Attending to Learn and Learning to Attend for a Social Robot

Lijin Aryananda
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
Email: lijin@csail.mit.edu

Abstract— **Our motivation is to create a robotic creature, Mertz, that 'lives' among us daily and incrementally learns from and about people through long-term social interaction. One of Mertz's main tasks is to learn to recognize a set of individuals who are relevant to the robot through ongoing human-robot interaction. We present an integrated framework, combining an object-based perceptual system, an adaptive multi-modal attention system and spatiotemporal perceptual learning, to allow the robot to interact while collecting relevant data seamlessly in an unsupervised way. Our approach is inspired by the coupling between the human infants' attention and learning process. We implemented a multi-modal attention system for the robot that is coupled with a spatiotemporal perceptual learning mechanism, which incrementally adapts the attention system's saliency parameters for different types and locations of stimuli based on the robot's past sensory experiences. We conducted and described results from a six-hour experiment where the robot interacted with over 70 people while collecting various data in a public space.**

I. Introduction

Our motivation is to create a robotic creature, Mertz, that 'lives' among us daily and incrementally learns from and about people through long-term social interaction. One of Mertz's main tasks is to learn to recognize a set of individuals who are relevant to the robot through ongoing human-robot interaction. This life-long developmental approach and social interaction in robotics have been widely explored[1], [2].

In this paper, we present an integrated framework, combining an object-based perceptual system, an adaptive multi-modal attention system and spatiotemporal perceptual learning, to allow the robot to perform the following tasks automatically and seamlessly in an unsupervised way:

1) operate for long periods of time in public spaces
2) interact visually and verbally with multiple passersby at a time
3) filter and collect relevant audiovisual sensory data
4) generate clusters of face and voice data for each individual
5) generate clusters of color histograms for a set of objects
6) generate clusters of frequently heard words and learn a simple bigram language model
7) learn spatiotemporal patterns of various audiovisual sensory events

In this setup, where there is no boundary between testing and training stages, the robot has to perform the parallel task of interacting with while collecting data and learning from the environment. This task is difficult for a number of reasons. Firstly, the robot's attention system faces conflicting tasks, as it has to be reactive to find learning targets in the environment but also persistent to observe targets once they are found. In the human's visual attention system, this dichotomy is reflected in two separate components: the bottom-up (exogenous) and top-down (endogenous) control [3].

The importance of an attention system for learning has been discovered in many research areas [4], [5]. Incorporating top-down control of attention has also been explored in [6], [7], [8]. However, the top-down attention control was mostly simulated manually in most of these systems. Many properties of the robot's attention system that we implemented were inspired by the Sensory Ego-sphere [7].

Secondly, attending to learn in an unconstrained social environment is a difficult task due to noisy perceptual sensors, target disappearing and reappearing, presence of multiple targets, and the target's or robot's own motion. Same person tracking in subsequent frames is an easy task for the human's visual system since we are very good in maintaining spatiotemporal continuity. Even when our heads and eyes move, we can easily determine what have moved around us. Unfortunately, for an active vision system, this is not the case. The robot essentially has to process each visual frame from scratch in order to rediscover the learning target from the previous frame. Tracking a person's face in order to learn to recognize the person is a somewhat convoluted problem. The robot has to follow and record the same person's face in subsequent frames, which requires some knowledge about how this person looks like, but this is exactly what the robot is trying to gather in the first place.

An additional complexity is introduced by the trade-off between timing and accuracy requirement of the interaction and learning process. The interaction process needs fast processing to allow for timely responses. The data collection process needs higher accuracy in terms of ensuring that the robot is collecting the correct data for the right person or object. Interestingly, this dichotomy is also reflected in the separate dorsal *'where'* and ventral *'what'* pathways in the human's visual system, for locating and identifying objects.

We have designed the robot's attention system to address some of the issues mentioned above, by incorporating object-

based tracking and an egocentric multi-modal attentional map based on the world coordinate system [9], [7]. The attention system receives each instance of object-based sensory events (face, color segment, and sound) and employs space-time-varying saliency functions, designed to provide some spatiotemporal short-term memory capacity in order to better deal with detection errors and having multiple targets that come in and out of the field of view. We also implemented two separate face trackers in order to cater to the needs of both interaction and data collection. One actively affects the attention system in real time, while the other runs independently to collect face data at a slower speed.

