Attentive Sensing for Long-Range Face Recognition

Helio Perroni Filho
helio@yorku.ca

Aleksander Trajcevski
atrajcev@yorku.ca

Kartikeya Bhargava
kkb@yorku.ca

Nizwa Javed
nizwaj@yorku.ca

James H. Elder
jelder@yorku.ca

Centre for AI & Society, York University
Department of Electrical Engineering and Computer Science
Toronto ON M3J 1P3, Canada
https://www.elderlab.yorku.ca/

Abstract

To be effective, a social robot must reliably detect and recognize people in all visual directions and in both near and far fields. A major challenge is the resolution/field-of-view tradeoff; here we propose and evaluate a novel attentive sensing solution. Panoramic low-resolution pre-attentive sensing is provided by an array of wide-angle cameras, while attentive sensing is achieved with a high-resolution, narrow field-of-view camera and a mirror-based gaze deflection system. Quantitative evaluation on a novel dataset shows that this attentive sensing strategy can yield good panoramic face recognition performance in the wild out to distances of ∼35m.

1. Introduction

To work well with human populations, a mobile robot must be broadly socially aware, able to detect and recognize the people around it in the environment, identify human attributes (e.g., age, gender), and estimate emotional states and intentions. These requirements suggest a very wide (ideally panoramic) sensory Field-of-View (FoV) to avoid blind spots, i.e., directions in which the robot is unaware of human occupancy or activity. At the same time, identifying individuals, estimating traits and understanding intent in the far field generally requires high spatial acuity, to support face or expression recognition or estimation of gaze direction, for example. For a fixed sensor pixel resolution, this generates a **resolution-FoV tradeoff**: expansion of the FoV to support wide-field awareness leads to a reduction in acuity needed for interpretation, especially in the far field.

How does the human visual system address this tradeoff? Humans have a wide-field binocular visual system, a fast and accurate oculomotor plant, spatial remapping and short-term memory systems that integrate over time. Foveation of the retina allows for instantaneous processing of fine spatial detail at selected gaze points in the scene, typically sampled at a rate of 2-3 fixations per second. Due to the exponential falloff in acuity with eccentricity, human visual performance depends profoundly on judicious selection of gaze points and accurate interception of gaze targets. For robots to operate successfully in social environments they must solve many of the same problems as humans, and in particular balance the tradeoff between whole-field 3D spatial awareness and the ability to process finer detail in the parts of the scene relevant to the task at hand.

Here we study a novel bio-inspired attentive sensor architecture to address this challenge, and evaluate it on the core task of long-range face recognition.

2. Related Work

Early work on attentive machine vision systems typically employed two cameras: A wide-field (e.g., 130 deg FoV) pre-attentive camera mounted in close proximity to a second narrow-field (e.g., 13 deg FoV) camera on a pan-tilt unit [12, 35, 22, 13, 14]. People are detected in the pre-attentive video stream and pan/tilt control is used to direct the attentive sensor to these saccadic targets, allowing high-resolution capture. Bellotto et. al [5] extended these ideas to more widely-displaced pre-attentive and attentive cameras, supporting active tracking and zoom control.

While an ultimate aim of this early work was to support biometrics, including face recognition in the far field [21], this was not demonstrated, since face recognition systems

of that time performed well only under close-range, controlled conditions. In one of the earliest attempts to use active sensing for facial biometrics, Li et al. [18], using active near-IR illumination and sensing to allow low-light operation, demonstrated an attentive system consisting of a single wide-angle pre-attentive camera and two narrow-FoV attentive cameras on pan-tilt units. In preliminary results, they reported an improvement in face recognition performance for attentive over pre-attentive sensing, but only for one image containing three people. Yu et. al [31] extended these ideas to larger networks of cameras, but only evaluated face detection and recognition on a single video and single subject.

Wheeler et. al [29] performed a more comprehensive study of the potential for attentive sensing to support face recognition at a distance, employing a standard pairing of fixed wide-FoV pre-attentive and narrow-FoV attentive cameras with PTZ control. People were detected and tracked in the pre-attentive stream and used as saccadic targets to control pan, tilt and zoom of the attentive camera. Pre-attentive person detection and recognition were demonstrated out to $\sim$ 65m and 22m, respectively, but quantitative face recognition accuracy was not reported.

While this prior work focused on fixed installations of cameras for surveillance applications, the maturation of mobile robot technologies raises the possibility of incorporating attentive sensing into robot architectures to improve social awareness in the far field. While a distinction between pre-attentive and attentive processing has been made for robot software design (e.g., [2, 8, 7]) and for robot sensor chips (e.g., [3]), and attention has been used as a framework for the control of binocular robot eyes (e.g., [19]), none of this work has addressed the resolution-FoV tradeoff.

