, and P. Varachavskaia. Social constraints on animate vision. *IEEE Intelligent Systems*, August 2000. to appear.

[5] S. Brzezowski, C. M. Dunn, and M. Vetter. Integrated portable system for suspect identification and tracking. In A. T. DePersia, S. Yeager, and S. Ortiz, editors, *SPIE:Surveillance and Assessment Technologies for Law Enforcement*, 1996.

[6] J. Farringdon and V. Oni. Visually augmented memory. In *Fourth International Symposium on Wearable Computers*, Atlanta, GA, 2000. IEEE.

[7] R. Feraud, O. J. Bernier, J.-E. Viallet, and M. Collobert. A fast and accurate face detector based on neural networks. *Pattern Analysis and Machine Intelligence*, 23(1):42–53, January 2001.

[8] E. T. Hall. *The Silent Language*. Doubleday, 1963.

[9] B. L. Harrison, H. Ishii, and M. Chignell. An empirical study on orientation of shared workspaces and interpersonal spaces in video-mediated collaboration. Technical Report OTP-94-2, University of Toronto, Ontario Telepresence Project, 1994.

[10] D. Hond and L. Spacek. Distinctive descriptions for face processing. In *8th British Machine Vision Conference*, pages 320–329, Colchester, England, September 1997.

[11] C. Iordanoglou, K. Jonsson, J. Kittler, and J. Matas. Wearable face recognition aid. In *Interntional Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2000.

[12] Y. Ivanov, C. Stauffer, A. Bobick, and E. Grimson. Video surveillance of interactions. In *CVPR Workshop on Visual Surveillance*, Fort Collins, CO, November 1999. IEEE.

[13] T. K. Leung, M. C. Burl, and P. Perona. Finding faces in cluttered scenes using random labelled graph matching. In *5th Inter. Conference on Computer Vision*, 1995.

[14] S. Mann. Wearable, tetherless computer–mediated reality: WearCam as a wearable face–recognizer, and other applications for the disabled. TR 361, MIT Media Lab, Cambridge, MA, February 1996.

[15] E. Marszalec, B. Martinkauppi, M. Soriano, and M. Pietikinen. A physics-based face database for color research. *Electronic Imaging*, 9(1):32–38, 2000.

[16] A. M. Martinez and R. Benavente. The ar face database. TR 24, CVC, June 1998.

[17] D. J. Moore. *Vision-based recognition of actions using context*. PhD thesis, Georgia Institute of Technology, Atlanta, GA, 2000.

[18] A. Nefian. *A hidden Markov model-based approach for face detection and recognition*. PhD thesis, Georgia Institute of Technology, Atlanta, GA, August 1999.

[19] N. Oliver, B. Rosario, and A. Pentland. Statistical modeling of human interactions. In *CVPR Workshop on Interpretation of Visual Motion*, pages 39–46, Santa Barbara, CA, 1998. IEEE.

[20] A. Pentland. Looking at people: sensing for ubiquitous and wearable computing. *Pattern Analysis and Machine Intelliegence*, 22(1):107–119, Jan 2000.

[21] A. Pentland, T. Starner, N. Etcoff, A. Masoiu, O. Oliyide, and M. Turk. Experiments with eigenfaces. In *Looking at people workshop: IJCAI93*, Chamberry, France, August 1993.

[22] Psychological image collection at stirling (PICS). available at:http://pics.psych.stir.ac.uk/.

[23] J. Reeves. The face of interest. *Motivation and Emotion*, 17(4), 1993.

[24] B. Rhodes and P. Maes. Just-in-time information retrieval agents. *IBM Systems Journal special issue on the MIT Media Laboratory*, 39(3-4):685–704, 2000.

[25] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1), January 1998.

[26] W. Rungsarityotin and T. Starner. Finding location using omnidirectional video on a wearable computing platform. In *International Symposium on Wearable Computing*, Atlanta, GA, October 2000. IEEE.

[27] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. In *2nd IEEE Workshop on Applications of Computer Vision*, Sarasota FL., December 1994. IEEE.

[28] H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *Computer Vision and Pattern Recognition*, pages 45–51. IEEE, July 1998.

[29] H. Schneiderman and T. Kanade. A statistical model for 3d object detection applied to faces and cars. In *Computer Vision and Pattern Recognition*. IEEE, June 2000.

[30] T. Selker, A. Lockerd, and J. Martinez. Eye-r, a glasses-mounted eye motion detection interface. In *to appear CHI2001*. ACM, 2001.

[31] B. A. Singletary and T. Starner. Symbiotic interfaces for wearable face recognition. In *HCII2001 Workshop On Wearable Computing*, New Orleans, LA, August 2001. Lawrence Erlbaum. to appear.

[32] T. Starner, S. Mann, B. Rhodes, J. Levine, J. Healey, D. Kirsch, R. W. Picard, and A. Pentland. Augmented reality through wearable computing. *Presence special issue on Augmented Reality*, 1997.

[33] T. Starner and A. Pentland. Real-time American sign language recognition using desktop and wearable computer based video. *Pattern Analysis and Machine Intelligence*, December 1998.

[34] T. Starner, B. Schiele, and A. Pentland. Visual contextual awareness in wearable computing. In *International Symposium on Wearable Computing*, 1998.

[35] K. K. Sung and T. Poggio. Example-based learning for view-based human face detection. *Pattern Analysis and Machine Intelligence*, 20(1):39–51, January 1998.