In addition, inspired by the coupling between the human infants' attention and learning process, we implemented a spatiotemporal perceptual learning mechanism, which incrementally adapts the attention system's saliency parameters for different types and locations of stimuli based on the robot's past sensory experiences. In the case of human infants, the attention system directs cognitive resources to significant stimuli in the environment and largely determines what infants can learn. Conversely, the infants' learning experience in the world also incrementally adapts the attention system to incorporate knowledge acquired from the environment. Coupling the robot's attention system with spatiotemporal perceptual learning allows the robot to exploit the large amount of regularities in the human environment. For example, in an indoor environment, we would typically expect tables and chairs to be on the floor, light fixtures to be on the ceiling, and people's faces to be at the average human height.

We conducted a six-hour experiment where the robot interacted with over 70 people in a public space. We evaluated how the robot directed its attention among competing stimuli and collected face, color, and speech data for future recognition during this experiment. At this time, the adaptation feedback from the spatiotemporal learning module to the attention system has been implemented in a minimal way. We present some preliminary results on how the robot's past experiences, which suggest where faces, color segments, and sound tend to appear allowed the attention system to favor certain spatial regions. In the last section, we discuss some future implementation plans to extend this framework to explore various behavioral adaptations based on the robot's past sensory experiences.

II. Implementation

A. Robotic Platform

MERTZ is an active-vision head robot with thirteen degrees of freedom (see Figure 1). Mertz has been designed with the goal of continuous long-term operation in various human spaces, as reported in [10]. The robot perceives visual input using a Point Grey Dragonfly digital camera per eye. The robot uses an Acoustic Magic array desk microphone to allow multiple people to speak to the robot. The robot is mounted on a portable wheeled platform to allow for experiments in different locations.

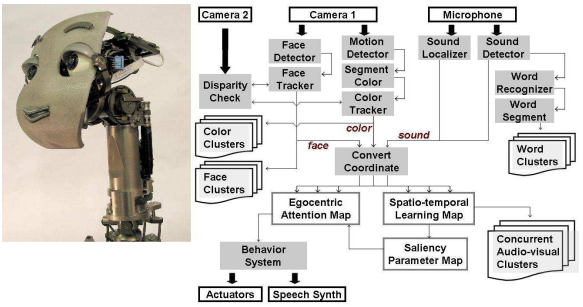

Fig. 1. Left: MERTZ, an active vision humanoid head robot with 13 degrees of freedom. Right: The robot's overall system architecture.

B. System Architecture

Figure 1 illustrates the robot's system architecture. The robot's visual system is equipped with detectors and trackers for relevant stimuli, i.e. people and colored objects. A large part of the robot's visual system was implemented using the YARP library [11]. The auditory system detects, localizes, and performs various processing on sound input. Each instance of face, color segment, and sound event is projected onto the world coordinates system using the robot's forward kinematic model and entered into both the egocentric attention and spatio-temporal learning map. The spatio-temporal learning process incrementally updates the attention's saliency parameters, which is then fed back into the attention map. The egocentric attention map's target output is passed onto the robot's behavior system to calculate the appropriate next step. In parallel, each perceptual event is also processed to generate clusters of individual's faces, color segments, and words for future recognition.

C. Face Processing

The robot is using a frontal face detector [12], complemented by feature tracking. We are combining a KLT tracker [13] for faster attentional processing with a SIFT tracker for slower but more accurate generation of clusters of individual's faces.

D. Color Processing

The robot detects colored objects by looking for moving color segments within some distance. First, the robot detects motion using [14], where a KLT tracker is used to estimate displacement of background pixels due to robot's own motion. Detected motion patch is then used to activate a color-histogram based tracker. A color histogram model is built using color segments tracked in each continuous sequence and stored into a color cluster database.