To summarize, while prior work has shown the potential for attentive sensing to support better social awareness in the far field, quantitative studies of attentive face recognition accuracy in the far field are lacking, and attentive sensing architectures suitable for social robot applications have not been deeply explored. The aim of this paper is to propose such an architecture and to quantitatively assess its accuracy on the task of face recognition in the far field.

3. Hardware Platform

Our Attentive Sensor hardware platform (Fig. 1) is mounted atop a mobile robot comprised of a Clearpath Dingo holonomic base and a custom aluminum frame that brings it close to human height (1.6m). The platform is comprised of three modules, responsible for pre-attentive sensing, attentive sensing and attentive gaze control.

Pre-Attentive sensing is provided by four RealSense D455 RGBD 1280×720 pixels cameras mounted horizontally at 90 deg intervals (Fig. 3(a)). With nearly 90° horizontal FoV they collectively provide a panoramic pre-attentive FoV. While the potential resolution of this pre-attentive stream is $5K$ horizontal, due to bandwidth limitations we are currently running the RealSense cameras at 640×360 resolution, yielding a pre-attentive panoramic resolution of 2560×360 pixels.

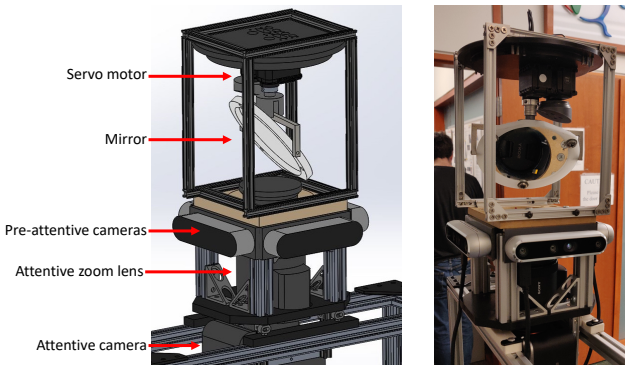


Figure 1: Attentive sensor design (left) and as-built (right).

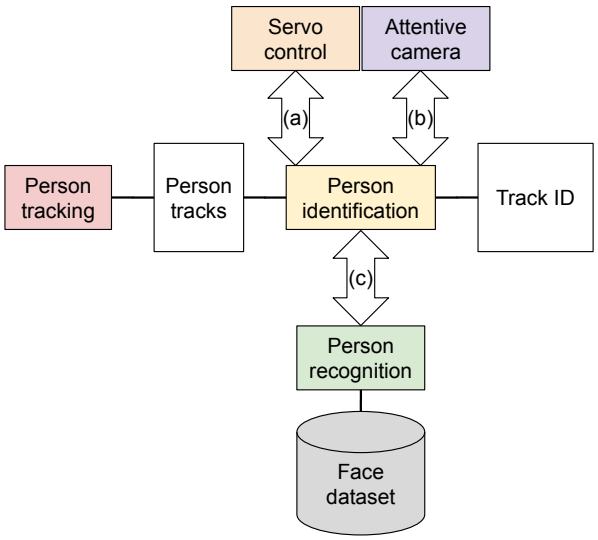
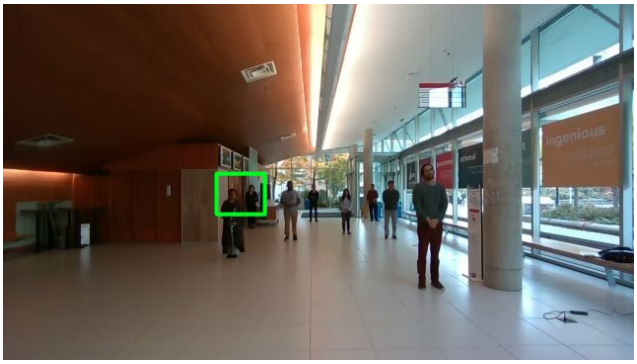


Figure 2: Vision pipeline. Colored boxes represent functional components, and white boxes describe information exchanged between components. The person tracker sends track updates to the person identification module; when prompted by an external source to find a person of given ID (not shown), the module selects a track and (a) directs the mirror servo to point to its latest coordinates, (b) retrieves an image from the attentive camera and (c) sends it alongside the sought-for ID to the person recognition module.

Attentive sensing is provided by a Sony Alpha 7C 3840×2160 pixel camera with a Sony E PZ 18-200mm powered zoom lens fixed at its longest focal length, yielding an 8 deg horizontal FoV (Fig. 3(b)). Adaptive focus is critical for acquiring in-focus images of human activity



(a) Pre-attentive image. The green box indicates the gaze direction for the attentive image below.

