Recognizing and Remembering Individuals: Online and Unsupervised Face Recognition for Humanoid Robot

Lijin Aryananda

Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA, lijin@ai.mit.edu

Abstract

Individual recognition is a widely reported phenomenon in the animal world, where it contributes to successful maternal interaction, parental care, group breeding, cooperation, mate choice, etc. This work addresses the question of how one may implement such social competence in a humanoid robot. We argue that the robot must be able to recognize people and learn about their various characteristics through embodied social interaction and thus proposed an initial implementation of an online and unsupervised face recognition system for Kismet, our sociable robotic platform. We show how specific features of this particular application drove our decision and implementation process, challenged by the difficulty of the face recognition problem, which has so far been explored in the supervised manner. Experimental results are reported to illustrate what was solved and lessons learned from the current implementation.

1. Introduction

The ability to recognize and remember individuals is crucial for complex interactions among social animals, such as preferential treatment, cooperative behavior, and reciprocity. Mantis shrimps have been observed to avoid empty cavities with the odor of those that have defeated them in the past, but enter those with the odor of individuals they have beaten. Caldwell [2] concluded that they're able to not only recognize other individuals, but also remember their reputation as fighters.

This paper addresses the question of how one may implement in a humanoid robot the ability to learn to recognize and remember things about people it interacts with. Such social competence leads to complex social behavior, such as cooperation, dislike, loyalty, affection, and attachment. As proposed by Dautenhahn [6], if robots have long-term contact with humans, it may be desirable to have them develop individual relationships, which is exactly the aftermath of this social dynamic. Moreover, the ability to distinguish among people allows the robot to build toward more complex social competencies where the idea of people as distinct individuals is crucial, including theory of mind and social referencing.

In order to recognize and remember characteristics of various people, the robot must be able to identify people and link contextual information (past behavior, beliefs, habits, affiliations, etc) with recognition memory of one's appearance. A substantial amount of research has been carried out in person identification technology [4, 9, 17]. Most of these works attempt to solve the identification problem: given a set of labeled training data and a set of test data, find the correct person label for the test data. We would like to focus and draw attention to the acquisition process of the training data. In most existing face recognition systems, the training images are collected and labeled manually. We argue that the contextual knowledge about other people that we acquire through our daily social experience is so rich and complex that manually encoding it into a database along with the corresponding person's facial images for the robot to memorize is very limiting. We propose that the robot must collect training data for learning about people's various characteristics through embodied social interaction, thus directly perceiving the richness of the environment.

As an initial attempt toward this goal, we implemented an online and unsupervised face recognition system, where the robot opportunistically collects, labels, and learns various faces while interacting with people (without any staged introduction session), starting from an empty database. We will show how specific features of this particular application drove our decision and implementation process, challenged by the difficulty of the face recognition problem, which has so far been explored in the supervised manner.

In the rest of the paper, we describe related works and briefly describe Kismet, our robotic platform. We then outline design issues and present an implementation of an online and unsupervised face recognition system for our robot, using the eigenfaces technique [5]. Experiments were performed to examine system behavior. We report details on experiment setting and results. Lastly, we introduce future works to implement other modalities and precursors based on the lessons learned from this implementation.

2. Related Work

Research in person identification technology has recently received significant attention, due to the wide range of biometric, information security, law enforcement applications, and Human Computer Interaction (HCI). Face recognition is the most frequently explored modality and has been implemented using various approaches [5]. A combination of face and body recognition has been proposed [9]. Speech recognition has also been widely investigated [17]. The use of multiple modalities has been observed [7,8].

In an image and video indexing work [16], using a neural network based face detector, extracted faces are grouped into clusters by a combination of a face recognition method using pseudo two-dimensional Hidden Markov Models and a k-means clustering algorithm. The number of clusters is specified manually. In contrast to this work, we require the recognition system to perform in an automatic and unsupervised manner. Thus, the number of clusters i.e. the number of individuals interacting with the robot at any given time is unknown.