E. Auditory Processing

The robot's auditory system was implemented using CMU Sphinx 2 [15]. An energy-based sound detection module determines the presence of sound events above a threshold. We also use the five indicator LEDs on the microphone to obtain the horizontal direction the sound source. Each recorded segment is then processed for word recognition. One or two

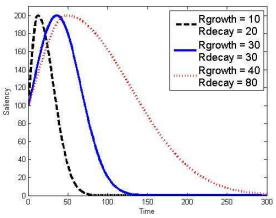

Fig. 2. The attention's system's saliency function for varying growth and decay rates

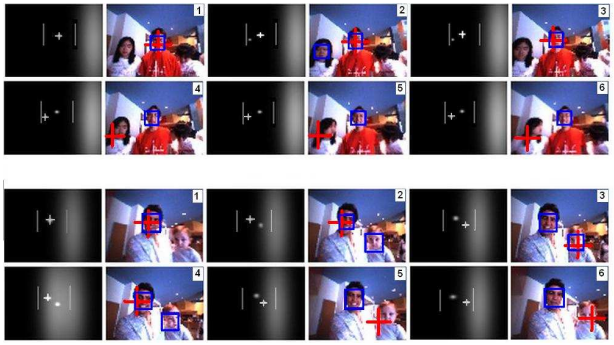

Fig. 3. Two sample image sequences and the corresponding attentional map, illustrating the attention system's output while interacting with two people simultaneously.

consecutive words are then selected to construct a simple bigram model, which generate a one or two-word response.

F. Egocentric Object-based Attention System

The robot's attentional map is a 2D rectangular plane, which is an approximated projection of the front half of the geodesic sphere centered at a robots origin (a simplified version of the Sensory Ego-Sphere implementation [7]). This plane consists of 280x210 pixels, indexed by the azimuth and elevation angles of the robot's gaze direction. The retinal location of each perceptual event is projected onto the attentional map's world coordinates using the robot's forward kinematic model. Each object is represented spatially in the attentional map by a 2D unnormalized gaussian, centered at the object's center, sized at $3 * \sigma$ = object's diameter, and initially scaled with magnitude of 100. The magnitude of this spatial 2D gaussian represents the saliency value of the corresponding object. If the same object is successfully tracked during the subsequent frame, the location of the corresponding gaussian is updated accordingly and the magnitude is modified using a time-varying saliency function:
$f(t,x,y,fov,p_{type}) = M_{max}\ e^{-(t-t0)^2/2R(x,y,fov,p_{type})^2}$,
t=time, x,y=location, fov= 1 if inside, 0 if outside field of view, p_{type}=percept type, M_{max}=200, $t0$=start time.
$R = R_g$ (growth rate) for $t < t_{peak}, t_{peak} = t0 + \sqrt{-2R_g^2 log(M_{init}/M_{max})}, M_{init}$=100. $R = R_d$ (decay rate) for $t > t_{peak}$. If the target object is outside the field of view (fov=0), $f(t,x,y,fov,p_{type})$=the last saliency value at time t-1, i.e. does not grow or decay.

The function $R(x,y,fov,p_{type})$ essentially determines the growth or decay rate parameter for a particular sensory input of type p_{type} and located at x,y. Initially, both $R_g(x,y,fov,p_{type})$ and $R_d(x,y,fov,p_{type})$ are set to 30 for all x,y, and p_{type}. As the robot gains experience in the environment, the spatio-temporal learning system incrementally updates both R_g and R_d for each type of perceptual input and its location in the egocentric map. If an object has not been tracked for some period, it is considered lost and its corresponding gaussian is then deactivated by setting $R_d = 0.2$.

Figure 2 illustrates the saliency function for varying values of saliency growth rate (R_g) and decay rate (R_d). The idea is that if a face or color segment is detected and subsequently tracked, its saliency value will initially grow and start decaying after a while. The saliency growth rate determines how good a particular stimuli is in capturing the robot's attention and the decay rate specifies how well it can maintain the robot's attention. The time-varying saliency functions and interaction among these functions for multiple sensory events generate a number of advantages. Firstly, since each object has to be tracked for some time to achieve a higher saliency value, the system is more robust against short-lived false positive detection errors. It also deals better with false negative detection gaps. The combination of decay rates and egocentric map's short-term memory provides some short-term memory capabilities to allow the robot to remember objects even if they have moved outside the robot's field of view. Moreover, the emergent interaction among various saliency functions allows the attention system to integrate top-down and bottom-up control and also to naturally alternate among multiple learning targets. Lastly, the system architecture provides natural opportunities to detect various spatio-temporal and multi-modal correlation in the sensory data. The incremental adaptation of the saliency parameters based on these observed patterns allows the attention system to be more sensitive to a set of previously encountered learning target types and locations.