(b) Attentive image. The trapezoidal attentive FoV arises from the oblique azimuthal orientation of the mirror relative to the camera plane axes.

Figure 3: Example pre-attentive and attentive images.

over a broad range of distances; here we operate the attentive camera in auto-focus mode. The (linear) visual acuity of the attentive stream is roughly 65 times higher than the pre-attentive stream (7.5 sec arc vs 8 min arc). The attentive camera is mounted vertically below the pre-attentive sensors and centred horizontally so that its lens passes between the pre-attentive sensors and its vertical optic axis roughly intersects their horizontal axes.

Attentive gaze control is accomplished by a mirror mounted at a 45 deg angle on a servo motor coaxial with the attentive optic axis. Rotation of the motor shifts the azimuth of the attentive FoV, providing attentive resolution in any direction of interest identified from the pre-attentive stream.

The servo motor and Sony camera are controlled by an onboard computer through USB connections. Pre-attentive and attentive frames are streamed over a dedicated Wi-Fi link to an external server that runs person detection, face detection and face recognition algorithms.

4. Vision Pipeline

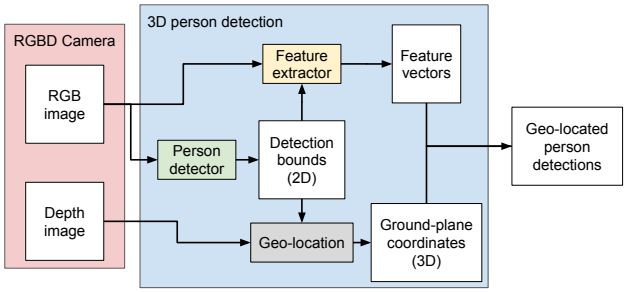
Our vision pipeline, implemented within the ROS framework [23], is composed of five components (Fig. 2):

1. A multi-camera RGBD tracking pipeline providing pre-attentive panoramic human tracking in ground-plane coordinates;
2. A servo controller that rotates the mirror to deflect attentive gaze to a person of interest;
3. A ROS bridge node for capturing attentive still frames;
4. A face detection / recognition pipeline for attentive person identification;
5. An identification module that coordinates the work of the previous components to tag person tracks with identification data.

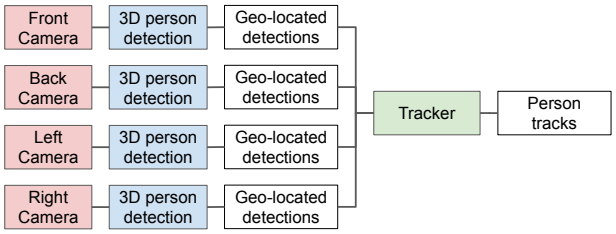
The person tracking pipeline — shown in Fig. 4 — is a modified DeepSORT [30] implementation that takes as input timestamped and synchronized images from the four pre-attentive RGBD cameras. RGB and depth images are acquired by a set of bridge nodes running on the onboard computer, then relayed to the external server. Person detection is performed on RGB images using the YOLOv4 object detector [6]. For each detected person, the corresponding image crop is extracted, and a feature vector encoding their appearance is computed by a wide residual network [32]. The detection is also geo-located through a hybrid strategy:

1. If the pre-attentive camera provides depth returns within the bounding box, and the mean of these returns is less than 6m, we geo-locate the person at the azimuth of the bounding box centroid at a distance determined by the mean of the depth returns within the bounding box.
2. If the bounding box contains no depth returns, or their average exceeds 6m (i.e., beyond the reliable range), we back-project the center of the bottom of the bounding box to the ground plane, using the pre-determined projection matrix of the pre-attentive camera.

After being computed, geo-located detections (composed of a feature vector and ground plane coordinates) are submitted to a tracker. Detections are paired to existing tracks through the Hungarian algorithm [17], based on a metric that combines appearance similarity (defined as the cosine distance between feature vectors) and Euclidean distance in ground plane coordinates. Any detection not assigned to an existing track is used to initialize a new track, and any tracks not assigned any detection for longer than a threshold period are discarded. A linear Kalman filter is



(a) Person detection and 3D projection. For each RGBD camera in the array, a projection pipeline detects people in the RGB stream, extracts feature vectors, and assigns them world 3D coordinates computed with the help of the depth map.



(b) Multi-camera person tracking. The projection streams computed for each camera are fed into DeepSORT, which maintains a collection of tracks representing individual humans as they move over the groundplane.

