REAPER: A Reflexive Architecture for Perceptive Agents

Bruce A. Maxwell, Lisa A. Meeden, Nii Saka Addo, Paul Dickson, Nathaniel Fairfield, Nikolas Johnson, Edward G. Jones, Suor Kim, Pukar Malla, Matthew Murphy, Brandon Rutter, Eli Silk

Swarthmore College
500 College Ave.
Swarthmore, PA 19081
maxwell@swarthmore.edu

1 Abstract

This paper describes the winning entries in the 2000 American Association for Artificial Intelligence robot contest. The robots, developed by Swarthmore College, all used a modular hybrid architecture designed to enable reflexive responses to perceptual inputs. Within this architecture, the robots integrated visual sensing, speech synthesis and recognition, the display of an animated face, navigation, and inter-robot communication. In the hors d'oeuvres serving event, a team of robots entertained the crowd while interactively serving cookies; in the Urban Search and Rescue event, a single robot autonomously explored a section the test area, identified interesting features, built an annotated map, and exited the test area within the allotted time.

2 Introduction

This year Swarthmore College entered robots in two events at the American Association for Artificial Intelligence [AAAI] robot competition. As any AI researcher knows, developing one robot for a real-world task like serving hors d'oeuvres to a crowded room can be a significant undertaking. Our plan was to have a team of three agents participate in the hors d'oeuvres serving contest and to have one mobile robot attempt the Urban Search and Rescue [USR] event.

Even with ten undergraduate students working on the project for eight weeks, developing robots that could accomplish both events at our desired level of performance was difficult. What made it possible, let alone successful, was that each of the agents used the same overall software architecture for integrating navigation and control with perceptual processing. Furthermore, this architecture was largely platform independent and modular, permitting different agents--including non-mobile agents--to use different capabilities with few changes to the overall system.

Using the same architecture for each agent meant we could distribute our efforts and focus on common capabilities such as visual information processing modules and facial animation modules that could be used on several platforms. This permitted us to give each agent a wide range of abilities and integrate them effectively.

The unique aspects of our hors d'oeuvres entry this year included:

- The integration of multiple sensors and modes of interaction in a single agent,
- A powerful, general purpose, real-time color vision module,
- Fast, creative, entertaining, and robust human-agent interactions,
- Facial animation--including tracking faces with the eyes--in sync with spoken text,
- Shirt color detection and identification,
- Fast, safe navigation in a crowded space using a reactive navigation algorithm, and
- Communication and interaction between the agents.

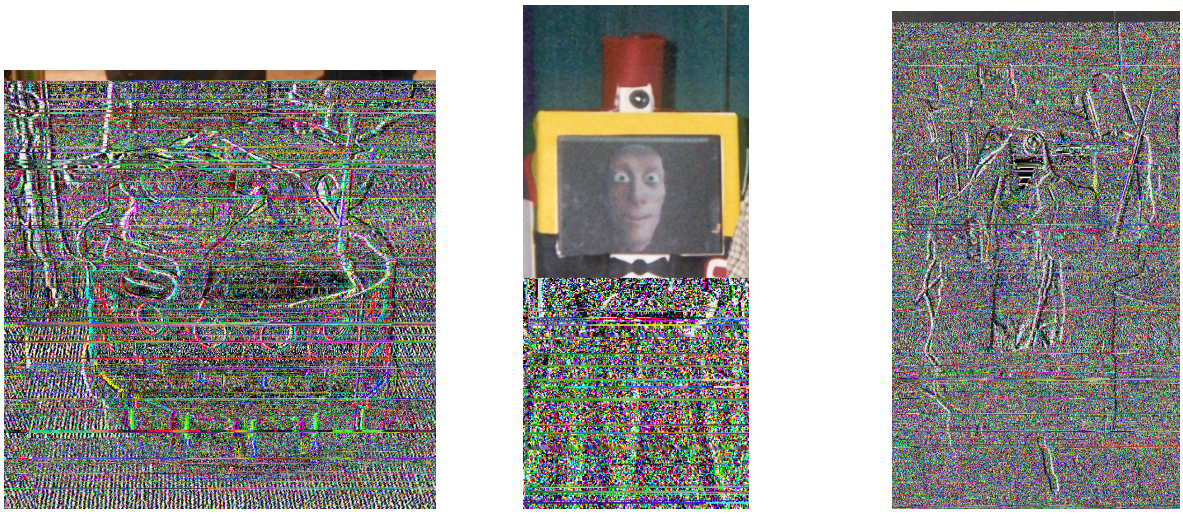


Figure 1 Alfredo (center) and his two “sons”. Mario (620 0 0 ons”9lA telsf th29.8see al40.56 tels

The same architecture also managed our USR entry. The only difference between Mario the bus-boy robot and Mario the rescue robot were the controlling modules. Otherwise, the vision, speech, and navigation modules were identical. The strengths of our USR entry were:

- Completely autonomous function,
- A robust reactive wander mode and “get out” mode using sonar and IRs,
- Providing a map built by the robot with connected annotated images, and
- The vision module, which could identify motion and skin-color.

The theme for our 2000 “Hors d’Oeuvres Anyone?” entry was based on our previous year’s entry. In 1999, Swarthmore’s waiter robot, Alfred, won the “Hors d’Oeuvres Anyone?” event. This year, Alfred graduated to italian restaurant owner--changed his name to Alfredo--and went back to the competition with his “sons” Santino and Mario. Alfredo was the maitre’d, Santino the waiter, and Mario the bus-boy

This year Alfredo was not a mobile robot, but a computer with a large monitor placed at the waiter’s refill station. It had speakers and a video camera, and would respond to different kinds of visual input. The monitor displayed a talking face, whose lips moved in synchronization with the speech. It had three special capabilities: 1) it could tell when you held your palm in front of the camera and would give you a palm reading. 2) it would comment on the color of your shirt (based on analysis of the video image), and 3) it would comment if you stayed in front of the camera too long. Otherwise, Alfredo would talk about various things, responding to its visual input.

Santino, the waiter, was a Nomad Super Scout II, a medium size mobile robot with an on-board computer. Santino was also outfitted with two cameras, a microphone, speakers, a 6” LCD display and a mechanical arm that could raise a tray up and down. Santino used the two cameras to look for people, look for brightly colored badges, and to check when his tray was empty. It would come up to a person, ask if they wanted an hors d’oeuvre and then lift the tray if they said yes. When his tray was empty, it would make his way back to the refill station. When Santino was happy a face on the LCD screen would smile. When it was grumpy or angry, it would frown.