Figure 3 shows two sample sequences of the attentional map output. On each attention map (left column), the two vertical lines represent the robot's current field of view. Two people were interacting with the robot. The blue box superimposed on the image indicates detected faces. The red cross indicates the current attention target. Once a person's face is detected, it is represented by a white blob in the attentional map, with time-varying intensity level determined by the saliency function described above. Thus, the blob often remains in the map even if the face is no longer detected for some time, allowing the robot to still be aware of a person despite failure in detecting his or her face. In the upper sequence, the female's face was detected only in frame 2, but was still present in the map in frame 3-6. Similarly, in the lower sequence, the infant's face was detected in frame 2-4 and remains in the map for the rest of the frames. Moreover, as shown in both sequences, after attending to the first person, the attention system switches to the second person after some time due to the temporal

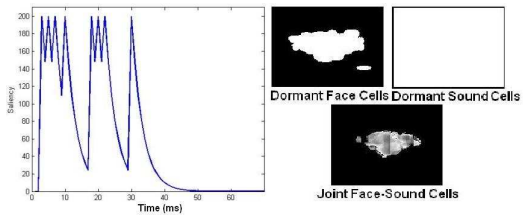

Fig. 4. Left: An example of the spatio-temporal learning system's activity function when an object is entered at time t=3,5,7,10,18,20,22,30 ms. Right: Sample dormant and joint occurence of face and sound map, constructed by the spatiotemporal learning module.

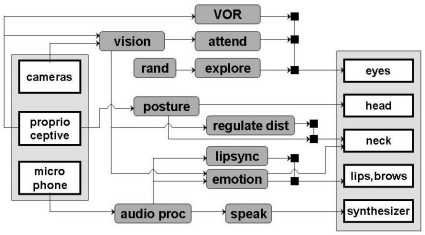

Fig. 5. The robot's behavior-based controller.

interaction among each blob's saliency function. In both upper and lower sequences, this attention switch from the first person to the second person in frame 4 and 5 respectively.

G. Spatio-temporal Sensory Learning

The spatio-temporal sensory learning map is very similar to the egocentric attention map, i.e. a 2D rectangular plane with 280x210 pixels, indexed by the azimuth and elevation angles of the robot's gaze direction. Each map pixel is a storage space containing up to three cells, one for each perceptual type (face, color, and sound). Each cell $C_{x,y,p_{type}}$ is associated with a time-varying activity function $A(t) = Me^{-D*(t-t_{start})}$,$t$=time, M=200, D=0.3. Initially, all cells are empty ($A(t)$=-1). The spatio-temporal learning system receives the same sensory input as the egocentric attention system, plus the object's velocity. Whenever an object of type p_{type} at time T' and location L is entered, cell $\mathrm{C}_{x,y,p_{type}}$ is activated by setting $t_{start} = T'$ which inserts a spike of magnitude M. Figure 4 Left illustrates a sample sequence of a cell's activity level when an object is entered at time t=3,5,7,10,18,20,22,30 ms. If a cell has not been activated for about 2 seconds, its activity level decays to 0 and the cell becomes dormant until activated again.

Using this simple mechanism, the map can be used to record various spatiotemporal pattern in the sensory input. Dormant cells provide spatial history of each perceptual type, which allows the robot to learn where faces typically occur, etc. We also collect joint occurrence statistics among different perceptual types within each cell. Figure 4 Right illustrates maps of all dormant face and sound cells and a sample joint occurence map showing regions in the robot's entire observable space, where face and sound events have frequently occurred together in the past. This joint occurence information is also used to generate clusters of concurrent pairs of audio-visual input, containing a set of face or color segment images and a number of temporally correlated sound files. The map also computes a measure of the average dynamic and presence duration for various input types and locations. At this time, for each input type and attentional map location, R_g and R_d are updated based on a simple function of the accumulated history of the corresponding input's occurence frequency, joint occurrence frequency, dynamic, and presence duration. We plan to explore more complex feedback strategies in the near future.