Figure 4: Person tracking pipeline for the pre-attentive camera array. Colored boxes represent functional components, and white boxes describe information exchanged between components.

used to predict the current position of each tracked person before tracks are compared to the latest detections.

Because tracks are located in ground-plane space, and not image coordinate space, a single track can seamlessly transition between cameras, allowing consistent person tracking across the entire panoramic pre-attentive FoV.

Person track updates are passed to the identification module. When this module receives a request to locate a person by their ID, it performs the following steps:

1. Select one of the current tracks and extract the individual's current azimuthal location;
2. Instruct the servo controller to deflect the mirror to the target azimuth;
3. Retrieve a frame from the attentive camera;
4. Submit the attentive frame and person ID to the person recognition module;
5. If the ID is positive, return the track ID;
6. Otherwise, return to step (1) and select a different track.

Our face recognition module is built on deepFace [25], a Python framework that provides a unified interface to a wide variety of face detection and recognition models. We provide access to a gallery of reference face images associated with unique IDs. When initialized, the module loads the gallery and builds a dictionary of face vectors indexed by ID, using a pre-selected recognition model. When prompted with a query ID, the module performs the following steps:

1. Retrieve the face vectors corresponding to the query ID;
2. Detect faces in the attentive image;
3. Compute the face vector for each detected face;
4. Measure the cosine distance between detected and query face vectors;
5. If the minimum cosine distance is below a pre-selected threshold, return a positive ID response;
6. Otherwise, return a negative ID response.

5. Experiment

11 volunteers agreed to participate as subjects in our experiment. Each read and signed a waiver form accepting the use of their data for this publication. For each participant we created a gallery face image dataset of images with the head in five different poses (Fig. 5), generated by asking the observer to rotate their head to direct their gaze toward five different markers on the wall, floor and ceiling, while maintaining a central position of their eyes in their head. Images were captured in uniform lighting against a blank wall, using a high-resolution DSLR camera.

We evaluated our attentive sensor system in two different indoor environments (Fig. 6): a relatively open 25×7m rectangular foyer, and a longer corridor that allowed us to stretch distances to 35m. Participants stood at various distances from the sensor, often looking towards it, but sometimes looking to the side as they talked to each other, or gazing down at mobile devices.

To collect the dataset we ran a modified form of the attentive vision pipeline (Fig. 2), detecting people in the attentive stream, and then fixating each detected person bounding box to collect and store an attentive image. In order to evaluate and compare face detection and recognition at pre-attentive and attentive resolutions, we manually annotated bounding boxes for each face within both pre-attentive and attentive streams and then ran each of the face recognition systems offline for each individual in the gallery, and over all annotated faces in the pre-attentive and attentive

Figure 5: Example set of reference images for face recognition. For each participant, five pictures were taken, while the participant looked forward, up, down, left and right.

streams. While face detection algorithms were evaluated on uncropped pre-attentive and attentive image streams, face recognition algorithms were evaluated on the manually annotated face bounding boxes, for fair comparison.

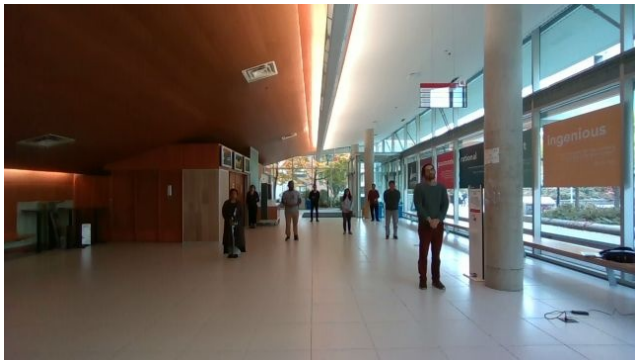
We tested six face detection systems:

- Haar Cascades [28];
- ResNet-10 [15];
- HoG + SVM [9];
- MTCNN [33];
- RetinaFace [10];
- BlazeFace [4].

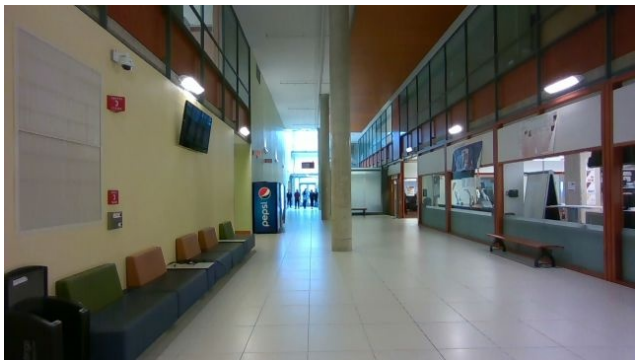
Each detector returns a confidence for each face detected; varying a threshold on this confidence sweeps out a precision-recall curve. We associate above-threshold detections with ground truth faces by solving for the assignment that maximizes average intersection over union (IoU), using the Hungarian algorithm [17]. Assignments with IoU over 0.5 are considered hits.