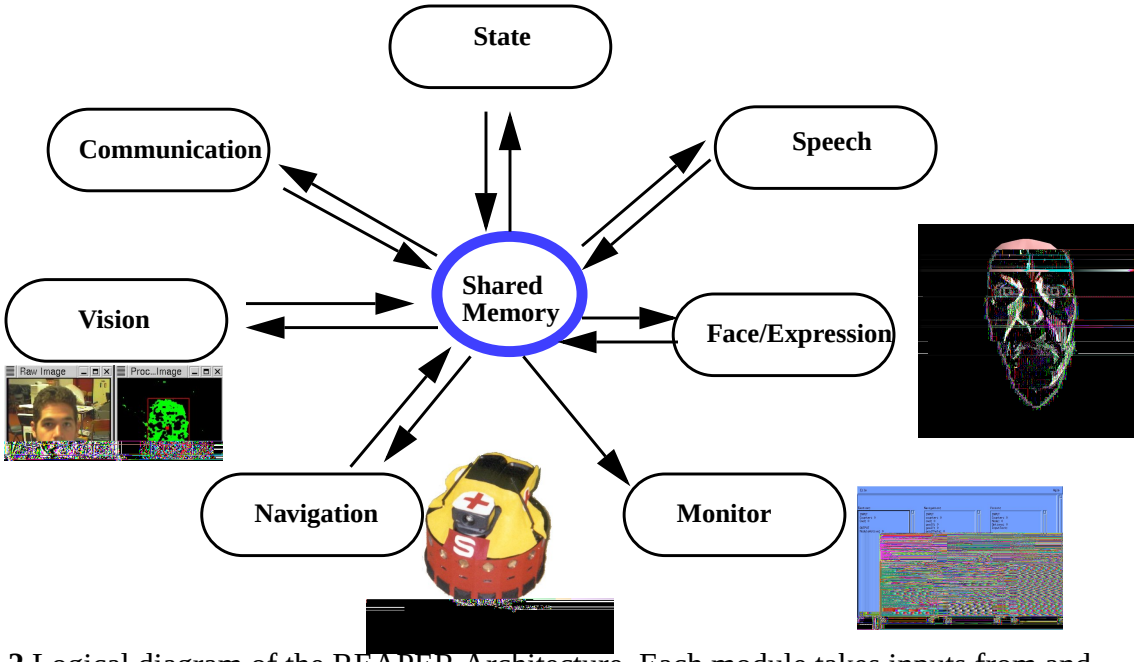


Figure 2 Logical diagram of the REAPER Architecture. Each module takes inputs from and writes its outputs to the shared memory. The State module is the central controlling unit.

Mario, the bus-boy, was a Real World Interfaces [RWI] Magellan Pro, a short mobile robot with a camera and speakers. His job was to provide entertainment by running around in the crowd. During the competition, it also had a plate of cookies on his back. In addition, it would shuttle back and forth between Santino and Alfredo, attempting to strike up conversations with them. The two mobile robots could identify one another by a red, white, and green flag that each carried (one with the red side up, one with the red side down).

This year Swarthmore not only competed in the "Hors d'Oeuvres Anyone?" event, but also in the Urban Search and Rescue event on a standard course prepared by the National Institute of Standards and Technology [NIST]. The robot Mario explored one section of the course autonomously, built a map, and connected annotated 360˚ panoramic images of the scene to map locations. The annotations identified image areas of interest by highlighting motion and skin-color. Mario then made its way out of the course within the allotted time limit (25 minutes).

It's worth taking a look at what was under the hood, so the rest of this paper examines the overall architecture and highlights the most important pieces.

3 REAPER: an Intelligent Agent Architecture

The system architecture--hereafter referred to as REAPER [REflexive Architecture for PErceptual Robotics]--is based on a set of modules. The purpose of each module is to handle one of: sensing, reflexes, control, communication, and debugging. The fundamental concept behind REAPER is that the central control module--whether it is a state machine or other mechanism--does not want a flood of sensory data. Nor does it want to have to make low-level decisions like how fast to turn each wheel ten times per second. At the same time it needs real-time updates of symbolic information indicating what the world around it is doing. The sensor and reflex modules gather and fil-

ter information, handling all of the preprocessing and intermediate actions between high-level commands or goals. This is similar to the way our brain seems to deal with a request to pick up an object. While we consciously think about picking up the object, our reflexes deal with actually moving our hand to the proper location and grasping it. Only then does our conscious mind take back control to decide what to do next. This approach has some biological justification [1].

The two sensing modules handle all vision and speech--based interaction. Their main task is to act as filters between the sensory data and the symbolic information required by the rest of the system. The reflex modules--navigation and face--handle the motion and appearance of the robot. The navigation module also incorporates sensing (sonar and infrared sensors), but its primary task is to control the motion of the robot, not to filter and interpret the sensory information for other modules. Central control of the robot is handled through a state module, and communication between robots is handled through its own module. Finally, the architecture has two modules for debugging purposes. One--the monitor--shows text fields that represent all of the information available to the system. The other--the visual monitor--is designed to show graphically the information being provided by the vision module.

The modules on a robot communicate through a shared memory structure, which provides a computationally efficient means of sharing information. They use a common framework for communicating and programming, including a handshaking protocol to ensure that information and commands are passed and read correctly. Communication between robots occurs through sockets between the agents communication modules over a--possibly wireless--ethernet system.

Central control of the robot is handled by a controller module, or state module. This module is started first, and it starts up all of the other modules it needs--each which of which is its own program. In our implementation, the state module then initiates a state machine process that specifies how the robot will interact with the world, what sensing and interaction modalities it will use, and the navigational goals it needs to achieve. To specify what the other modules should do it uses our handshaking protocol to send information and commands. The other modules, in turn, maintain blocks of output information that the state machine uses to determine what to do next and when certain actions are complete.

The state module is the most difficult to develop because of timing and synchronization issues. In particular, our state machine has to avoid switching between states too quickly. Since the it does not include any of the low-level sensing or interaction, it iterates extremely quickly and can move between states before other modules have any chance to react to the previous state. Thus, it has to watch flags from the other modules to determine when actions are complete before moving on or making a decision. The strength of this approach is that the state machine can sit back and sample high-level information asynchronously, reacting to changes in the world smoothly and quickly.

The REAPER design methodology follows a number of the lessons learned about robot architectures over the past two decades. In particular, it follows the recommendations of Bryson that autonomous agent architectures need:

- a modular structure,
- a means to control action and perception sequences for complex tasks, and
- a mechanism for reacting quickly to changes in the environment [4].

REAPER is related to behavior-based systems, such as those proposed by Brooks, in that it distributes low-level interpretation and control of sensors and actuators to modules [3]. These mod-

ules run in parallel, and those connected to actuators are designed to react to changes in the environment on their own. Within the modules, a subsumption approach or weighted mixing of a set of behaviors are useful ways of generating appropriate low-level actions.

