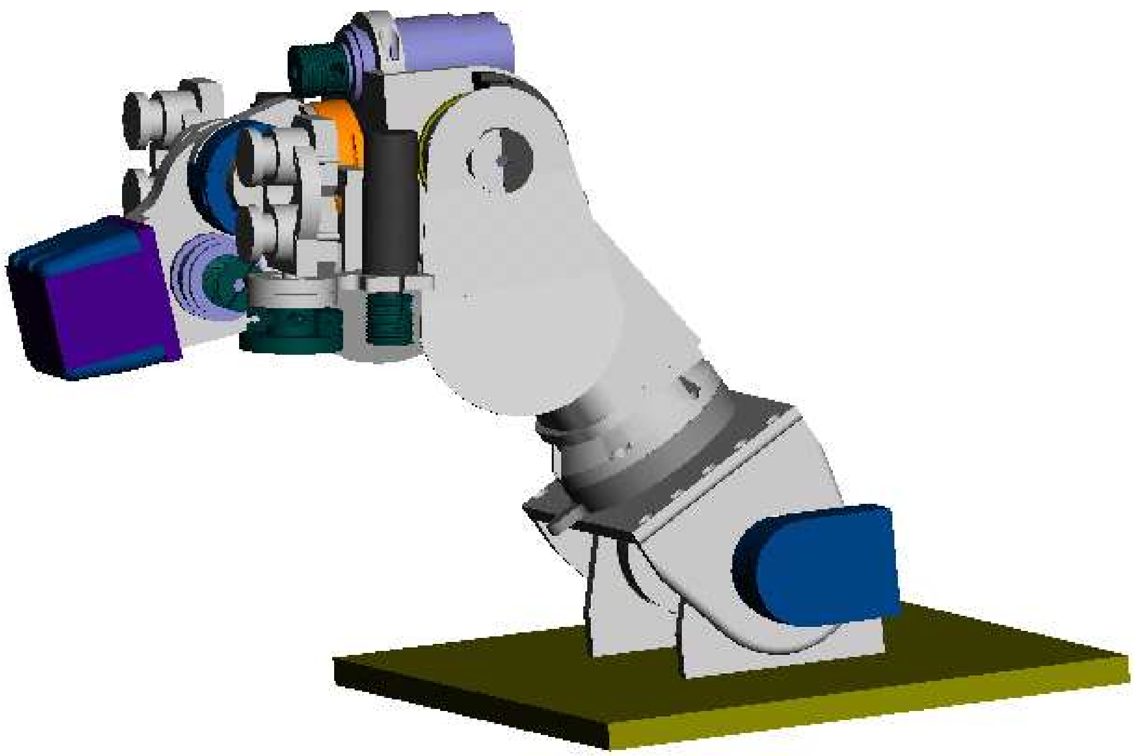
M4 Project

Head and Brain Development



Primary Sponsor: Air Force Aerospace Research - OSR
Sponsor: IS Robotics contract number F30602-96-C-0280

Adding an Active Vision Head to the M4 Robot

Report, March 2000

Introduction

This report describes the development of M4, a mobile robot project in joint development with the MIT Leg Lab. Because M4 is a quadruped automaton it must incorporate a vision system that allows for real-time interpretation of its surroundings and response to its own initiatives in those surroundings. This mutual feedback and correlation between the active vision and sensory-motor systems is imperative for such biologically inspired designs [3].

1. Perception for the Robots

There has been much machine vision work done involving many different visual modules. However, most of this work has been applied to slow or stationary unintegrated systems with less complex motility requirements. We will adapt as much of this previous work to our system as is feasible, such as elements of vergence and object tracking. Furthermore, we will integrate formerly tested VOR and sensory-motor mapping algorithms. Currently we have initial original versions of the face detection [2] and gaze stabilization algorithms [1] and a working motion tracking algorithm [2].
A careful selection of the hardware and a mechanical design of the head have also been finalized. Each vision module is first being separately developed. Once we have working versions of these programs, they will be combined into the unified sensor-actuation-vision system of the head and neck. Refinement of this stage will then enable us to integrate this composite with the body:

- *Image Acquisition* - The vision software is invariably linked to the optical hardware, which exploits biological binocular apparatus, each eye consisting of one B/W camera for wide range image acquisition at low resolution and one foveal color camera for narrow fields of view. The continuous input to these sensors must be analyzed with computationally efficient and synchronous algorithms so that veridical assessments can be made as the robot navigates through its environment.

- *Attention mechanisms* - The success of a convincing intelligent robot is greatly gauged by its putative social skills. Thus, all of the tracking mechanisms previously illustrated will be of maximal value when they contribute to the robot's capacities for human interaction, inference, and attention.
- *Stabilization Algorithm* — Consists in stabilizing the head and cameras while the robot is in movement, through the use of inertial sensors. Needs to be executed in less than 10 ms. The kinematics and dynamic study of the head are concluded, and experimental results were obtained from simulation.
- *Vergence, Smooth Pursuit and Saccades* - These three algorithms are tightly related. An approach based on active vision merges these three behaviors, running at 30 Hz.
- *Obstacle Detection*: Should run at 15 Hz. Obstacles may be detected using features saliency, such as motion, form and color saliency.
- *People Detection* will run only on regions where objects were detected. The detection of obstacles and people will be done on the small resolution images. *Face detection* will be executed whenever people are detected, and people gaze direction will be extracted from the high-resolution color images.
- *Human gaze direction* - provides cues for robots head gazing. This module is important for the robot to infer in which direction people are looking and to what are they paying attention.
- *Matching* — corresponds to finding the correspondences between small and high-resolution images.
- *Rough Terrain and Slope Detection* - operate at small frequencies, namely ~ 8Hz (but might run at ~ 2Hz if necessary due to computational requirements). Visual information will be fused with filtered inertial information. The Fast Terrain Navigation module will run at 15 Hz and provide rough estimatives for save directions of movement.