H. Behavior System

We have implemented the robot's behavior based control system in CREAL (CREAture Language) [16], designed to implement behavior based programs (see Figure 5. The emergent interaction among these modules generates the following high-level behaviors: random exploration when nothing seems interesting, orient gaze to interesting stimuli (faces, color segments, and sound), maintain interaction distance by approaching or pulling back, display facial expression and posture based on a simple stimuli-based emotion system, and utter a single or pair of words generated by a simple bigram language model, that is learned from the word sequence input. We have shown in [10] that with these perceptual and behavior mechanisms, the robot was able to engage in interaction with a large number of passersby in a set of public locations.

I. Face and Voice Clustering for Recognition

One of the robot's main goals is to recognize individuals in an unsupervised way through natural social interaction. Face recognition, especially when unsupervised, is still a very difficult problem. We are utilizing accurate face tracking and multi-modal sensory binding to simplify this task. As mentioned above, we are using a SIFT-based [17] same-person face tracking which provides automatic generation of face clusters from each continuous tracking session. Since the tracking process terminates as soon as the robot loses track of the face, we are using the same tracking technique to perform furthering offline clustering to merge face clusters that belong to the same individual. The spatio-temporal binding process between faces and sounds, as described above, complements these face clusters with speech samples from the corresponding individuals. This will essentially allow us to rely on existing supervised face and speaker recognition technology.

III. Experimental Results

We conducted an experiment to evaluate the system as the robot interacts with passersby for 6 continuous hours. The robot was placed at the MIT Stata Center building lobby, with a poster requesting for people to interact with the robot. A set of colored toys were placed next to the robot. People were free to approach the robot at anytime and we observed from a distance to minimize interaction constraints.

Figure 6 Left illustrates various detection and attention output obtained during the experiment. We recorded data in 55097 frames which span a total of 17351.5 seconds. In order

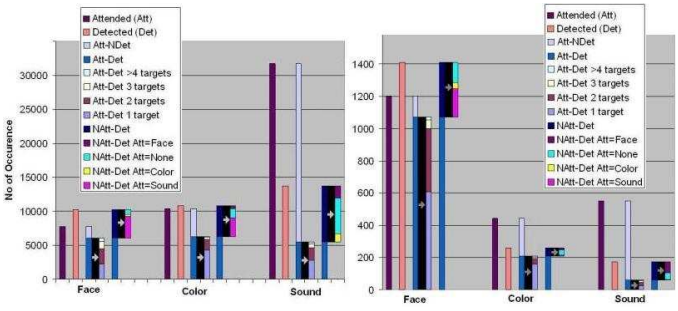

Fig. 6. Left: Detection and attention output of each perceptual type during the experiment. Right: The same output collected during a shorter interaction session inside the laboratory where the robot is developed

to evaluate what the robot perceived, we counted the number of times each input type (face, color, sound) was detected and subsequently tracked. We then counted the attention system's output to evaluate how it selected from the perceived sensory input. The label *Att* indicates that a particular input was attended, i.e. the attention system selected this object as the next target for that particular frame. The label *Det* represents objects that have been detected and subsequently tracked.

As mentioned above, we are using two separate face trackers, one for the attention system and the other for more accurate generation of face clusters. We will first report results of the KLT-based tracker which is actively affecting the robot's attention system and gaze direction. Manual count of this face tracker's output indicates that the robot detected and tracked 77 individuals during the experiment. The face tracker erroneously merged two individuals in a tracking sequence twice. Of the entire recorded sequence, faces were detected/tracked for 10230 frames and attended for 7754 frames. Faces were both detected and attended simultaneously (*Att-Det*) for 6016 frames. In 37.4% of these frames, the corresponding face was attended because it was the only possible target. In 38.2%, 17 %, and 7.4 % of these frames, the face was selected by the attention system among two, three, and more than four possible targets respectively. Of all faces that were selected as the attention's next target, roughly 1738 frames (22.4%) were not actually detected during the same frame. This is made possible because the attention system has been implemented with some capacity for short-term spatiotemporal memory. Conversely, of all the detected ones, faces were not selected as the next attention target in 4214 frames (41.2%). From this *NotAtt-Det* set, the attention system instead selected nothing for 17.7% of the time, a color segment for 6.1% of the time, and sound for 76.2% of the time.