We also tested eight face recognition systems:

- VGG-Face [20];
- Facenet [24] with 128- and 512-dimension vectors;



(a) Environment 1: a large indoor area with glass walls to the right and front of the Attentive Sensor.



(b) Environment 2: a relatively narrow corridor with uneven lighting.

Figure 6: Environments used for image collection.

- OpenFace [1];
- DeepFace [27];
- DeepID [26];
- ArcFace [11];
- Dlib [16], a customized version of ResNet-34 [15];
- SFace [34].

To evaluate a face recognition model, we select in turn each of the 11 individuals in our gallery as a query ID. We then consider each in-the-wild attentive and pre-attentive facial image, identifying its minimum cosine distance to the five gallery images of the query ID. Varying a threshold on this minimum cosine distance sweeps out an ROC curve for the model.

6. Results and Discussion

6.1. Face Detection

Four of the six detectors tested failed to detect *any* faces in the pre-attentive stream, and RetinaFace and MTCNN

managed to detect only a few faces, achieving AP scores of 0.074 and 0.004, respectively.

Face detection performance was much better in the attentive stream (Fig. 7), demonstrating the benefit of attentive sensing for accurate panoramic face detection. Ranking of the detectors on this task is consistent with prior reports: RetinaFace achieved near-perfect performance, followed closely by MTCNN. ResNet-10 and BlazeFace, light models that trade accuracy for speed and are meant for use in mobile devices to locate close faces, achieved the worst results. Interestingly, Haar Cascades and HoG + SVM, older methods not based on deep learning, achieved intermediate results — notably worse than SoTA deep models for face detection in the wild, but still better than lighter-weight deep models.

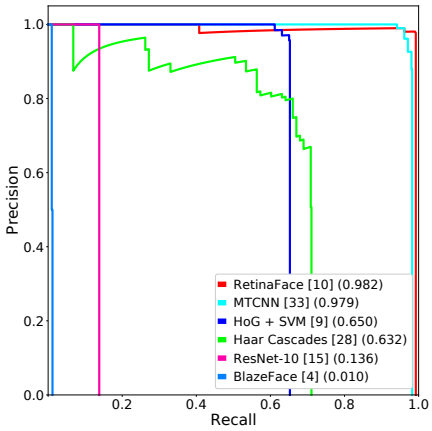


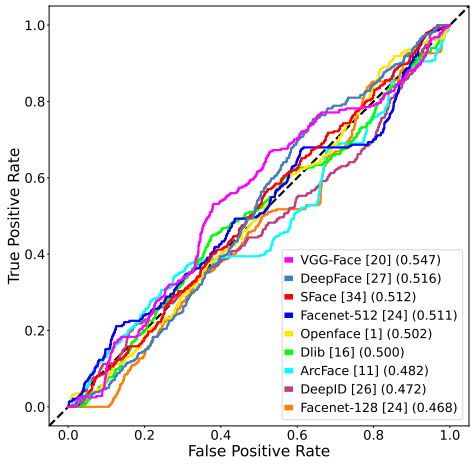
Figure 7: Attentive face detection performance. RetinaFace achieves near-perfect performance, followed closely by MTCNN. Haar Cascades and HoG + SVM, older methods not based in Deep Learning, did considerably worse. ResNet-10 and BlazeFace, light models that trade accuracy for speed, achieved the worst results. Results for pre-attentive images are omitted, since no model could detect faces on them to any relevant degree.

6.2. Face Recognition

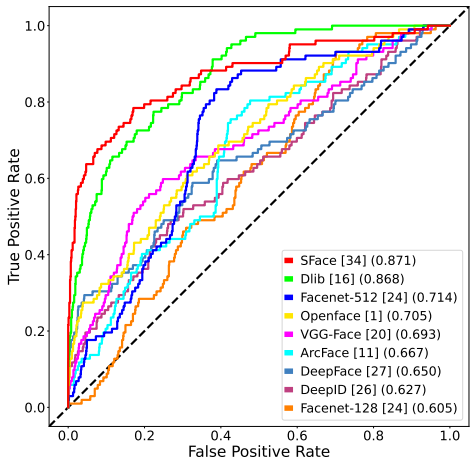
All face recognition models performed at or near chance on the pre-attentive stream (Fig. 8a).