One approach of AI is the development of robots whose embodiment and situatedness in the world evoke behaviors that obviate constant human supervision [9,10,11,12]. With this in mind, M4 was designed to walking, run, and turn in unstructured environments. Thus, its vision system must comprise methods of assessing the constantly changing terrain. Navigation for obstacle and potentially treacherous landscape avoidance, slope detection, and gaze stabilization are all requisites for such competence. Furthermore, in order to be a convincing social participant, the vision system will also allow for person detection and inference of human gaze direction. All of these vision modules should be accessible to each other and concurrent with other sensor input from the rest of the body.

Attention Mechanisms — Selection of Stimulus

A special issue that remains to consider is the activation of the different modules. Some of the modules will run in parallel in different processor units (like the control/stabilization and Obstacle detection). Others will be executed sharing the time of the processor, such as Vergence and Easy Navigation. Finally, People Detection will run only if an obstacle has been detected, and Face Detection will run only after a person has being detected.

How the robot will share the attention between multiple targets? Although several people will be detected simultaneously, Face Detection will require a decision by the system on which person should be selected. Furthermore, the vergence process will require decisions on each direction to look, and the stabilization and control decisions on each kind of motion should be applied each time. Attention to a certain kind of stimulus will be dictated also according to the state of the robot and the task being performed. As a matter of fact, the robot will be looking constantly for salient stimulus that may be navigation related, such as presence of obstacles, or for more general stimulus, such as movement, people or face detection. Attention mechanisms are already implemented at our lab, [40]. In this project, the goal will be to adapt attention mechanisms to the functionality and necessities of a mobile platform.

During software development, special efforts will be put to give learning capabilities to the system, mainly by investigating the use of neural networks, Support Vector Machines, and Look-Up-Tables. Finally, we will try to simulate the plasticity in the nervous system, which will increase the flexibility and capabilities for learning and performing the desired tasks.

Active Vision Head & Stabilization

A preliminary version of the stabilization algorithm was concluded in December 10th, 1998. The strategy implemented is flexible, being the same algorithm applied both to image stabilization and to saccades and smooth pursuit motion. The algorithm implemented exploits in a straightforward way *redundancy*, and a detailed description may be found in [1]. Further advantages include the use of the same algorithm for implementing different behaviors for body expressions, such as: approaching an object of interest; looking to a desired surface of an object or performing exploratory nearly circular movements around it, or even to saccade far away from the object in case of danger. The same strategy may also be applied for controlling the vision head to keep multiple targets on the visual fovea, among other behaviors for controlling head movements, including the implementation of VOR for head/neck coordination using a neural network. Future developments includes switching between different *body expressions* in a natural way, using, for example, learning.

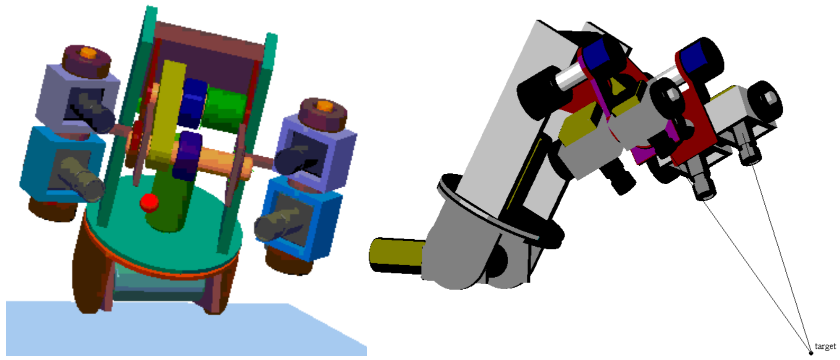


Figure 1 Approaching postures (curiosity behaviors - left), and object tracking (right)

Stabilization has to operate at frequencies on the range of 100 Hz. This requires a simple optimized algorithm — like the one implemented, that fulfills such requirement. Having the head stabilization operating about three times the frequency of visual acquisition allows stabilizing the image outside the transient motor response.

Current simulation results are available for perfect measurements. Inertial sensors will be used to stabilize the Head, which requires three gyroscopes that measure angular velocity. This biologically inspired strategy controls the head independently of body motion.

Redundancy — from biological systems to robots

Redundancy consists of extra freedom of choice to accomplish a certain task. For motor tasks, it is often dealt as an optimization problem to be solved, although nature found a clever solution by combining primitives to generate the movements. Functional modules in the human cerebral cortex have layers of nervous cells performing the same type of information processing. In case of brain damage, the living organism can reconnect layers of cells to carry a different function, therefore recovering from the damage. This process is also essential for aging people. As an e.g., although men loose more nervous cells than woman after fifties, they recover also better from this lost, because their brain reconnects more efficiently layers of cells (larger plasticity). Redundancy is also notorious in the motor system. For e.g., our arms, legs and head have extra degrees of freedom that permit them to move much more efficiently.

Thus, redundancy is crucial for plasticity and learning. Indeed, Central Nervous System neurons stop growing in human beings after three months of gestation. It is the simultaneous divergence and convergence on layers of neurons that facilitates learning, through reorganization of brain structures and creation of new synapses. Therefore, redundancy plays an essential role in biological organisms, and research on the different aspects (such as controlling, deciding and learning) that are related to it may bring new insights to the study of brains and bodies.

Visual Tracking — Vergence, Smooth Pursuit and Saccades

The main methods that have been presented in the literature on active vision for tracking a moving object over a sequence of images are optical flow, image correlation, and deformable contours. The use of optical flow involves segmentation of the object to be tracked from the scene background,~\cite{SantosVictorSandini}. This problem is far from being solved and the existing algorithms (typically segmenting various motion models) are very time consuming. The study of insect visual systems such as that of bees, which relies on optical flow for navigation, is especially important for these types of systems.

Coombs, [5], suggests the use of a zero disparity filter to segment the object from the background, assuming that the object is on the fovea. Bernardino et al, [6], follows a similar approach using log-polar images and multi-scale disparity. A method is presented in [35] to extract corner points of the target image, which are then tracked using correlation. This assumes that the target moves with known constant velocity relative to the camera. Besides their use in active vision, correlation techniques have also been widely used in stereo, [18]. Correlation was used extensively for the first version of the active vision system.