REAPER is not a pure behavior-based system, however, because it is designed to facilitate the mixing of symbolic and subsymbolic, or reactive strategies. In particular, the sensing and acting modules receive symbolic commands from a controller module that tells them their current set of low-level goals. Thus, REAPER also falls into the category of hierarchical architectures, which facilitate the development of action and perception sequences for complex tasks. While our implementation of REAPER has so far been a two-level architecture, our fundamental approach can easily support a three-level architecture with a knowledge-representation and reasoning system as the top-level. The overall controller, in this case, could be either the middle level, as in Gat's ATLANTIS architecture [6], or the top-level as in Bonasso's 3T architecture [2].

Finally, REAPER's modular design facilitates cross-platform and multi-platform development and simplifies debugging and testing of the individual modules. This is particularly true in the case of the sensing modules, which we extensively tested on their own before putting them on the robot platform. As noted by Bryson, modularity in design has become a staple of robot architectures for just these reasons [4].

At the most fundamental level, however, the REAPER approach evolved because it enables the agent to gather--or reap--information at a tremendous rate from a large number of sensors and present that information in a useful, symbolic form to the controlling module. Since our concern was real-time interaction between people and a robot integrating vision, speech, sonar, infrared, and bump sensors, it was essential that we off-load the responsibility for sensor interpretation to individual modules. These individual modules could then plan and schedule processing according to their current goals, providing the most information possible given computational constraints.

3.1 Overall module structure

The modules other than the controller all contain the same basic program structure. After start-up and initialization, each enters an event loop--initially in an idle state. Each time through the event loop, the module tests if the controller has issued a new command. If so, the transition to executing that command take places. Otherwise, the module processes the current command. When it completes the current command, the module transitions itself back to an idle state and indicates to the controller via a flag that it is in an idle state. In some cases, such as sensing commands, the module continues to process and update sensory information until told to do something else.

The goal of all of the modules is to make the event loop as fast as possible. In the navigation module, our goal was to maintain a control loop of at least 10Hz; in the vision module, our goal was to maintain 30Hz, or real-time visual processing.

3.2 Reflexes: Navigation

The navigation modules on the Scout and Magellan have to be platform-specific because of the differences between the two robot's low level interfaces. From the point of view of the controller modules, however, they appear similar. Different groups developed the navigation modules, so, while they both use a reactive architecture, they differ in the specifics.

3.2.1 Scout Navigation

The navigation requirements for the Scout were fairly simple for the hors d'oeuvres event. It had to move slowly and safely, be able to get to a goal location, and be able to avoid obstacles on the way there. In addition, it had to have a mode where it actually stopped for an obstacle in case it was a person to serve.

The navigation module is setup as a 2-layer reactive system. The sensors available to the navigation module are the sonars and bump sensors, including five bump sensors on a low front bumper we added to Santino. The bottom layer contains a set of behaviors that react directly to these inputs. These behaviors include the following.

- Goal achieving
- Obstacle avoidance
- Wander
- Free-space finding
- Front bumper reaction

Each of these behaviors returns a fuzzy priority, speed, and heading. The controller layer combines the speed and heading values based on its mode and the currently active behaviors.

The modes/commands for the navigation system include: Idle, Stop now, Stop slowly, Goto Avoid, Goto Attend (stop for obstacles), Arm Up, Arm Down, Wander, Track Attend, Track Avoid, and a set of commands for resetting the odometry and controlling orientation.

The most interesting of these modes are the track modes. Their purpose is to directly connect the vision system and the navigation system without controller intervention. It can be used to follow a judge's name-tag badge or to track a target in real-time. Once the vision module finds a badge or target, the controller initiates the mode in both the vision and navigation modules. Once initiated, the vision module continues to track the object and update the object's position. The navigation module, in turn, reacts as quickly as possible to the visual information and tries to orient and follow the target. The two modules continue to track the target until either the target is lost, the controller ends the tracking, or an obstacle appears (in the case of Track Attend mode).

3.2.2 Magellan Navigation

The Magellan Pro--Mario--is a small round robot with symmetrically opposed wheels which allow it to rotate on its axis. The basic sensor array consists of a three rings of 16 bump (contact), sonar, and IR sensors mounted around the sides of the robot. In addition Mario has a Sony DV30 pan-tilt camera and external speakers. The on-board computer is a Pentium II running Linux 2.2.10, and it communicates with the robot's rFlex controller over a 9600 baud serial line.

Because of the lack of a low-level software library, we developed an interface for the Magellan which we called Mage. Mage communicates directly with the rFlex controller of the robot. The rFlex accepts a simple set of motor control commands and is also responsible for transmitting the sensor data of the robot back over the serial line. We were able to extract or deduce most of the protocol for this communication from some example code that Real World Interfaces [RWI] provides for updating the CMOS on the rFlex. At our request, RWI sent us code snippets containing information relevant to the IR sensors, which allowed us to enable and read the IR range values. During this time we also developed and integrated a controller for the Sony pan-tilt-zoom camera on the robot, which was controlled over a separate serial line.

In general the Mage API closely resembles the API for the Nomad Super Scout--due to the fact that we have extensive experience with the scouts--although it has a simplified command set and uses as the units of distance thousandths of meters and the units of rotation thousandths of radians (as opposed to tenths of an inch and tenths of a degree on the Scouts).

In keeping with the Nomad API, all sensor and motor control data is maintained in a large state vector. For example, the statement State[STATE_SONAR_0] returns the most recent value of the forward-pointing sonar sensor. This state vector is updated continuously by a thread which handles new data passed from the robot controller. Although the rFlex controller supports a request-based protocol, the simpler method is to ask it to continuously stream data from the sensors as fast as it can. This approach ensures that the sensor data is as up to date as possible. In order to send motor commands, the API includes a method which sets the contents of an output buffer. The same thread which handles incoming data also watches this buffer and transmits its contents to the rFlex controller. As a note, this motor data is transmitted immediately if it changes and then transmitted periodically to keep the rFlex controller alive. Serial communication with the pan-tilt-zoom mount of the camera is implemented in the same way.

The navigation module sits on top of the Mage API and is responsible for reporting the basic sensor data and for actually getting the robot from point A to point B without running into obstacles. The Magellan navigation module has several different modes, but they are all based on a reactive kernel. The robot decides how much to translate and rotate based on four lines of code.

- Translate = Translate - Distance to nearest object in front
- Translate = Translate + Distance to nearest object behind
- Rotate = Rotate - Distance to nearest object to the right (assuming clockwise rotation)
- Rotate = Rotate + Distance to nearest object to the left

To make the robot wander, we give Translate a forward bias. To go to a goal point, we calculate the Translation and Rotation biases required to push the robot towards the goal point. To track an object, the navigation module monitors the relative position of the object--stored in the vision module--and feeds this information straight into the biases. This approach proves to be robust as long as the biases do not exceed the maximum repulsion of obstacles.