3. Robotic Platform

Kismet is an expressive robotic creature with perceptual and motor modalities tailored to natural human communication channels. Infant-level social competencies have been implemented on our robot, Kismet, as building blocks for exploring socially situated learning between Kismet and its human caregiver [1].

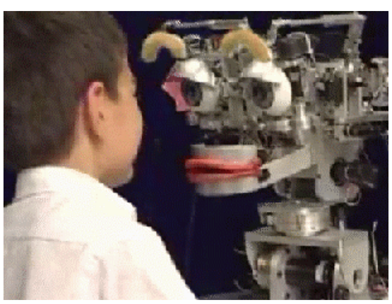

Figure 1: *Kismet is an expressive robotic creature designed for natural social interaction with human [1].*

To interact with its caregivers, Kismet uses four color CCD cameras and an unobtrusive wireless microphone. All cameras move with respect to the head. The positions of the neck and eyes are important both for expressive postures and for directing the cameras towards behaviorally relevant stimuli. The hardware and software control architectures have been designed to meet the challenge of real-time processing of visual signals (approaching 30 Hz) and auditory signals (frame size of 10 ms) with minimal latencies (500 ms). Kismet's vision system is implemented on a network of nine 400 MHz commercial PCs running the QNX real-time operating system. Kismet's emotion, behavior, and expressive systems run on a collection of four Motorola 68332 processors. The speech module, including synthesizer and speech processing software runs on Windows NT and Linux.

4. Design Issues and Considerations

4.1 Performance Criteria

Current state of the art in face recognition technology allows for a recognition accuracy of 95% on more than 1000 frontal mugshot-like images when taken in the same day and 80% when taken with a different camera and lighting condition [11]. Our performance criteria, however, is less ambitious in terms of recognition accuracy per image. Our goal is for the system to be able to consistently recognize and remember people who are relevant to Kismet and interact with it on a regular basis. In the same way, we also do not remember every single person we pass on the street.

4.2 Failure Modes

We consider two possible failure types: clustering error and failure to learn to recognize a person despite frequent encounters. While some errors are unavoidable, it is less harmful to place an individual's faces into multiple clusters than it is to cluster multiple individuals into one class in the training set. In the latter case, the robot will constantly treat multiple people as the same person and any effort to learn additional characteristics of these people will be misleading.

4.3 Eigenface Method

We decided to implement the face recognition module using the eigenface technique [5] because it is widely implemented and well known for its simplicity and computational efficiency. We plan to explore other types of representations as well. The eigenface method uses an information theory approach of coding facial images, where it attempts to find the principal component of the distribution of faces, or the eigenvectors of the covariance matrix of the set of face images.

Observation of the recognition system's performance on a random sample of images indicates similar findings to [12] and [13]. Performance is highly dependent on correlation between face alignment in the training and test set. Also, performance degrades in the presence of facial expressions and changes in scale. In [13], subjects were requested to minimize head motion in order to ensure proper alignment. This is not an option in our case because imposing restrictions on subjects' posture and expression will greatly restrict social interactions with the robot.

5. Implementation

As shown in figure 2, visual input from the camera is passed as input to the face detector system [14]. Face detector sends information regarding location of faces found within the robot's visual range, if any, for further pre-processing. Based on the training set, the pre-processed image is then recognized to generate an initial hypothesis[1]. After the recognition process, post-processing may be performed on the test image in order to either include unknown individuals or incrementally improve the existing training set.