Face recognition performance was much better on the attentive stream (Fig. 8b), demonstrating the benefit of attentive sensing for panoramic face recognition. Ranking of the models on this task is generally consistent with prior reports (e.g., [34]). We found the strongest model to be SFace [34], which uses a MobileNet backbone and is trained using a loss function robust to outliers.

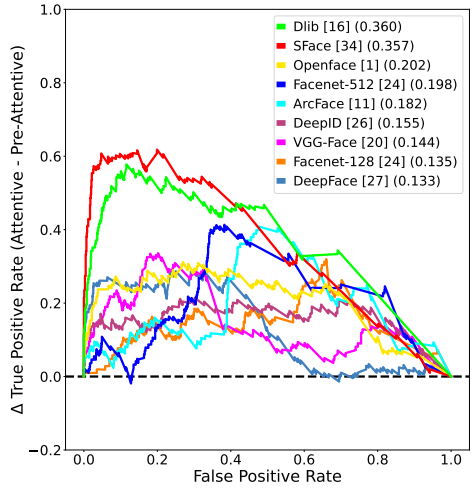
Fig. 8c differences the attentive and pre-attentive ROC curves to show the improvement in face recognition performance due to attentive sensing. All models enjoy a sub-



(a) Pre-attentive.



(b) Attentive.



(c) Difference.

Figure 8: Face recognition performance

stantial boost from attentive sensing. To test the statistical significance of this attentive boost, we measured the equal-error-rate accuracy separately for each model and each individual in our gallery, using both pre-attentive and attentive streams, and then performed for each model a matched-sample t-test of the mean equal-error-rate accuracy for attentive vs pre-attentive sensing. We found (Fig. 9) that attentive sensing produces an *attentive boost* in equal-error-rate accuracy of up to 30%. Although our experiment was based on a relatively small gallery of 11 individuals, our statistical testing suggests that, for 5 of the 9 models (Dlib, SFace, ArcFace, Facenet512 and OpenFace), this attentive boost should generalize to new datasets.

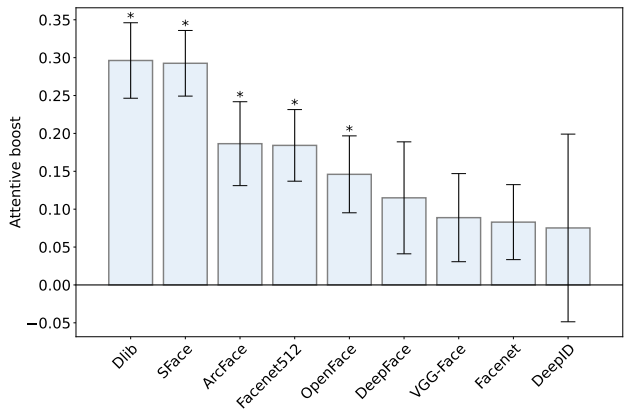


Figure 9: Attentive boost in face recognition performance. Bars and error bars indicate mean and standard error of the increase in equal-error-rate accuracy for attentive vs pre-attentive sensing. '*' indicates statistical significance at the $p = .05$ level.

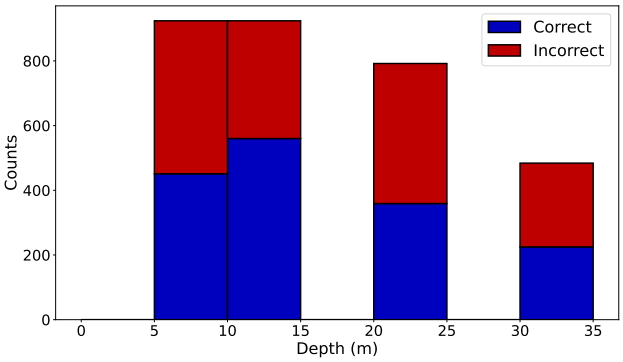
Fig. 8c shows how equal-error-rate SFace performance varies as a function of range. For pre-attentive sensing, error rate remains roughly constant at 50% (chance) at all distances. For attentive sensing, error rate is much lower, ranging from 13% in the near field (5-10m) to 24% in the far field (30-35m).

Table 1 shows the distribution of gallery head poses matched by SFace: Slightly less than half were frontal, with the remaining distributed across the other four head directions.

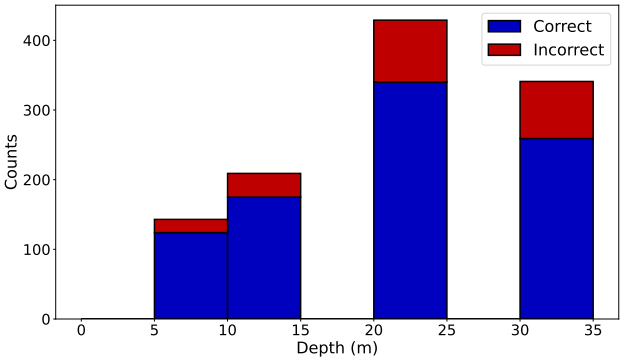
Fig. 11 shows failure modes for SFace in the attentive stream. False negatives often arise due to imperfect focus, while false positives often occur when poses are non-frontal and multiple faces are present.