The first version consists of using the monocular and binocular modules supporting each other (Figure 2). The monocular module estimates the position of the moving object in each image, from which it is possible to extract the disparity. This estimated disparity is applied by the binocular algorithm to correlate pairwise left and right points at the same disparity. Hence, the points corresponding to smaller correlation values are assumed as the ones corresponding to the object of interest, since background points have different disparities, and therefore a higher correlation cost is expected for those points.

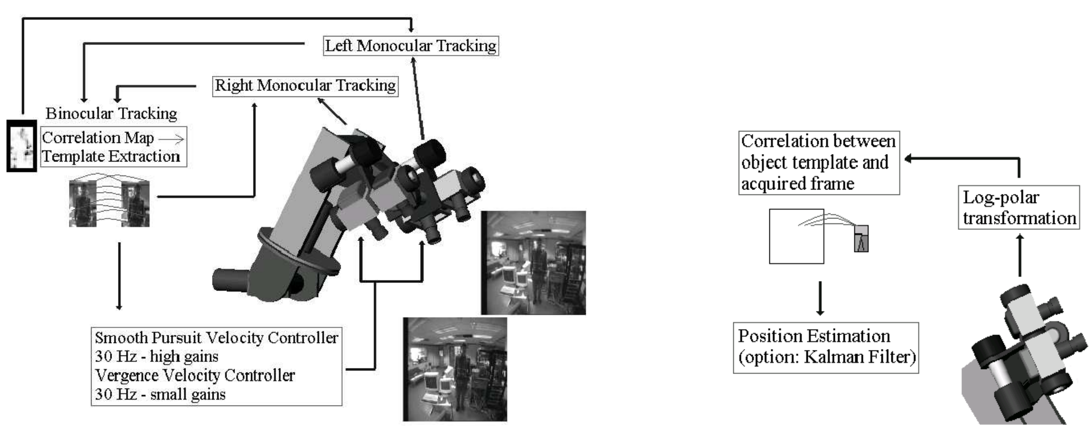


Figure 2. Active binocular tracking (1st version).

This version was tested on the humanoid robot Cog. Samples of one experiment are shown in Figure 3.

Figure 3. Cog tracking a person.

A second version of this algorithm, more sophisticated, is already fully developed, but untested. Velocity estimations are determined after the calculation of the Optical Flow in an image window. The optical flow is determined by filtering spatially (using a Sobel edge detector) and temporally the images, and then the aperture problem is solved through optimization, [23].

A different version of the 0-disparity filter [15], which was named N-disparity filter, was developed to segment a moving object from the background, by applying correlation techniques. This algorithm is applied to log-polar images (Figure 4), so that information in the neighborhood of the predicted disparity is heavily weighted. To speed up computations, a pyramidal image resolution is used to constrain the search space (exemplified in Figure 5).

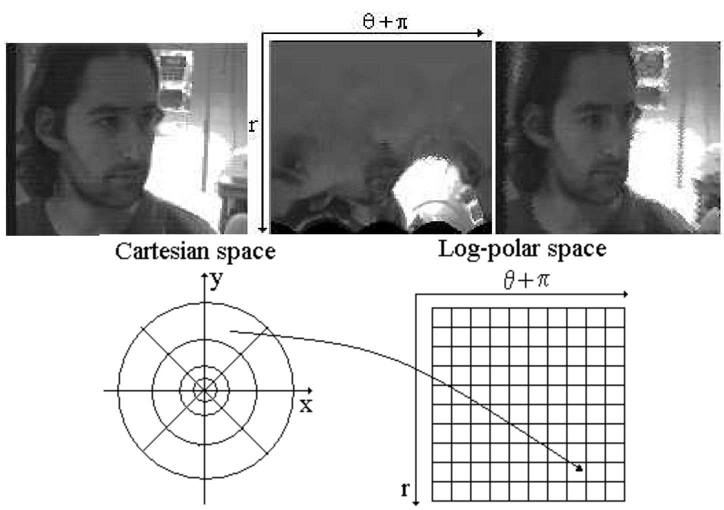


Figure 4. Log-polar transformation.

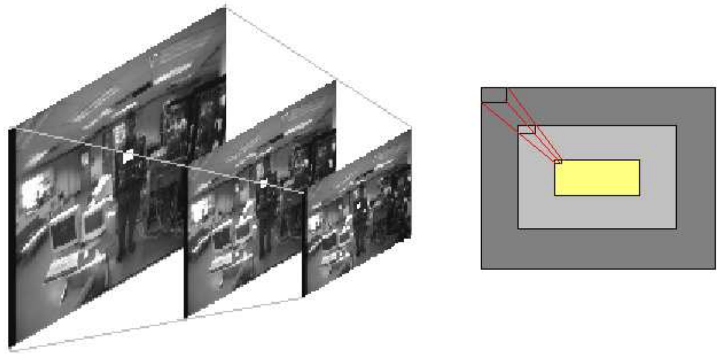


Figure 5. Pyramidal image resolution.

Easy Terrain Navigation

Navigation of the robot over smooth and horizontal terrain may be accomplished by balancing the left and right optical flow. Such approach is computationally expensive. However, the tracking approach already determines the optical flow, and thus this solution may be feasible for images with enough texture. Hence, a simple navigation algorithm will be implemented, based on the type of terrain.

A first cue related to the terrain may be obtained from both eyes by comparing the present image with the previous one (for example, by image difference). Another cue is related to homogeneity in the images: for homogeneous images, it is most probable that no obstacles exist on the path and that the material that constitutes the terrain remains the same as previously. Whenever significant changes on the image occur, then the rough terrain navigation module should become active, and the attention process is directed towards obstacle detection.

As an alternative, an algorithm similar to the one presented by Lorigo et al [28] might be useful to estimate the contours for a path or to detect obstacles that lie on the ground.