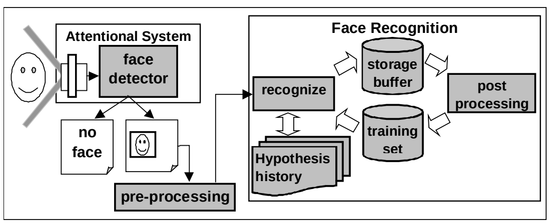

Figure 2*: The schematic of the online and unsupervised face recognition system. The system receives video stream as input and learn to recognize individuals in an online and unsupervised manner.*

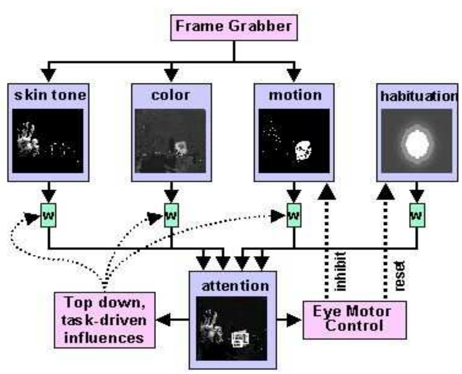

Figure 3*: Kismet's visual attention system [15] picks out low-level perceptual stimuli (highly saturated colors, motion, face, and skin tone) that are particularly salient and direct the robot's attention to gaze toward them.*

5.1 Attentional System: An Interface to the Environment

Kismet's attention system acts to direct computational and behavioral resources toward salient stimuli and to organize subsequent behavior around them [15]. As shown in figure 3, the pre-attentive system processes information about basic visual features across the entire visual field. These low-level features influence the direction of Kismet's gaze. This greatly simplifies the interface between the face recognition system and the environment. The face recognition system is simply activated when a face is detected within the visual field. If the face is the most salient stimulus at the time, the cameras will track it, maintaining it within the visual field.

Face Detector. Given periodic camera input, the face detector outputs the location of existing faces, if any. Clearly, a robust, accurate, and fast face detector is key in this implementation. We used a frontal face detection system developed by [14] which satisfies both the performance and real-time criteria on our slightly different environment: 128 by 128 pixel images on 800 MHz PCs. Since both the camera and target move around, the face detector's output is relatively noisy. There is a large variation in distance, viewing angle, and lighting. For higher detection accuracy, we utilize the skin tone detector and accept the detected face region if it contains enough skin color.

5.2 Pre-processing

In order to minimize variations generated by the noise in the environment, each face image found within the visual range is pre-processed. Each image is scaled to a 40x40 greyscale pixel image. A simple alignment procedure is applied to roughly correct off-center faces (only effective on frontal images). Each pixel in the image is normalized in order to alleviate lighting variations. Lastly, each face image is masked using 2-D Gaussian, thereby diminishing background variations.

5.3 Face Recognition

As mentioned, we are particularly concerned with the shortcoming of storing a set of manually labeled faces for the robot to learn without grounding them to direct sensory experience. We also would like to avoid a *staged* introduction session where individuals present their faces and provide some labels to the robot. This poses a requirement for our recognition system to learn in an online and unsupervised manner starting with an empty training set, which translates into the following tasks: 1. Given a sequence of unlabeled data containing facial images of an unknown number of individuals (produced by face detector), we must implement a somewhat reliable process to cluster them. 2. The resulting clusters may then be used as training data. 3. Once a training set is formed, each new input is classified as either a known class or new individual. 4. The system must decide when to include these new individuals into the existing training set. Details of the recognition process are described below.

---

[1] The face recognition was adapted from Turk and Pentland's eigenface-based face recognition system [5].