To build a map--used in the USR event--the navigation module uses an evidence grid approach [12]. We integrate sonar readings into a probabilistic map that can then be classified into free space and obstacles for interpretation by a person. The evidence grid technique worked well in our test runs, but in the actual event small objects on the floor and tight paths between obstacles caused sufficient wheel slip to significantly throw off the odometry. Thus, local areas of the map were correct, but globally it did not reflect the test situation. This is a common problem with evidence grids, and suggests that for this task we either need alternative means of localization, or we need to take a topological approach to map-building.

3.3 Reflexes: Face

Robot-Human interaction is the key component that distinguishes the Hors d’Oeuvres Anyone? competition from other robot competitions. the goal of creating a fully-functional intelligent agent with the capabilities of any average human is far from realized. Our robot team this year began to make strides in developing our own synthetic characters to better solve the difficult task of the competition by incorporating an animated, 3-D graphical model of a human head with interactive capabilities.

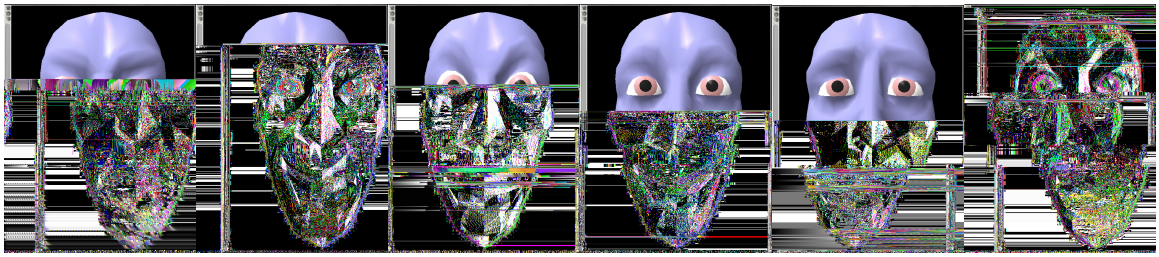


Figure 3 The faces of Santino. From left to right: anger, disgust, fear, happy, sad, surprised.

A growing amount of work has been dedicated to the creation of synthetic characters with interesting interactive abilities. Each year the competitors in the robot contest find better ways to explicitly display complex interaction s with humans. We considered a number of graphical models with the capability to display emotion and the flexibility to add increasingly more complex abilities. The Dragon Wing, for example, is a facial modeling and animation system that uses hierarchical b-splines for the generation of complex surfaces [5]. The technique provides an incredible amount of flexibility, but was too complicated for our needs. Instead we utilized a muscle model for facial animation and facial geometry data available on the web [14]. We ported the system to OpenGL [13] on Linux.

The facial model is a simple polygon representation that uses 876 polygons. Only half the face is actually described in the input data file since symmetry is assumed between the right and left sides. Reading the data and rendering it is straightforward in OpenGL. The system permits the user to view the face data in a number of ways, including: transparent, wire frame, flat shading, and smooth shading. In addition, the user can orient and rotate the face.

The model we used includes a simple muscle model to animate the face. The muscles are defined by specifying the beginning and ending points, as well as a zone of influence. Each muscle can be relaxed or contracted, affecting all those vertices within its specific zone of influence. The face module has a set of predefined expressions which consist of a set of contractions for each muscle in the facial structure. We move between expressions by interpolating the differences in the expression vectors. Our system uses a total of 18 different muscles and 6 unique expressions.

Beyond the structure of the face, we added a couple of features to increase the interactivity of the system. First, the jaw has the ability to move in order to synchronize the mouth with speech generation. The jaw is able to rotate vertically by specifying jaw polygons and then rotating them about a central horizontal axis. The mouth is also able to move horizontally from puckered lips to a wide mouth through the addition of a virtual muscle that contracts the polygons of the mouth. Our speech generation program, IBM's ViaVoice™ Outloud, generates a mouth data structure--containing mouth height and width--in parallel to the sound synthesis [7]. We pass this information to the face module and use it to update the mouth state in synchronization with the robot's speech.

The second capability we added was to give the face eyes--half-spheres colored appropriately with an iris and pupil. In addition, the face module orients the eyes according to the output of the vision module. This simulates the effect of the eyes tracking people's faces or focusing on their conference badges.

We presented the faces on Santino and Mario using color LCD displays at a resolution of 640x480 in 8-bit color. On Alfredo--a dual processor workstation--we presented the face on a 17” monitor with 8-bit color at a resolution of 800x600 pixels. The complete animation capabilities were only used on Alfredo because of the more limited processing power on the mobile robots. On Alfredo, with the full capabilities--and the vision module running simultaneously--the rendering system was able to run at approximately 9 Hz, which was at the low end of acceptable quality.

Overall, the facial animation system greatly enhances the interactively capability of the trio of intelligent agents. It gives people a central focus when interacting with the robots and helps to keep their interest throughout an interaction.

3.4 Senses: Speech

To serve people, a server must be capable of interacting with those being served. The server must signal his/her presence and offer the objects being served, the servee must be able to signal acceptance, and the server must serve. On Santino, we chose to make speech the main modality of communication and wanted Santino to be capable of asking people if they wanted an hors d’oeuvre and responding correctly to their response.

Santino’s speech generation and recognition capability comes from commercially available development software. For recognition it uses the ViaVoice™ SDK for Linux, which we also used in our 1999 robot entry [11]. For speech synthesis, the robots use ViaVoice™ Outloud, which provides the capability to adapt voices and intonation, in addition to being easy to integrate with the ViaVoice™ recognition system [7].

The major difficulty in using the ViaVoice™ software is that it is designed to be used with a single person speaking clearly into a microphone in a mostly quiet room. The hors d’oeuvres competition does not meet that criteria. Instead, we expect several hundred people chatting amongst themselves, and some people not knowing to speak directly into the microphone. Building on our experience in 1999, this year’s module keeps recognition interactions brief and includes some additional filtering to remove unnecessary noise as described below.

To complement the minimal recognition ability, the robots have as much personality as we can give them. This includes writing humorous and sometimes outrageous expositions, using different voices for each robot, and using the phrasing abilities of the Outloud software. All three of the agents--Santino, Mario, and Alfredo--use the same basic software. Alfredo and Mario do not try to recognize speech, but engage in a monologue that responds to different inputs. Santino engages in short interactions involving both recognition and synthesis.