Rough Terrain Navigation

Several issues must be considered for certain types of terrains:

- Slope of the terrain: since the robot s locomotion system allows the robot to run, it will be essential to estimate the slope of the terrain relative to an inertial referential. This is important both for controlling robot s legs as well as for implementing emergency stops in case of a dangerous slop, or even to select the direction of movement with the smaller slope.
- Terrain roughness estimates: the robot must avoid navigating on a soft terrain, which may present unpredictable holes or may not resist to the robot s weight. Thus, the direction of movement will be selected according to the terrain.

The algorithm that may provide terrain roughness estimates may also solve for the determination of the terrain s slope. A gyroscope mounted on the head will measure the direction of the acceleration vector relative to its posture on the head. However, the robot s acceleration will change the direction of the final acceleration vector, adding a component to the gravity acceleration vector. This component may be compensated by receiving information from the legs, since the leg s dynamics is known.

The direction of the gravity acceleration vector does not correspond to the direction of the normal to the terrain s slope, because the gravity acceleration vector direction remains the same if the slope varies and the body compensates such variations. Thus, the visual image system must provide the direction of the normal to the terrain s slope relative to the head s frame of reference, while the inclinometer determines the rotation of this vector from the head s referential to a world referential.

The determination of the terrain s slope relative to the head referential may be accomplished through vergence. Gazing of both eyes to a point of interest gives the 3D position of that point relative to a cyclope eye viewed from the middle of both eyes baseline. Therefore, the procedure may be as follows: whenever is required to update the slope, detect a point of interest in the fovea of one image (a point of interest is defined as a point in which neighborhood significant changes of features occur, such as intensity gradient), and the correspondent point on the other image. The correspondent eyes angles give a 3D position. Next, gaze the head to a new point of interest in the periphery of the previous point, but out of the fovea. Redirect the head to the new point, gazing both eyes to the new point of interest. A new 3D position is obtained. After a few acquisitions, 10 or 20 points have been acquired and the normal of the slope determined at these samples. This procedure gives a very coarse estimate of the path s slope.

The idea underlying the slope determination is inspired in biological systems. For example, human s eyes, when looking to an image - e.g., a portrait, as described in [25] — execute a number of very small movements along the image, especially in the neighborhood of features rich in information (such as the eyes and the mouth). Since slope determination may be obtained at frequencies of 2Hz (and only whenever necessary), the algorithm is not computationally expensive — humans, when driving, should not loose attention more than 2 seconds. Then, 0.5 seconds for this algorithm seems a reasonable limit. In addition, slope information may be used to support the estimation of terrain s roughness and to select the best direction to follow. Additionally, the coarse slope map might be also used to estimate holes in the path (if the slope map is not uniform) or significant changes on the terrain relative to the previous map (by comparing both slope maps). Significant changes on the terrain relative to the previous map may indicate that the terrain where the robot is navigating changed, and thus it might be useful for the robot to reduce the velocity.

Finally, significant changes within the slope map (temporary map, not stored) may also indicate obstacles (or a person along the path that should be identified by the robot). Thus, the slope map can also support the obstacle detection algorithm. Errors in the slope may be compensated using sensory feedback from the force sensors in the legs. Although this information is obtained with delay, it can provide an important measure to correct for systematic errors.

The slope map, although not providing information concerning the materials that constitute the terrain, provides information relative to the previous slope map and also relative to the information received from the legs. Disparity maps may also support this algorithm.

Obstacle Detection

In a first phase, the slope map may provide coarse information to avoid large obstacles on the path. However, it must be developed an obstacle detector at high frequencies (more than 15 Hz of operating frequency), using vision. A possible implementation could involve the detection of points of interest in the right/left image, and correlation with neighbor points on the other left/right image to find matches. Indeed, the tracking and obstacle detection algorithms may share the same structure, both directing the robot attention for movement. The difference between the two modules resides on the fact that the former will direct the robot to engage and interact with people, while the latter module initiates motor commands that change robot direction for safe navigation.

Detecting People and Faces

Operating times are estimated on the order of 10 Hz (similar to the rate achieved by Poggio et al.). The same software, with some modifications, may be applied to face detection. The people detection algorithm runs only in regions in a neighborhood of an obstacle detected by the obstacle detection algorithm. After, face detection is activated once candidate locations for people have been identified, using the color narrow range images (high resolution). Redundancy of the active head may be exploited to coverage a large field of view (or the larger number of targets in the field of view). The coarse depth map computed in the previous section may support people detection, by providing a lower bound on the scale of possible humans in each direction.

An obvious prerequisite to the higher-level behaviors is the basic ability to detect people. Such a mechanism can then be piggybacked on the motion detection and tracking modules since people are relatively active, especially those to which we direct visual attention. The fact that the human face is such a rich source of non-verbal communication and information makes it a particularly attractive focal point for person detection. For instance, facial expressions, gaze direction and vergence on distal objects, eye contact and other perceptual facets of joint attention are extremely valuable components of social reciprocation, [5]. Furthermore, demonstrations of infant preference for face-like arrangements and examples of syndromes involving face recognition suggest that facial cue processing may even be relegated to specialized areas of the brain, [34]. This biological data evidences the natural salience of the face as a visual stimulus and thus offers a bearing for the person detection algorithm.

However, when it comes to recognition of a 3D object, especially of a detailed face, from 2D digitized video camera images, computer vision is challenged by variability of appearance. Not only does the non-rigidity of a face due to speech and expressions contribute to these difficulties, but also so do variations in illumination conditions, relative position, and scale. Thus, many face detection techniques employing pixel-based pictorial templates either require storage of an inordinate number of facial views, [7], or necessitate strict experimental controls, such as presentation of fixed frontal images in which the bilateral symmetry of the face is exploited, [36,44]. In attempt to mitigate these constraints, other face detection approaches rely on parameterized deformable templates [30,45], elastic feature matching [26], and snake or snakelet contour detection [27].

Because the ultimate goal of the face localization system is for application in a real-time, real-world vision system, it was important to find a solution computationally inexpensive, efficient, and that allows for reliable location of faces in cluttered, populated scenes.