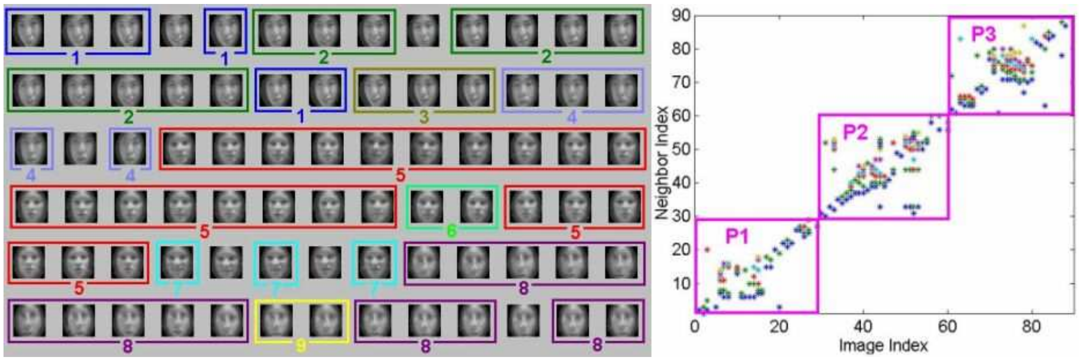

Figure 5*: A sample heuristic clustering process with 3 individuals in an input batch. Image 0-28 belongs to person 1. Image 29-59 belongs to person 2. Image 60-87 belongs to person 3. On the left are all images and the resulting nine clusters. In the right graph, the x-axis represents each image and the y-axis represents other images in the set that are within the threshold value Dc to it.*

Pre-Training-Set Phase. Without any training set, the recognition system employs a heuristic clustering method to examine incoming input data, containing facial images of an *unknown* number of individuals and cluster them. Essentially, the system collects input data into batches and for each batch, the system iteratively holds out each image and computes the eigenfaces as well as each coefficient vector of the remaining images (treating each image as a class). Figure 4 shows each input's distance to roughly estimated face space and distance to its closest neighbor in the batch, calculated from several initial batches. Threshold values for the maximum allowable distance to the face space (Ds) and a known face class (Dc) are determined empirically based on the assumptions that most of the input images contain faces and there are a few images per individual inside each batch.

Figure 5 illustrates a sample heuristic clustering process where a batch containing facial images of three individuals is separated into eight clusters. The issue of dealing with multiple clusters per individual will be addressed in the next section. As shown in the lower graph, no two individuals are mistakenly placed in the same cluster. Again, the clustering process works without knowing how many individuals there are in an input batch. Thus, we do not need to notify the recognizer whenever an individual comes and leaves, allowing the recognizer to perform in an automatic and unsupervised manner. In the current setting, cluster #5 is considered big enough to be placed into the training set as a new class. The minimum threshold value for cluster size is empirically determined based on the predictions that small clusters are more likely to contain non-representative face images of an individual and large clusters of similar images are difficult to come by due to moving camera and subjects.

Online Training Phase. Initially, the training set is quite small, consisting of only a few individuals. Each input image is projected into the eigenface basis and the distance between its coefficient vector to the face space and each known class is calculated. Image is immediately discarded if its distance to the face space > Ds. If its distance to a known class < Dc, it is classified as the corresponding individual. Otherwise, it is perceived as an unknown. Similar to the pre-training phase, each unknown image is collected in batches for post-processing, where the same clustering procedure is applied (distributed across several processors).

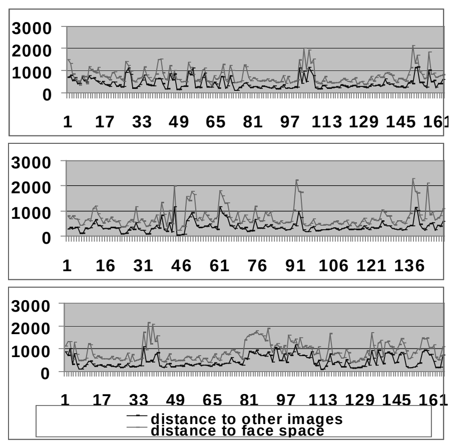

Figure 4: *Data analysis during pre-training phase. The x-axis represents each image in an input batch, taken from an interaction sequence. The y-axis represents each image's distance to the estimated face space and distance to its closest neighbor.*

If a large enough cluster is found, it goes through another processing step which essentially calculates the average of distances among images in the cluster and the average of distances among the newly clustered images combined with images in each known class in the existing training set. The cluster is then either added as a new individual or into an existing class depending on how different it is from known individuals. In some cases, the system decides to ignore the newly clustered images if the difference between the calculated averages is not significant enough.