3.4.1 Santino’s speech module

Santino’s speech module integrates both ViaVoice™ and ViaVoice Outloud™ for recognition and synthesis, respectively. One of the major improvements on the speech system suggested by last year’s hors d’oeuvres competition is that our robotic waiter agent can detect when the background noise exceeds a threshold, thus making speech recognition undesirable. When the environment is too noisy, the speech module shuts down its speech recognition component and switches into a different mode that uses only speech synthesis. This noise detection ability greatly improves speech recognition rates since the robot attempts recognition only in reasonable environments.

The noise level detection function calculates the average power of a ten second sound recording from an omni-directional microphone and compares it to a threshold value [8]. The threshold

value is determined at the conference hall just prior to the competition. In determining appropriate threshold values, the function uses the peak power of a sound waveform as a guide to prevent us from specifying a threshold that would never be exceeded. The threshold value is set so that speech recognition will still occur with some background noise.

In addition to making the speech module more robust through the noise level detector, the module uses a simple finite impulse response band-pass filter to eliminate frequencies that are beyond a specified range (~200Hz - 2kHz) [8]. Mechanical objects--like ventilation fans in a conference hall--mainly produce low frequencies, while high frequencies occur from electrical interference in the sound card--which is integrated on the single board computer on the robot. To ensure module independence and speed, the band-pass filter is integrated into the ViaVoice™ Speech Recognition audio library. This bypasses the necessity to first record the speech utterance to a pulse code modulated (PCM) wave file, perform filtration and then pass the output to the recognition engine.

The most important part of the competition for Santino is interacting with a person during a serving scenario. There are several classes of spoken phrases used during the each serving scenario. When the state machine signals speech to begin an interaction, it says something that asks the person if they would like something to eat, usually in an interesting and occasionally rude way. When the robot finishes speaking, the recognition engine gets control of the sound device to record the response of the person. If the recognition engine registers a yes or no response, the speech module reports the response to the controller module, which then instructs speech to respond appropriately and end the interaction.

If the recognition fails, the robot will either say something about the color of the person's shirt--if vision has managed to detect shirt color--or something non-committal. Santino will then ask the person again if they want an hors d'oeuvre and listen for a response. A second failure causes speech to say something to just get out of the interaction, and state starts to look for someone else to serve. If the robot hears nothing at all, the speech module will comment that the person is probably a box being mistakenly served and move on.

During the competition we found this shortened interaction to be better than the long interactions we used in the 1999 competition. People did not lose interest or become confused about what to say. This seemed to make the robot more personable and desirable to interact with.

When Santino is not in an interaction, it mutters, which is a running commentary about whatever the robot is doing at that moment. When the robot is in the GOTO_SERVE state and not serving anyone, it mutters about all the food that it has to give out. In the GOTO_REFILL state, it mutters and tells people to not bother it; there is no food to be had. To get this functioning properly on the actual robot, we had to make synchronous calls to both ViaVoice™ programs telling them to stop controlling the audio device in order to deal with a slow turnaround time switching from input to output on the sound card.

The speech module acquitted itself very well at the competition. Recognition rates in the crowded hall were fairly high, at about 70-75%, which included misrecognitions of people not talking into the microphone, or saying something with absolutely no resemblance to a yes-no response. Given the loudness and the large numbers of people, the robot did just a little worse than a human might have in the same circumstance. Its worst mistakes were when it appeared that a variable was not getting properly cleared, causing the robot to respond to a no response as if it were a yes response, but this only occurred once or twice. We attribute the strong performance of the speech module to

the fact that we could run it and debug it as a stand-alone program prior to incorporating it into the overall REAPER architecture. This, in itself, is one of the strengths of the REAPER approach.

3.4.2 Mario's speech module

Because Mario does not attempt speech recognition, its speech module ia a simplified version of Santino's. The speech module serves a mainly diagnostic function, encoding information about the internal state of the robot into natural-sounding phrases, as well as a means for the robot to communicate its goals and interact with humans. The speech output is expressed as strings and then we render the speech using IBM's ViaVoice™ Outloud. Although the speech module does have the functionality to read and speak a phrase directly from the state module, we often used a more flexible mutter mode. In the mutter mode the speech module monitors the shared memory information fields and makes its own decisions about what to say. Once properly configured, the mutter mode picks an appropriate phrase out of a pool of possibilities every few seconds. To a practiced ear this is informative about the robot's internal state but at the same time it reduces the risk of hearing the same boring phrase over and over.

3.5 Senses: Vision

Being able to sense the visual world gives numerous advantages to a robot, especially one involved in human interaction. Visual capability allows the robot to find and locate objects, detect motion, and identify visual object characteristics. One of our goals in both contests was to make the robots react to their world as quickly as possible. Thus, the navigation module maximized the number of times per second it executed the control loop. Likewise, our goal in the vision module was to maximize the frame rate while still providing a rich array of information.

The structure of the vision module is similar to the others. After initialization, the event loop checks if there is a pending command from the controller. If so, it transitions to the new state according to the command. Otherwise, it continues to execute the current command set.

The vision module includes a rich set of operators for converting images into symbolic information. The three general classes of operators are: object detection, motion detection, and object characteristic analysis. Each command to the vision module indicates a general mode and the set of operators that should be turned on. The controller can then scan the relevant output fields of the vision module for positive detections, motion, or object characteristics. Each output field includes information about where an object was detected in the image and when it was detected as determined by a time stamp. The controller can then decide what information requires a response.

The set of operators provided by the vision module include:

- Person detection based on skin color and gradients
- Motion detection across multiple frames
- Color blob detection, focused on conference badge detection
- P-similar pattern detection
- Red, white, and green flag detection
- Palm detection
- Orange arrow detection
- Shirt color analysis (dependent upon detecting a person)
- Person identification (dependent upon detecting a person)
- Calculation of how much food was on the robot's tray (using the tray camera)
- Take a panoramic image (on Mario only)

Which operators are available depends on the mode the controller selects. The modes relevant to the competition are: IDLE, LOOK, TRAY, and PANO. The LOOK mode is the primary mode of operation and permits all but the last two operators to be active. The TRAY mode activates the second camera input and analyzes how much of the tray is filled. The PANO mode works with the pan-tilt-zoom camera on Mario to generate a 180˚ panoramic image that concatenates eight frames together while simultaneously applying the motion and person detection operators.

While in the LOOK mode, there is clearly no way we can maintain a high frame rate and execute all of these operators on each image. To solve this problem the vision module includes a scheduling algorithm that only applies a few operators to each frame. This solution works because the controller usually doesn't really need to know that there is a badge in view--or whatever other object--30 times per second. Frame rate is usually a lot faster than the robot can react to things since reactions often involve physical actions or speaking. Likewise, most of the other operators do not benefit from continuous application. Since we supply a time stamp with each piece of information, the information sticks around and the controller can decide based on the time stamp whether a piece of information is recent enough to warrant a response.