It was decided to use the ratio template face detection algorithm currently employed by Cog, [38,39], as a springboard for our methodology. Geared for face localization in situations of social engagement, Cog's system detects frontal views of faces by comparing potential candidates to a template of relational grayscale-averaged regions. Such a template capitalizes on the inveterate brightness relations of face structure, not absolute intensities, to achieve immunity from illumination conditions and easily accommodate for scale invariance. The notion of such an invariant attribute template approach attractive in its potential for both diminished storage requisites and resistance to computationally onerous image transformations. Even more alluring was that its extraction of relational correspondences between a whole stimulus face and a stored representation relied on approximations of qualitative attribute dimensions. This latitude allows for broad application due to the generalized nature of the referential face representation, [13,42]. Thus, such an approach involving relational templates would permit interpolation of approximate facial aspects irrespective of attitude. In other words, the right single template could obviate the need for a comprehensive database of image templates over all head pose orientations.

Furthermore, the approach seemed plausible from a biological and situated standpoint. Often the real world does not avail us precise or singularly adequate perceptual information due to sensor (eye) noise, occlusions, object and environment novelty, etc; we often have to use comparative information to construct unique interpretations. In fact, studies suggest that almost all forms of available information are used in human face recognition [14] and even that a viewpoint between the full face and the profile is neurophysiologically preferred in such tasks, [41].

Therefore, it may be that at the even more generalized level of face detection, specific templates for each pose orientation are not necessary. Even though the ratio-template currently employed by Cog's face detection system does not perform well over changes in head rotation, modification of the underlying mechanism seemed capable of achieving such proficiency. So in order to realize the maximum flexibility of the technique, it seemed necessary to employ the simplest possible template: an outline of an oval. The template was chosen based on the observation that the face and head maintain a somewhat oval contour over most angular rotations. (For the initial version of our program, however, we concentrated mainly on rotation about the spinal axis).

Figure 6. Oval template for face detection.

As Figure 6 demonstrates, in almost all poses, the actual facial features are contained within the surface area of the face; therefore, the template would only have to approximate for ears, hair, and parts of the nose. This template is then compared to raw grayscale images to find correlative contours representing potential faces. The configuration of the template oval (TO) was designed to be ‘‘incrementally evaluated’’ in order to allow for positive correlations even in the case where a segment of the face image outline may belie the inherent contour of the oval, such as in the profile case. This technique also allows for partially occluded faces to be detected because small interruptions in contour would not disqualify an oval from face candidacy.

This technique may be combined in the future with skin detection (both for people and faces detection), to unify both algorithms.

2. Building the Head

Several requirements impose constraints especially on head s weight and size. Therefore, the main processing units will have to be placed on the body. However, due to the same requirements for the body, processing units have to be small and light. Furthermore, cameras have to be synchronized and close enough from the frame-grabbers to avoid delays and corrupting data when transfers. Other requirements are specified for the video cameras, such as focus, focal length, and resolution, among others. Special attention was put on compatibility between all the components that will be integrated into the system.

Head Design

The mechanical design of M4 head resembles a dog. Except three aesthetic components (two plastic ears and a small plastic piece on the nose, above the gyros), all the other components are functional.

The weight of the head (without electronic hardware and motor amplifiers, but including motors, gears, gyros) is 3.55lbs (~1.6Kg). The height of the head is approximately 7.5in, and the distance between the eyes is approximately 3.5in. The head has seven degrees of freedom (four in the neck, three for the eyes), including two eyes, and two CMOS miniature color board cameras at each eye. The eyes have a common actuated revolute joint (a tilt joint) and independent pan actuated joints (one for each eye). The two cameras at each eye rotate with same pan and tilt angles and acquire depth information from their binocular geometry.

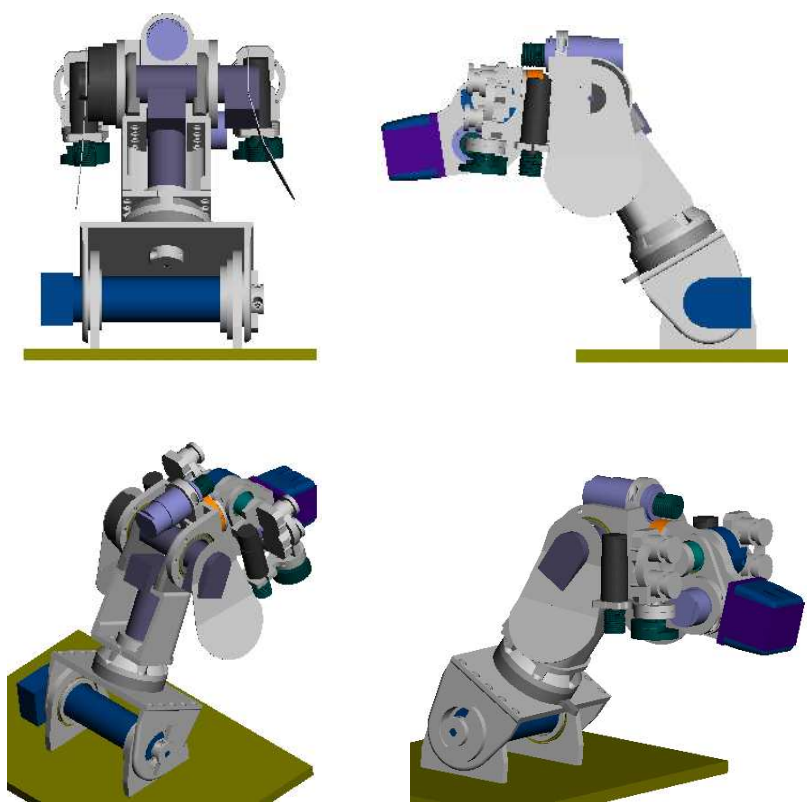


Figure 7. Various views of the design.