Figure 7*: The sequence of system output for P9. Grey = unknown. Black = not a face. White column (of length x) = a known individual x.*

This method allows the system to learn to recognize new individuals and incrementally improve its training images over time. The more variations in expressions, lighting, and pose there are in an individual's training set, the better the recognition performance is. Thus, the more one interacts with the robot, the more likely it is for the system to obtain their snapshots with various expressions, pose, etc. However, if one just happens to pass by, they would be represented poorly in the database and thus, not easily recognized.

6. Experimental Results

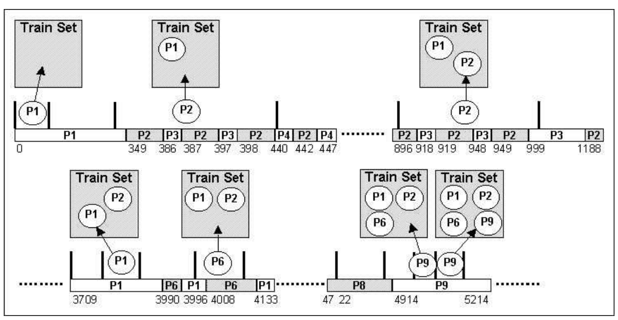

Figure 6*: Experimental results with nine subjects. Number below each bar indicates the starting image index of each individual sequence. The thick vertical lines above the sequence bars indicate the start and end of batches collected for post-processing. Note that in the third addition to the training set, since the training set already 'knows about' P2, the new P2 cluster is simply merged into the existing P2 class in the training set.*

An experiment was carried out in order to evaluate system performance. Nine subjects were asked to play with the robot and interaction time was not regulated. Table 1 shows the number of images received per individual. Not all images are processed because the system filters out images that are too small and too large. Interaction mode is either 1 (one on one) or x > 1 (x people concurrently).

Table 1: *Experiment Data Statistics Per Individual*

	P11	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**
#images	803	1842	1201	228	282	246	34	498	575
mode	1&2	4	4	1&4	4	2	1	1	1

Figure 6 illustrates the results of this experiment. The system was able to learn about 4 (P1,P2,P6,P9) out of 9 subjects. The rest of the subjects were never learned, meaning their images were never put in the training set. No learning error took place, meaning the system never clusters one individual into multiple clusters and vice versa. The system failed to learn about P3 despite lengthy interaction sequences. Note that due to limited space, we only include interesting interaction sequences where new data is added into the training set.

Figure 7 illustrates the detailed output sequence produced while the robot interacts with P9. Essentially, output is either unknown or not a face until a large enough cluster of P9's faces was placed in the training set. From this point on, the system starts to recognize P9.

7. Discussion and Future Work

Experimental results show that the system exhibits a potential to incrementally learn to recognize a few people's faces. More extensive testing is to be done on more subjects to further verify this claim. Results also indicate that the system behaves very differently across individuals. Observations of interaction sessions indicate that there is a large variation in the way people interacts with the robot. Some people like to move around; others like to use toys to attract Kismet's attention. These variations greatly impact the behavior of the system.

As mentioned above, the system failed to learn about P3 despite lengthy encounters. Good clusters of P3's images were actually found but never added into the training set because the system could not reliably decide whether or not P3 was the same person as P1 or P2. Ideally, these ignored clusters should be stored such that the system will be able to go back to them after collecting additional data. This may be analogous to priming in human's cognitive process, which facilitates memory performance on the basis of having a single prior exposure to stimuli that can enhance it.

Improvements have been made such that the current system runs faster, handles concurrent interactions, and deals better with longer interaction sequence compared to our previous implementation [3]. Consequences include larger post-processing batch size, more realistic testing scenarios, and better tuning of threshold values.

So far, we have not paid much attention on the system's final hypothesis about who is currently in front of the robot. In figure 7, we have only shown that the system is able to recognize P9 once the system trained on his images, but still considers him as an unknown about half the time. Our hope is that by incrementally improving each individual's training images, the recognition process will be more accurate. Moreover, a better alignment procedure and the use of other modalities (i.e. speaker recognition, sound localization, and gaze detection) should significantly improve recognition performance.