The SIFT-based face tracker passively processed each incoming image frame, as determined by the robot's attention system. This tracker generated face clusters, each of which was obtained from each same-person tracking sequence. Due to the large amount of data, we have so far manually counted a part of the entire set (1082 clusters from 60 individuals). Despite having to deal with simultaneous interaction with multiple people, the tracker never merged two individuals

Fig. 7. Some examples of the automatically generated individual's face (Left) and color (Right) clusters.

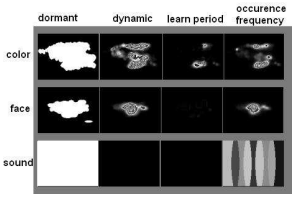

Fig. 8. Various spatio-temporal patterns acquired during the experiment for each perceptual type.

in a cluster. In 9 clusters, the tracker made a mistake by mixing faces with background or other non-face objects. In 247 clusters, the entire sequence contains only background or other non-face objects. Figure 7 Left shows some examples of these generated clusters. Each cluster contains quite a high variability in orientation and facial expression, which is desirable for future face recognition purposes.

As shown in figure 6, color segments were segmented/tracked for 10849 frames and selected as the next attention target for 10428 frames. Color segments were both detected and attended for 6280 frames. From these *Att-Det* occurences, the corresponding color segment was the only available target for 68.6% of the time and was selected as a target among two, three, and more than four possible targets for 24%, 5.1%, and 2.4% of the time respectively. About 4148 frames (39.8%) of all attended color segments had not been tracked during the same frame. From all color segments that were segmented and tracked, they were not subsequently selected as the next attention target in roughly 4569 frames (42.1%). Instead, the attention system selected nothing for 29% of the time, a face for 10.2% of the time, and a sound segment for 60.8% of the time.

During the entire experiment, the robot collected over 210 clusters of color segments. Some of the color clusters are very large, containing a few hundred images, while many are small

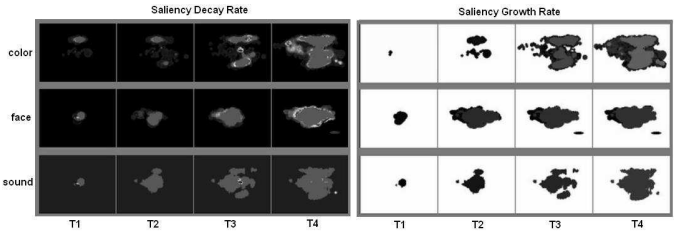

Fig. 9. The progression of saliency decay and growth rates during phase I of the experiment.

with only a few images. These clusters capture a range objects, i.e. toys, walls, ceilings, clothing, faces, etc. Figure 7 Right shows some examples of these color clusters.

Sound sequences were detected for 13745 frames and attended for 31761 frames. Sound segments were both detected and selected as the next attention target for 5462 frames. From these frames, the sound segment was the only possible target for 51.2% of the time and was selected from two, three, and more than four available targets for 33.9%, 10.3%, and 4.7% of the time respectively. From all sound sequences that were selected as the next attention target, 26299 frames had not been detected during the same frame. About 8283 frames of all detected sound sequences were not subsequently attended. The attention system instead selected nothing for 63.4% of the time, a face for 22.2% of the time, and a color segment for 14.5% of the time.

As mentioned above, the robot segmented one or two words from each speech utterance. Each word is then stored in a dictionary and the word sequences are used to learn a bigram language model. Unfortunately, due to human errors, we lost the data acquired during the first half of the experiment. In the second half, the robot stored 59 words in the dictionary. Some of the most frequently heard words are: *you, are, he, good, eight, I, it, no, the, good, name, hey*. Some of the words (single and pair) collected are: *good boy, no, hello, am robot, you yes, four I, wrong, no eight, my name,* and *bye robot*.

Figure 8 and 9 describe various results generated by the spatio-temporal learning mechanisms. We recorded these data in two phases. Phase I covers roughly the first 9700 seconds and phase II covers the rest. The spatio-temporal learning system was reset in between the two phases. We recorded a set of spatio-temporal patterns observed in the sensory input sequence (shown in figure 8), which were used to adaptively alter the attention's saliency growth and decay rate as the robot experienced the world. The dormant cells show all regions where each perceptual input has ever occured in the past. Sound cells are dormant everywhere because the sound localization module only provides five different horizontal direction around the robot. The different face maps indicate that faces tend to occur and move around in the middle front area. Color cells cover a larger area, as the robot actually segmented many color segments from the wall, ceiling, and floor. This causes the learning period of some regions of the color map to be very high, as color segments from the ceiling and floor tend to stay fixed for long periods of time. Figure 9 shows the adaptation progress of both the growth and decay rates during Phase I. We can see that as time passed and the robot gathered more experience in the world, the attention system's saliency decay and growth rates adapted to more sensitive to face, sound, and color segments from certain regions where they have frequently appeared in the past.