Our scheduling algorithm is based on the premise that running two operators per frame does not reduce the frame rate. This puts an upper bound on operator complexity, although in the case of motion analysis we get around the limitation by pipelining the process. In the standard LOOK mode, the module randomly selects two of the active operators based on a probability distribution. To create the probability distribution, each process has a weight associated with it, with processes requiring higher frame rates receiving higher weights. Most of the vision operators have small, relatively equal weights. Once selected, the module executes the two operators and updates their information. On average, each operator is executed according to the probability distribution.

The motion detection operator is the most difficult operator to integrate within this framework because it requires multiple frames--at least three for robust processing--and requires a significant amount of processing for each frame. Our algorithm uses Sobel gradient operators to calculate edge images, and then subtracts adjacent (in time) edge images to locate edges that moved. It then locates the bounding box of areas of motion that exceed a certain threshold. We have found this algorithm to be quite successful at locating people in the hors d'oeuvres event [11][10].

To avoid breaking the overall structure of the vision module, we pipeline the algorithm across multiple event loops. The motion algorithm takes five event loops to calculate a result--with the first three capturing images and calculating the Sobel results. To ensure the motion algorithm is called frequently enough, we give it a high weight in the probability distribution. On average, the motion algorithm produces a result 5-6 times per second and the vision module maintains a frame rate of greater than 20Hz. When the motion operator is active, it is usually selected as one of the two scheduled operators.

A secondary mode within the LOOK mode permits tracking using one operator in addition to looking for other objects. To engage tracking, the controller specifies a single tracking operator and the regular list of other active operators. The vision module scheduler then puts the tracking operator in one of the two execution slots and randomly selects the other operator from the active list. This guarantees that the vision module will look for the object being tracked every frame, providing the fastest update rate possible. As noted above, in the tracking mode the navigation module can look directly at the vision module output and adjust its control of the robot accordingly. Mario used this ability to follow badges during the competition.

The scheduling algorithm and overall structure were extremely successful as a way to manage a robot vision system. Even with all of the other robot modules running, the vision module was able to maintain a frame rate of at least 20Hz during the competition. Information updates occurred regularly enough that each agent was able to attend to multiple aspects of its environment with real time reactions.

The interesting new capabilities and algorithms we developed this year were: person detection and identification, shirt color identification, food tray analysis, and Italian flag detection. For details on the motion, color blob, and P-similar pattern detection see [11], [10], and [16].

3.5.1 Person detection and identification

Person detection is one of the most important capabilities for an interactive robot to possess. We used two independent techniques to accomplish this: motion and face detection. Our motion detector was straightforward and is described above, but we took a slightly novel approach to face detection that resulted in a fairly robust technique in the hors d'oeuvres domain.

The basis of our face detection system is skin-color blob detection. The key to skin detection is effective training, since lighting conditions can strongly affect the appearance of colors. We developed a fast, interactive training algorithm that gives the user direct feedback about how well the system is going to perform under existing conditions. The output of the training algorithm is an rg fuzzy histogram, where r and g are defined as in (1).

$$r = \frac{R}{R+G+B} \quad g = \frac{G}{R+G+B}. \tag{1}$$

A fuzzy histogram is a histogram with entries in the range [0, 1] that indicate membership in the colors of interest. You can create a fuzzy histogram by taking a standard histogram--which counts the occurrences of each rg pair--and dividing each bucket by the maximum bucket value in the histogram [17].

We use fuzzy histograms to convert standard images into binary images that contain only pixels whose colors have high fuzzy membership values. For skin-color blob detection we train the fuzzy histogram on skin-color regions of some training images and then keep only pixels with membership values above a specified threshold. To get blobs we run a 2-pass segmentation algorithm on the binary image and keep only regions larger than a certain size [15].

The result of the blob detection is a set of regions that contain skin-color. In previous competitions we ran into trouble using just blob detection because the walls of the hors d'oeuvres competition areas in 1998 and 1999 were flesh-tones. While this was not the case in the 2000 competition, there were other sources of skin-color besides people in the environment.

Our solution to this problem is to multiply a gradient image with the skin-color probability image prior to segmentation. The gradient image, however, is pre-filtered to remove high gradient values (i.e. strong edges). The result is a gradient image where mild gradients are non-zero and all other pixels are zero or close to it. Faces are not flat and contain mild gradients across most of their surface. However, they do not tend to contain strong edges. Thus, including the mid-range gradient values effectively eliminates walls--which are flat and tend to be featureless--but leaves faces. We found the combination to be robust and it reduced our false positive rate to near zero while still reliably locating people.

In the 1999 competition our robot--Alfred--tried to remember people based on texture and color histograms. This worked ok at the competition, but it relied on the person standing directly in front of the camera, which was rarely the case. This year we decided to integrate the person identification with the face detection and shirt color identification. We also decided not to store a permanent database of persons, but instead to only recall people for a short time period. The purpose, therefore, of the person identification was to discover if a particular person was standing in front of the robot/agent for an extended period of time.

After a successful face detection, if the memory feature is activated and called then the memory algorithm extracts a bounding box around the person's body based on the location of their face. It then extracts a short feature vector from that box to represent that person's identity. The feature vector is the top five buckets in an rg histogram--as defined in (1)--the top five buckets in an IB (Intensity, Blue) histogram, the average edge strength as determined by X and Y Sobel operators, the number of strong edge pixels, and the number of significant colors in the rg histogram. These 12 numbers provide a nice key with which we can compare people's appearance.

Once the system extracts a key, it compares the key to all other keys recently seen. The system stores the 100 most recent unique keys. If it finds a probable match, then it will send the matching ID to an output filter. If it finds no match, it will add the key to the data base and then call the output filter with the new ID value. The output filter simply returns the most common key identified in the past 10 calls. However, if no single key has at least three matches in the past 10, it returns a null result (no match). The output filter guarantees that, even in the presence of a person's motion and schizophrenic face detection results (jumping between people), if a person is standing in front of the camera for an extended period of time their key will register consistently.

We used this capability with Alfredo. If a person was standing in front of Alfredo for a minimum period of time, it would comment that they should go do something else. Clearly there are other applications, but we did not pursue them for lack of time.

3.5.2 Shirt color identification

Like the memory operator, the shirt color operator depends upon a successful person detection (face or motion). Once a person is detected, the algorithm selects a section of the image that corresponds to the person's shirt, as shown in Figure 4(a). The algorithm then analyzes a histogram of this region to determine the dominant color. The difficult aspects of this task are selecting a histogram space to use, and attaching color labels to regions of that space.