The material used for all the functional parts of the head is Aluminum 6061. The Capstan parts were also ionized for extra wear resistance. This solution was preferred over reinforced plastics (namely graphite-epoxy). Although graphite-epoxy has the largest strength to density and stiffness to density ratios, the material is anisotropic, and thus large ratios are obtained only along the direction of the reinforced fibers. Furthermore, inserts in Aluminum are necessary for mechanical fastening, making the design more complex. In addition, parts are more expensive and more difficult to produce, which introduces severe limitation on the shape of the final design.

Thus, by careful design, the specifications were met using Aluminum, by manipulating the geometry of the head, and selecting motors appropriately according to that geometry and to the power requirements. At four of the joints (eye joints and upper neck tilt — see Figure 8), the motors are connected to the moving element through tension cables (to reduce backlash). At the other joints, the motors are connected directly to the moving element, because of the small backlash on the gears head used with these motors, the large torques required for these motors, and also due to constraints in the geometry of the head. A pair of Ball Bearings protects each motor from radial torques.

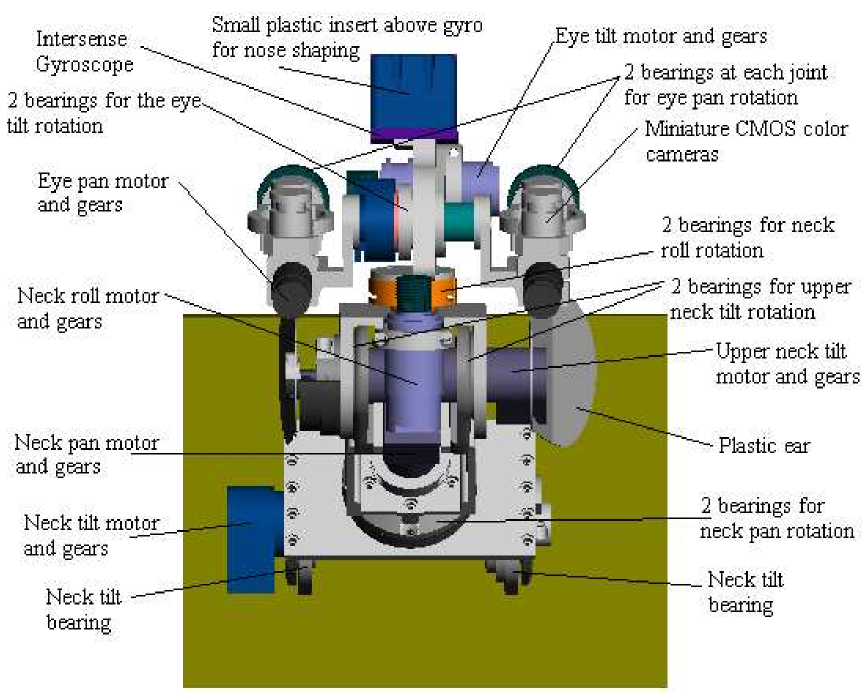


Figure 8. Top view of M4 head, with a description of several elements.

The *SolidWorks* software package was used for the design, and simulations of both the kinematics and dynamics were carried out using the *WorkingModel* simulation software. All the mechanical drawings of the parts are being generated by *SolidWorks*. The simulation in *WorkingModel* was very useful to determine motor power requirements, stresses on the structure, and mechanical and material properties, among other features, such as detection of colliding elements.

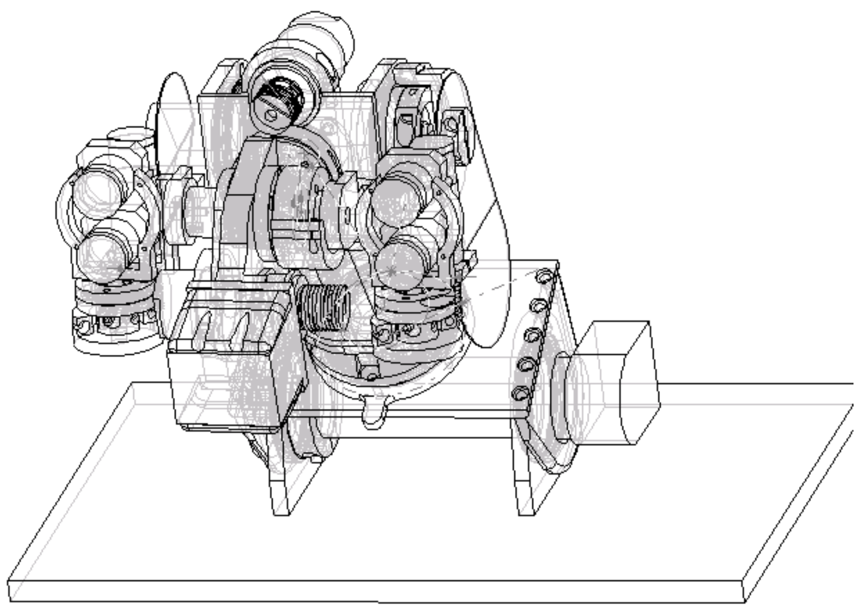


Figure 9. SolidWorks wire view of the head.

Computational Requirements

In order to build a hardware architecture, it is essential the definition of the communication requirements (data transfers) between the processing modules:

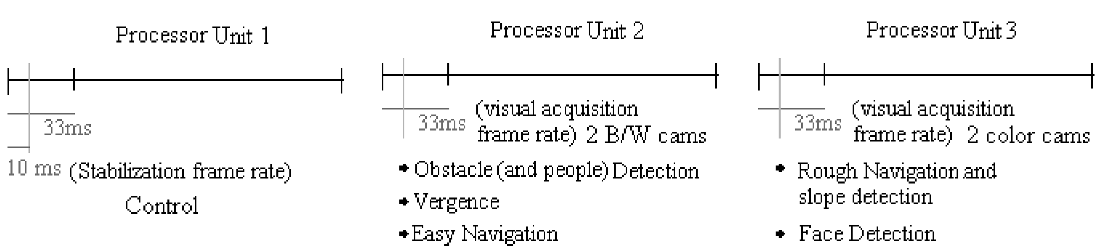


Figure 10. Processing Units.