Lastly, an immediate extension of this work is to implement the ability to correlate simple contextual information about individuals along with their faces. For example, the robot may learn to correlate its emotional state with the presence of certain individuals or remember individual's favorite toy and frequently used words.

8. Conclusion

The ability to distinguish among different individuals is crucial for all social animals. We have implemented an online and unsupervised face recognition system to address the issue of how one may implement such social competence in a humanoid robot. Experiments have been performed and results indicate the system is capable of learning to recognize a few individuals interacting with the robot. Lessons learned and future directions are discussed.

9. Acknowledgement

This work was funded by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" project under contract number DABT 63-00-C-10102. The author gratefully acknowledges Paul Viola and Michael Jones for their assistance in porting their face detector to Kismet. We would also like to acknowledge usage and include the copyright notice of Turk and Pentland's face recognition system (Copyright 1992, Massachusetts Institute of Technology. All Rights Reserved).

Reference

[1] C. Breazeal, Sociable Machines: Expressive Social Exchange Between Humans and Robots. Sc.D. dissertation, Department of Electrical Engineering and Computer Science, MIT (2000).
[2] Griffin, Donald R, Animal Minds. Chicago: the University of Chicago Press (1992).
[3] Aryananda, L. Online and Unsupervised Face Recognition for Humanoid Robot: Toward Relationship with People. Proceedings of the 2001 IEEE-RAS International Conference on Humanoid Robots (2001).
[4] W. Zhao, R. Chellappa, A. Rosenfeld, P. Phillips, Face Recognition: A Literature Survey.
[5] M. Turk, A. Pentland, Eigenfaces for Recognition. Journal of Cognitive Neuroscience, Vol. 3 No. 1, pp. 71-86 (1991).
[6] K. Dautenhahn, Getting to Know Each Other - Artificial Social Intelligence for Autonomous Robots. Robotics and Autonomous Systems 16, pp. 333-356 (1995).
[7] R. Brunelli, D. Falavigna, Person Identification Using Multiple Cues. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-17, No. 10 (1995).
[8] J. Kittler, Y. Li, J. Matas, M. Ramos Sanchez, Combining Evidence in Multimodal Personal Identity Recognition Systems. International Conference on Audio and Video-based Biometric Person Authentication, Switzerland (1997).
[9] C. Nakajima, M. Pontil, B. Heisele, T. Poggio, People Recognition in Image Sequences by Supervised Learning. A.I. Memo No. 1688, C.B.C.L. Paper No. 188. MIT (2000).
[10] J. Weng, C. Evans, W. Hwang, An Incremental Learning Method for Face Recognition under Continuous Video Stream. Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France (2000).
[11] A. Pentland, T. Choudhury, Personalizing Smart Environments: Face Recognition for Human Interaction. IEEE Computer. Special issue on Biometrics (2000).
[12] G. Sukthankar, Face Recognition: A Critical Look at Biologically-Inspired Approaches. http: //www.cs. cmu.edu/~gitars/16-721/final/final.html (1999).
[13] Y. Yacoob, H. Lam, L. Davis, Recognizing Faces With Expression. International Workshop on Automatic Face and Gesture Recognition, Zurich (1995).
[14] P. Viola, M. Jones, Robust Real-time Object Detection. Technical Report Series, CRL 2001/01. Cambridge Research Laboratory (2001).
[15] C. Breazeal, B. Scasselati, A Context-dependent Attention System for a Social Robot. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, pp. 1146-1151. Stockhold, Sweden (1999).
[16] S. Eickeler, F. Wallhoff, U. Iurgel, G. Rigoll, Content-based Indexing of Images and Videos using Face Detection and Recognition Methods. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City , Utah (2001).
[17] S. Furui, An Overview of Speaker Recognition Technology. ESCA Workshop on Automatic Speaker Recognition Identification Verification, Switzerland, 1-9 (1994).