7. Summary

Our experiments clearly demonstrate the value of attentive sensing for both face detection and face recognition. While it is always possible to demonstrate long-range recognition for a small FoV simply by employing a lens with a long focal length, our attentive vision system is unique in demonstrating long-range face recognition over a *panoramic* FoV, which can be particularly important for social robot applications.



(a) Pre-attentive



(b) Attentive

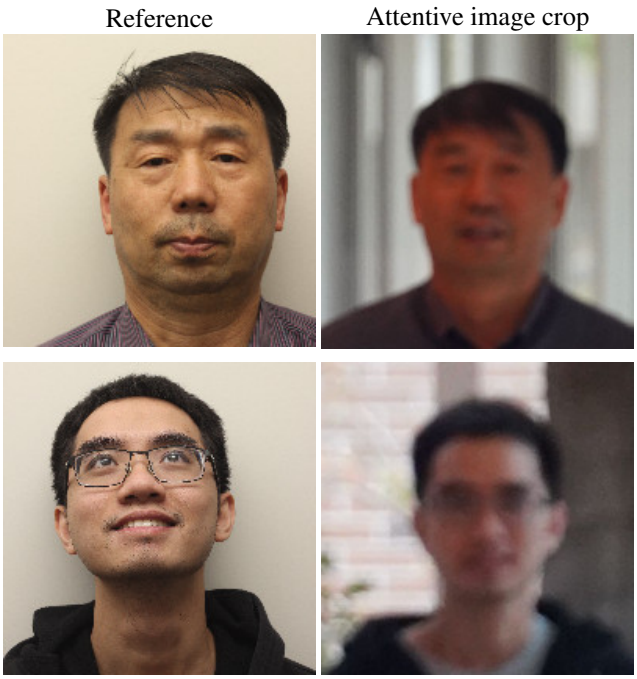
Figure 10: Face recognition performance of SFace as a function of range.

Direction	Attentive	Pre-attentive
Forward	45.1%	46.6%
Up	10.0%	5.9%
Down	12.6%	5.6%
Left	15.4%	19.7%
Right	16.9%	22.2%

Table 1: distribution of gallery head poses matched by SFace.

8. Societal Impacts

As for many technologies, robots with attentive sensing and face recognition capabilities could have both positive and negative impacts on society. Positive application domains may include long-term care, educational environments, security and de-escalation, while negative impacts



(a) False negatives are commonly correlated to imperfect focus, which can happen if another object or person is standing near the target and closer to the camera.

(b) False positives can be caused by a combination of a slightly de-focused image and a partially obstructed face.

Figure 11: SFace failure examples.

could include unwarranted surveillance for political purposes, invasion of privacy and racial profiling. It is important to maintain an open dialogue about this mix of potential benefits and risks to help guide policy as these technologies mature.

9. Future Work

One limitation of our attentive sensor design (Fig. 1) is that the four posts that support the mirror-motor assembly generate small regions of occlusion within the attentive FoV. Interestingly, due to the proximity of these posts to the lens and our substantial lens aperture, the volume of space that is completely occluded by these posts is relatively small, and the posts only get to *shade* objects of interest, rather than occlude them. In future work, we will explore methods for de-mixing the blurred images of the posts from the captured attentive images.

Another limitation of our system is that only the azimuth of gaze is controlled; the gaze elevation is fixed to horizontal. Incorporating even a small degree of gaze elevation control would allow the system to continue to function for small children as well as adults who are sitting or lying down.

Due to bandwidth limitations between the attentive sensor onboard computer and our remote server, we are currently operating our pre-attentive cameras at half-resolution (640×360 pixels). Work is underway to migrate the computer vision algorithms to onboard computing, thus removing the data transfer bottleneck and allowing us to operate the pre-attentive cameras at full resolution.

Currently, acquisition of each attentive image takes 3 seconds, including the time required for the mirror to rotate to its target angle and settle, and for the attentive camera to focus on the face. We are investigating smaller mirror assemblies that will allow faster gaze control and active smooth pursuit to allow real-time tracking of faces when subjects are moving. Updating to an SLR camera that allows more fully-programmable focus would also allow us to focus the camera as we rotate the mirror, based upon the estimated distance to the target face. More directly controlling the camera focus would also allow us to more reliably focus on target faces.