- Image size - 128pixels×128pixels. B/W image - 16384 bytes. 24 bits per pixel color image - 49152 bytes. So, this gives 133 Kbits for B/W and 400 Kbits for color cameras, that corresponds to 32Mbps.
- Processing unit 1 does not send data to other processing units besides motor controllers. It receives data from inertial sensors (Gyroscopes and Accelerometers mounted on the head), and from joint encoders. This processing unit will also receive visual slip information (two integers or two floats) from processing unit 2 (vergence module). No bandwidth constraints on Processing unit 1, since data transfers are computationally insignificant.
- Processing unit 3 has to process two color images in 33 msecs, while Processing unit 3 process two B/W images. These units may have to share visual information, thus requiring a fast Ethernet link between the processors.

The stabilization algorithm has to be executed in less than 10 mili-seconds. Since motors can be controlled at much higher frequencies, the same processor may be used for both tasks, because motor control is very fast. Therefore, there are still available ~ 10msecs for the stabilization algorithm. The algorithm implementation allows for saccades and smooth pursuit implementation (but at a frame rate at 30 frames per second), and thus the same processor may carry these computations.

Vergence must run at 30 frames per second in another processor, as well as the easy navigation module. It would be desirable to detect obstacles (and people) at a frame rate of 30 frames per second. Nevertheless, obstacle detection must run at a speed higher than 15 Hz.

The rough terrain and slope detection module may operate at slow frequency, namely ~ 8Hz (but might run at ~ 2Hz if necessary because of computational requirements). Since this algorithm will not run when the face detector becomes active, the same processor may perform both tasks.

Therefore, three main processing units are enough to carry the necessary processing. Additional processing may be required if required for image acquisition. Another approach could be the use of a processing unit for each eye and a third one for the controls and stabilization. However, this solution would increase the transfer of images between processors, and would result in a decrease in flexibility.

A further alternative is the distribute processing of the images, by splitting several pieces, and distribute each piece to a different processor. This would require more processors and the development of a distributed parallel system.

Electronic Hardware

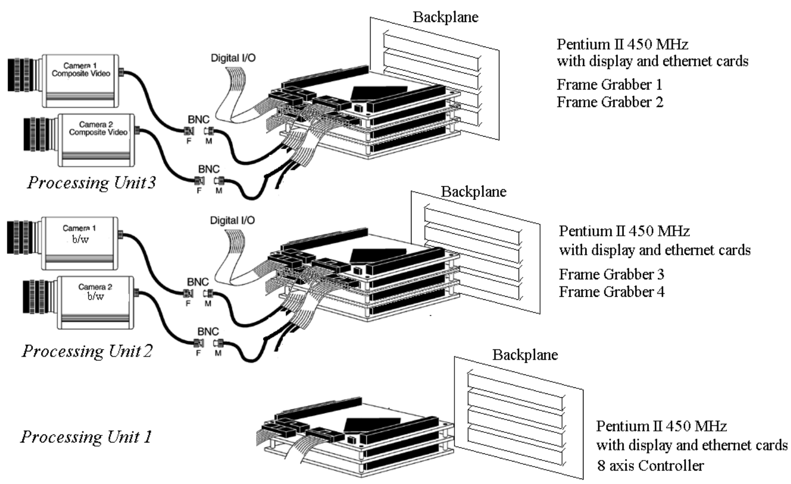


Figure 11. Hardware architecture.

Various solutions were considered for the design of the system hardware to accomplish the requirements. The size of the desktop motherboards (ATX) does not allow a system consisting on such cards to be mounted inside the vehicle. Therefore, smaller cards with Pentium processors will determine the choice. Another requirement is the bus bandwidth. Although exist frame-grabbers on the market that acquire real-time images for ISA bus (8bit, ~5-8 Mbytes of bandwidth), PCI bus compatible frame grabbers are mandatory for the acquisition of four images in real time.

Since QNX is the operating system that will be used with this solution, the frame grabbers selected are QNX compatible. In addition, the manufacturers of both the motherboard and the Ethernet card have also available drivers for QNX. Thus, the CPU card selected has the following features:

Company / Model	**Processor**	**Chipset**	**Price (without RAM)**	**Weight ***	**Size (in inches)**	**Ethernet/ QNX support & driver**	**Display / QNX support & driver**
Hester HS-6200	PII 450MHz (w/L2 cache)	Intel 82443BX	$1175	375lb all	7.36 x 5	Intel 82558A QNX — net.ether82557	C&T 69000 (2MB) QNX –Photon 113 CRT/ LCD

* Minus bracket s weight

Backplane	**Peripheral Bus**	**RAM (not included on CPU card)-max**	**HD Control**	**BIOS**	**Stock**
$68 4 slots —1CPU & 3PCI or $70 4 slots - 1CPU & 2PCI & 1 ISA 0.375 lb	PISA (PCI & ISA)	2 DIMM =1 GB /SDRAM/PC100 access t: 10ns = 100MHz 128 MB - $167 - 168 pin Accepts unequal amounts in each socket. Not all sockets need to be filled. Expandable	2 x Ultra DMA33 —IDE	Plug and play Y2K compliant Award BIOS	Yes Delivery in 5 days

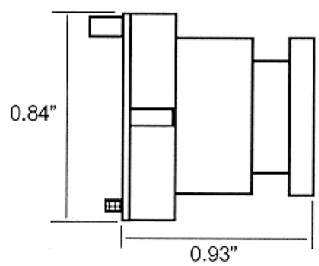
HD Control	**BIOS**	**Stock**	**Serial ports**	**Memory on chip**	**CPU Bus**	**.**	**Trial Period**
2 x Ultra DMA33 — IDE	Plug and play Y2K compliant Award BIOS	Yes Delivery in 5 days	2 RS 232 1 E-parallel port	2-72 MB 72 MB - $390 Msystems Flash MD-2201-D72	100 MHz	0-55 C	30 days eval. 1 Year warrantee

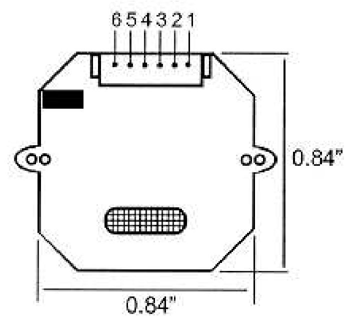
Option: PC104 / PC104plus bus

The PC104+ cards (3.8"x3.6"x2") are extremely modular - as many stacks as desirable may be joined to the system — and because of recent developments, are still considered as an alternative for the hardware system.