IV. Conclusion and Future Work

We present an integrated framework, combining an object-based perceptual system, and adaptive multimodal attention system, and spatiotemporal perceptual learning, for a sociable robot. The robot interacted with over 70 people during a six-hour experiment and collected various face, color, and speech data for future recognition. In the near future, we plan to extend the framework to explore various behavioral adaptation opportunities based on the robot's past sensory experiences. For example, we are interested in utilizing the coupling between attention and spatiotemporal learning to discover various multimodal co-occurence patterns and improve the robot's perceptual system. Figure 6 Right illustrates the equivalent data shown in figure 6 Left for a shorter session inside the laboratory. One can immediately observe the acoustical differences between the two locations. The laboratory is much more quiet and the building lobby is very noisy, thus the typical threshold-based sound detection cannot work well in both locations. We plan to implement hebbian learning within each spatiotemporal map location to learn various properties of sound and face (e.g. sound energy), whenever they occur together. This would essentially allow the robot's sound detector to adapt incrementally according to the current environment.

References

[1] M. Lungarella, G. Metta, R. Pfeifer, G. Sandini, "Developmental robotics: A survey," *Connection Science, vol.00 no.0:1-40*, 2004.

[2] T. Fong, I. Nourbakhsh, K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems 42:143-166*, 2003.

[3] M. Posner, "Orienting of attention," *Quarterly Journal of Experimental Psychology, 32:3-25*, 1980.

[4] D. Ballard. C. Yu, "Exploring the role of attention in modeling embodied language acquisition," *Proc. of the Fifth International Conference on Cognitive Modeling*, 2003.

[5] R. Simmons, A. Bruce, I. Nourbakhsh, "The role of expressiveness and attention in human-robot interaction," *Proc. AAAI Fall Symp. Emotional and Intel. II: The Tangled Knot of Soc. Cognition*, 2001.

[6] S. Frintrop, G. Backer, E. Rome, "Selecting what is important: Training visual attention," *Proc. of the 28th German Conference on Artificial Intelligence*, 2005.

[7] R. Peters , K. Hambuchen, K. Kawamura, D. Wilkes, "The sensory egosphere as a short-term memory for humanoids," *IEEE Int. Conf. on Humanoid Robots, pp 451-459*, 2001.

[8] L. Itti, "Models of bottom-up and top-down visual attention," *Ph.D. Thesis, California Institute of Technology*, 2000.

[9] J. Albus, "Outline for a theory of intelligence," *IEEE Trans. on Sysems, Man, and Cybernetics, vol. 21, no. 3*, 1991.

[10] L. Aryananda, J. Weber, "Mertz: A quest for a robust and scalable active vision humanoid head robot," *IEEE Int. Conf. on Humanoid Robots*, 2004.

[11] L. Natale, G. Metta, P. Fitzpatrick, "Yarp: Yet another robot platform," *Int. Journal of Advanced Robotics Systems, special issue on Software Development and Integration in Robotics, Volume 3(1)*, 2006.

[12] M. Jones, P. Viola, "Robust real-time object detection," *Technical Report Series, CRL2001/01. Cambridge Research Laboratory*, 2001.

[13] C. Tomasi, J. Shi, "Good features to track," *IEEE Conference on Computer Vision and Pattern Recognition, pp 593-600*, 1994.

[14] G. Foresti, C. Micheloni, "Real-time video-surveillance by an active camera," *Workshop sulla Percezione e Visione nelle Macchine, Universita di Siena*, 2002.

[15] R. Mosur. Sphinx-ii user guide. [Online]. Available: http://cmusphinx.sourceforge.net/sphinx2

[16] R. Brooks. (2003) Creature language. [Online]. Available: http://www.ai.mit.edu/people/brooks/creal.pdf

[17] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision, 60, 2, pp. 91-110*, 2004.