While we do intend to increase the size of the dataset as the project develops, it will remain of modest size due to the ethical need to secure signed consent from all participants.

One of the most interesting questions is how the attentive nature our system might affect public perception of the robot. While it might be perceived as sinister under some circumstances, it might also convey the impression of a more socially-engaged intelligent agent.

References

[1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[2] J.P. Bandera, R. Marfil, A.J. Palomino, A. Bandera, and R. Vázquez-Martín. Visual perception system for a social robot. In *2010 IEEE Conference on Robotics, Automation and Mechatronics*, pages 243–249, 2010.

[3] Chiara Bartolozzi, Neeraj K. Mandloi, and Giacomo Indiveri. Attentive motion sensor for mobile robotic applications. In *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, pages 2813–2816, 2011.

[4] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019.

[5] N. Bellotto, E. Sommerlade, B. Benfold, C. Bibby, I. Reid, D. Roth, C. Fernández, L. Van Gool, and J. Gonzàlez. A distributed camera system for multi-resolution surveillance. In *Proc. of the 3rd ACM/IEEE Int. Conf. on Distributed Smart Cameras (ICDSC)*, Como, Italy, 2009.

[6] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[7] Esther Luna Colombini, Alexandre da Silva Simões, and Carlos Henrique Costa Ribeiro. An attentional model for autonomous mobile robots. *IEEE Systems Journal*, 11(3):1308–1319, 2017.

[8] Esther Luna Colombini and Carlos Henrique Costa Ribeiro. An attentive multi-sensor based system for mobile robotics. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1509–1514, 2012.

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[12] F. Dornaika and J. H. Elder. Image registration for foveated omnidirectional sensing. In *Proc. European Conference on Computer Vision (ECCV)*, pages 606–620, 2002.

[13] J. H. Elder, Y. Hou, S. D. Prince, and M. Sizinstev. Pre-attentive face detection for foveated wide-field surveillance. In *Applications of Computer Vision and the IEEE Workshop on Motion and Video Computing, IEEE Workshop on*, volume 1, pages 439–446, Los Alamitos, CA, USA, jan 2005. IEEE Computer Society.

[14] J. H. Elder, S. J. D. Prince, Y. Hou, M. Sizintsev, and E. Olevskiy. Pre-attentive and attentive detection of humans in wide-field scenes. *International Journal of Computer Vision*, 72(1):47–66, Apr 2007.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Davis King. High quality face recognition with deep metric learning. http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html, 2017. Last accessed 18 Nov 2022.

[17] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[18] Xiaokun Li, Genshe Chen, Qiang Ji, and Erik Blasch. A non-cooperative long-range biometric system for maritime surveillance. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, 2008.

[19] B.G.D.A. Madhusanka and A.G.B.P. Jayasekara. Design and development of adaptive vision attentive robot eye for service robot in domestic environment. In *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 1–6, 2016.

[20] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. *BMVC 2015 - Proceedings of the British Machine Vision Conference 2015*, 2015.

[21] Simon Prince, James Elder, Yunhe Hou, M. Sizinstev, and E. Olevsky. Towards face recognition at a distance. In *Crime and Security. The Institution of Engineering and Technology Conference*, pages 570 – 575, 07 2006.

[22] S.J.D. Prince, J.H. Elder, Y. Hou, M. Sizintsev, and Y. Olevskiy. Statistical cue integration for foveated wide-field surveillance. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 603–610 vol. 2, 2005.

[23] Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. ROS: an open-source robot operating system. In *ICRA Workshop on Open Source Software*, 2009.

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[25] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020.

[26] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014.

[27] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[28] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

[29] Frederick W. Wheeler, Richard L. Weiss, and Peter H. Tu. Face recognition at a distance system for surveillance applications. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, 2010.

[30] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[31] Ting Yu, Ser-Nam Lim, Kedar Patwardhan, and Nils Krahnstoever. Monitoring, recognizing and discovering social networks. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1462–1469, 2009.

[32] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[33] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[34] Yaoyao Zhong, Weihong Deng, Jiani Hu, Dongyue Zhao, Xian Li, and Dongchao Wen. Sface: Sigmoid-constrained hypersphere loss for robust face recognition. *IEEE Transactions on Image Processing*, 30:2587–2598, 2021.

[35] Xuhui Zhou, Robert T. Collins, Takeo Kanade, and Peter Metes. A master-slave system to acquire biometric imagery of humans at distance. In *First ACM SIGMM International Workshop on Video Surveillance*, IWVS '03, pages 113–120, New York, NY, USA, 2003. Association for Computing Machinery.