Based on experimentation, we selected the rgI histogram space to represent color, where

$$I = \frac{1}{3}(R + G + B) \tag{2}$$

I is intensity, and r and g are the normalized coordinates defined by (1).

(R, G, B) are the raw pixels values returned by the camera for a given pixel. The benefit of using the rgI space is that the color--represented as rg--is then independent of the intensity--represented in the I axis. We used 20 buckets in each of r and g, and 4 buckets in I.

Because different camera settings and different lighting affect where a color sits in the rgI space, we calibrate the system using a MacBeth™ color chart prior to each situation in which the robot

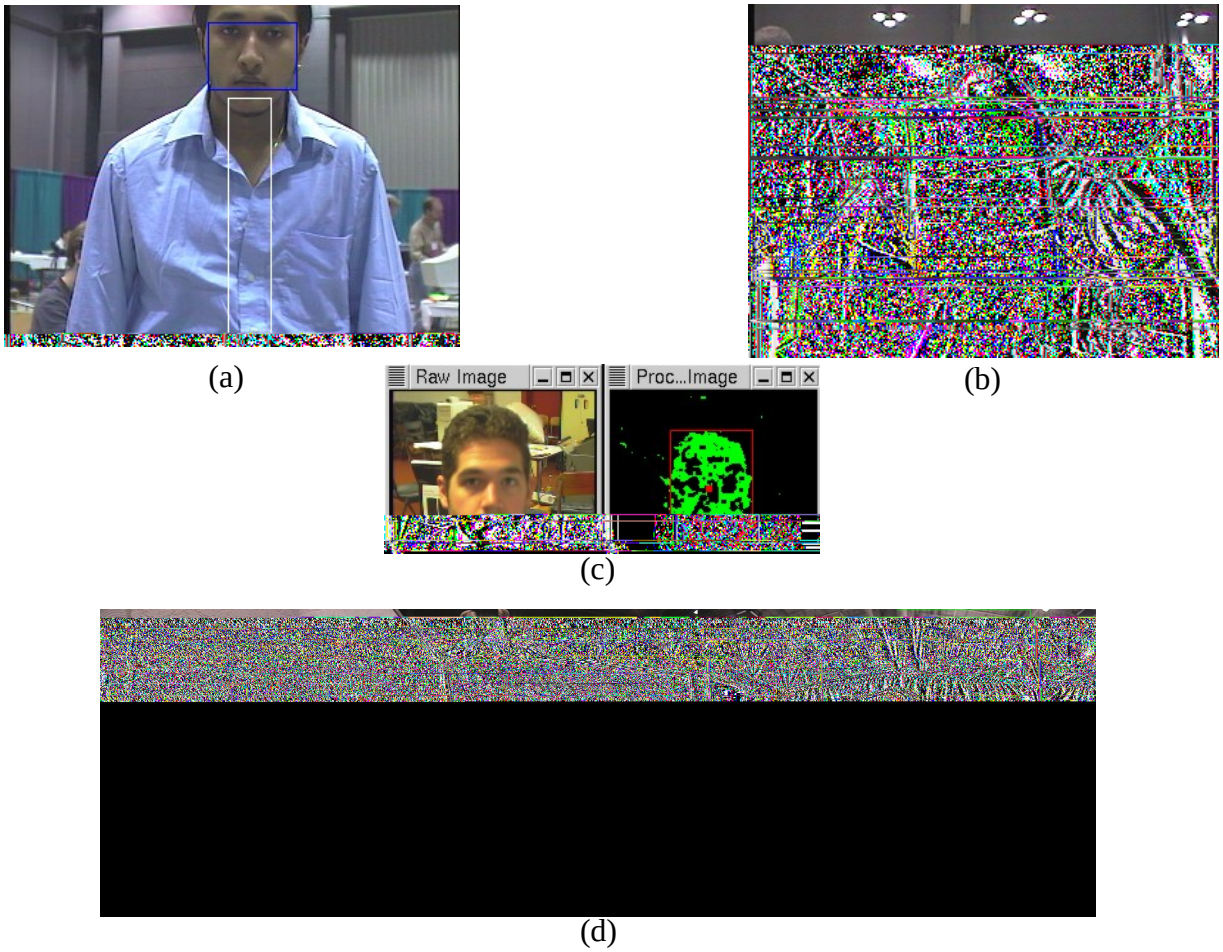


Figure 4 Examples of the vision module in action. A) Successful face detection and the corresponding box used for shirt color and person identification. B) Successful flag detection. C) Training system for face detection system. D) Panoramic image from the USR contest: the green and blue boxes indicate possible motion and skin color respectively. Note that the skin-color on the manniquin's arm--on which we trained--is grey, which is why the walls and floor get highlighted.

interacts. Using a picture of the color chart under the appropriate illumination we identify the centroid in the rgI space for each of the 24 colors on the color chart.

After identifying the region of interest--i.e. the shirt region--the system identifies the most common color in the rgI histogram. The system then finds the closest--in a Euclidean sense--color centroid and returns its text color label as the output. Alfredo used this system to great effect during the competition. It correctly identified numerous shirts, including Dr. Maxwell's mother, who was wearing a purple shirt. It made the computer appear cognizant of its surroundings in an engaging manner.

3.5.3 Food tray analysis

The food tray analysis is a simple, but effective algorithm. We have two cameras on the robot connected to an Osprey 100 framegrabber card with multiple composite video inputs. Upon entering the TRAY mode, the vision module switches to analyzing the input from a small greyscale camera

mounted on top of the tray looking across it. During the competition we used a white napkin to cover the tray and served dark brown or black cookies.

The tray analysis algorithm works on the middle 1/2 of the image, in which the tray dominates the scene. The operator simply counts the number of dark pixels and calculates the percentage of the visible tray that is full. Using pre-calculated minimum and maximum values, the operator sets a flag that specifies FULL, EMPTY, or a percentage in between. This turns out to be a good proxy for how many cookies remain on the tray. Since the small camera includes an auto-gain feature, this method works even when someone blocks the direct lighting by leaning over the tray or standing so it is in shadow.

Based on the percentage-full values returned by the vision module, the controller can smoothly transition from pure serving, to serving while heading towards the refill station, to heading directly to the refill station because the tray is empty. This feature worked well during the competition.

3.5.4 Vertical Italian flag (red-white-green) detection

Finally, we gave the robots for the hors d'oeuvres event the ability to strike up conversations with one another. To make this capability realistic it should only happen when the robots are near to each other which means they need the ability to visually recognize each other. We originally considered putting p-similar patterns--easily recognizable targets--on each robot. However, this would have detracted from the robot's appearance, which was something close to formal dress.

Since our theme was an Italian restaurant, we used the italian flag colors--red, white, and green--as our identifying feature. Santino had a flag draped vertically from his serving tray, and Mario had one placed on an antenna about 4 feet above the ground. Alfredo could also initiate conversations when it saw one of the mobile robots in his camera. To differentiate the two mobile robots we reversed the order of the colors for Mario and Santino from top to bottom.

The technique we use for flag recognition is based on traversing columns--since the colors are arranged vertically. Along each column a state machine tracks the order of the pixels. The state machine only outputs a positive identification if it finds a vertical series of red, white, and green pixels (or in reversed order). Each color has to be mostly continuous and contain a sufficient number of pixels. The state machine allows a certain number of invalid (not red, white, or green) pixels as it traverses the colors. However, too many invalid pixels invalidates that particular recognition and resets the state machine.

This method, since it is based on single columns, turns out to be robust and easy to execute in real time. The recognition system worked well both in test runs and in the competition. Because Santino was almost continuously engaged in serving during the competition, however, it was never able to respond to Mario during the competition. For us, watching the robots engage one another prior to the competition was one of the highlights of the experience.

4 Lessons learned and looking to the future

The products of our experience that we will continue--and are continuing--to use are the overall REAPER architecture, the navigation modules, the face module, and the vision module. All of these provide us with generic scaffolding on top of which we are building other capabilities and systems. All of them are extendable and easily integrated with one another. We also now have

excellent debugging tools that permit us to track all of the information and messages that pass between modules during execution. For us, this infrastructure is the real outcome.

What we also learned is that designing the controller module is still more art than science. From a practical point of view, if we continue to use the state machine approach we need to build a set of standard techniques for managing and passing information around the system. Some of this we have already started, but it needs to be approached in a more formal manner. One alternative is to start building a generic state controller that uses a knowledge management system and a set of rules to determine its actions. This method would implement a true three-layer architecture where the controller sits between a reactive system and a deliberative symbolic system [9]. The REAPER architecture is ideal for this situation since the perceptual information is provided to the controller module in a symbolic and real-time manner.

Looking to the future, if the Hors d'Oeuvres Anyone? event continues then the challenge is to push the envelope. On the interaction front, one challenge is to develop a more generic speech interaction system that can engage in and follow conversations, albeit within a limited domain. A second is to fully implement an emotional subsystem that can affect the whole range of robot behaviors. A third is to more closely link visual recognition of features--such as shirt color--with the interactions in a natural manner. We came close to that goal this year, but to be smooth it must be integrated with a more generic speech interaction system.

On the navigation front, coverage of the serving area has only been achieved by Mario, mostly because it never stopped to talk. Combining Mario's ability to move in a crowd with a more effective Santino will be difficult, because at some point the robot has to take the initiative, stop interacting and move on.

Finally, the multi-robot system proved to be both entertaining and successful at solving the task. Future competitions should encourage multiple robot interaction--two teams attempted it this year. They will have to deal with the fact that it is difficult for the robots to get to one another, but it should be possible.

In the USR task, the challenge is clear. The autonomous entries covered only a small amount of the test area, mostly because of limitations in their ability to sense and interpret the realities of the situation. The tele-operated entry, on the other, did not give much responsibility to the robots. Building meaningful maps, correctly flagging important features or injured people, and simply getting out of the test area within the time limit should be minimal goals for future entries. We believe the techniques exist to accomplish these goals, but their integration in a single package has yet to be completed.

5 References

[1] E. Bizzi, S. Giszter, E. Loeb, F.A. Mussa-Ivaldi, and P. Saltiel, "Modular organization of motor behavior in the frog's spinal cord", *Trends in Neuroscience*, 18:442-446.

[2] R. P. Bonasso, R. J. Firby, E. Gat, D. Kortenkamp, D. P. Miller, and M. G. Slack, "Experiments with an architecture for intelligent, reactive agents", *J. of Experimental & Theoretical Artificial Intelligence*, 9(2/3):237-256, 1997.

[3] R. A. Brooks, "A robust layered control system for a mobile robot", *IEEE J. of Robotics and Automation*, vol. 2, no. 1, 1986.

[4] J. Bryson, “Cross-Paradigm Analysis of Autonomous Agent Architecture”, *J. of Experimental and Theoretical Artificial Intelligence*, vol. 12, no. 2, pp 165-190, 2000.

[5] D. R. Forsey and R. H. Bartels, “Hierarchichal B-spline refinement”, in *Computer Graphics (SIGGRAPH ‘88)*, 22(4):205-212, August, 1988.

[6] E. Gat, *Reliable Goal-Directed Reactive Control of Autonomous Mobile Robots*, Ph.D. thesis, Virginia Polytechnic Institute and State University, 1991.

[7] *IBM ViaVoice™ Outloud API Reference Version 5.0*, November 1999.

[8] E. C. Ifeachor and B. W. Jervis, *Digital Signal Processing. A Practical Approach*, Addison Wesley Publishing Company, 1995.

[9] D. Kortenkamp, R. P. Bonasso, and R. Murphy (ed.), *Artificial Intelligence and Mobile Robots*, AAAI Press/MIT Press, Cambridge, 1998.

[10] B. A. Maxwell, L. A. Meeden, N. Addo, L. Brown, P. Dickson, J. Ng, S. Olshfski, E. Silk, and J. Wales, “Alfred: The Robot Waiter Who Remembers You,” in Proceedings of *AAAI Workshop on Robotics*, July, 1999. Submitted to *J. Autonomous Robots*.

[11] B. Maxwell, S. Anderson, D. Gomez-Ibanez, E. Gordon, B. Reese, M. Lafary, T. Thompson, M. Trosen, and A. Tomson, “Using Vision to Guide an Hors d’Oeuvres Serving Robot”, *IEEE Workshop on Perception for Mobile Agents*, June 1999.

[12] H. P. Moravec, A. E. Elfes, “High Resolution Maps from Wide Angle Sonar”, *Proceedings of IEEE Int’l Conf. on Robotics and Automation*, March 1985, pp 116-21.

[13] J. Neider, T. Davis, and M. Woo, *OpenGL Programming Guide: The Official Guide to Learning OpenGL*, Addison-Wesley, Reading, MA, 1993.

[14] F. I. Parke and K. Waters, *Computer Facial Animation*, A. K. Peters, Wellesley, MA, 1996.

[15] A. Rosenfeld and J. L. Pfaltz, “Sequential operations in digital picture processing”, *ACM*, 13:471-494, October 1966.

[16] D. Scharstein and A. Briggs, “Fast Recognition of Self-Similar Landmarks”, *IEEE Workshop on Perception for Mobile Agents*, June 1999.

[17] H. Wu, Q. Chen, and M. Yachida, “Face Detection From Color Images Using a Fuzzy Pattern Matching Method”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 6, June 1999.