Visual acquisition

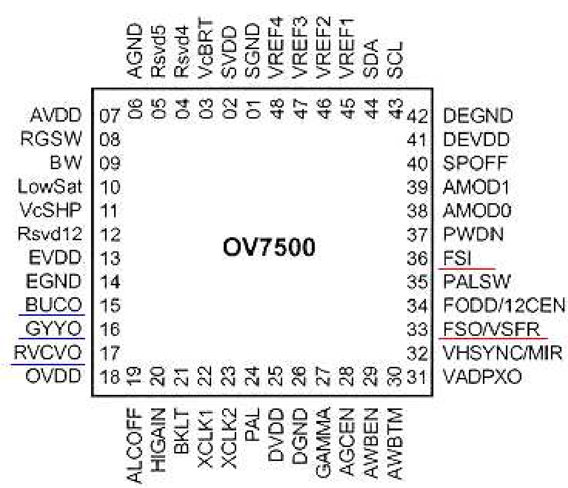
Features of the miniature CMOS color camera V-X0095-LH, with external synchronization.





SPECIFICATIONS	
Imager	CMOS Sensor
Format	1/3"
Array Size	NTSC - 510 × 492 PAL - 628 × 582
Resolution	310 TV Lines
Scanning	2:1 Interlace
Video Output	2 Vp-p, 75 ohm
Auto Shutter	1/60 to 1/15,000 sec.
Image Area (mm)	NTSC - 4.95 × 3.54 PAL - 5.96 × 4.23
Gamma Correction	0.45 - on/off
S/N Ratio	>38dB
Minimum Illumination	15 lux (f1.4)
Supply Voltage	5VDC
Power Requirements	150mW

Pin Diagram



In addition, it also includes on-chip auto exposure for automatic light compensation, switch able gamma correction, gain control, aperture correction, and auto white balance.

The lens selected for the foveal image is the V-4308, for which we get a field of view of 31H per 23V (in degrees, V-vertical and H-horizontal). The maximum relative aperture is 2, and the focal length is 8mm. For the wide range image, the lens selected is the V-4308, for which we get a field of view of 110H per 83V, maximum relative aperture of 2, and a focal length of 2.1mm.

Motors, gearheads and encoders

Neck base tilt

Maxon DC motor RE025, Graphite brushes, 20 Watt, number 118752 (24V of nominal voltage). Connected to a planetary gearhead 032mm, with a ratio of 190:1 (number 114483). Average backlash < 1.3 degrees. The digital encoder used is the HEDS 5540, 500 counts per revolution (CPT or CPR), 3 channels.

Neck upper tilt and neck pan

Maxon A-max 022mm, Graphite brushes, 6 Watt, number 110164 (24V of nominal voltage). Connected to a planetary gearhead 022mm, with a ratio of 370:1 (number 110340). Average backlash < 1.3 degrees. Digital encoder 22mm, 100 CPT, 2 channels (due to the high gear ratio, this encoder will give enough resolution on the measurements).

Neck roll and eyes tilt

Micro Mo DC series 2224R, 3.8 Watts. Gearhead series 22B, with a gear ratio of 27:1 (for the eyes tilt motor) and 81:1 (for the neck roll motor). These ratios are increased by a factor of 2.35 because of the capstans used on the head. Backlash < 3degrees (which is reduced by the tension cables). Micro Mo magnetic encoder IE2-512 with 512 CPR, 2 channel digital output.

Eyes pan

Micro Mo DC series 1524R, 1.4 Watts. Gearhead series 15/5, with a gear ratio of 22:1. These ratios are increased by a factor of 2.14 because of the capstans used on the eyes. Backlash < 4degrees (which is reduced by the tension cables). Micro Mo magnetic encoder IE2-512 with 512 CPR, 2 channel digital output.

Bearings

MPB precision bearings model:

Eyes pan: *sr4h* (.196in .25in .625in)
Eyes tilt and neck roll: *s1014* (.156in .625in .875in)
Upper neck tilt: *s2024* (.156in 1.25in 1.5in)
Pan neck: *s2428* (.156in 1.5in 1.75in)
Tilt base neck: *s2226* (.156in 1.375in 1.625in)

The size is indicated by (height inner radius outer radius). Two bearings are used at each joint.

Software

Real-time Operating System

It is required a multi-tasking operating system. Therefore, QNX was the operating system selected (and compatible with all the hardware selected), which is available for Pentium processors. This real-time operating system has been widely used in AILab projects.

Programming Language: C / C++.

Future Work

Once the robot is fully assembled and working, there is a lot of opportunity to make the vision system more robust to greater terrain variations. It can be made adaptable to increased agility, speed, and sensor acquisition of the robot. It may also be possible to implement learning strategies to enhance this flexibility. In addition, a thermal camera will be added to the robot, facilitating person and obstacle detection and enabling night vision.

Bibliography

[1] Arsenio A. Active Vision Head Stabilization — A redundant manipulator model (unpublished). AILab-MIT, 1998.

[2] Arsenio A. and Jessica B., People Detection and Tracking by a Humanoid Robot (unpublished). AILab-MIT, 1999.

[3] Arsenio A. and Jessica B. An Active Vision System for a Mobile Robot. MIT 1999 AILab abstract.

[4] Bach-Y-Rita, Collins. The Control of eye movements , Academic Press, 1971.

[5] Baron-Cohen S. Mindblindness: An essay on autism and theory of mind. The MIT Press, 1995.

[6] Bernardino A., Santos-Victor J., Visual Behaviours for Binocular Tracking. Robotics and Autonomous Systems, Elsevier, 1998.

[7] Beymer D., Face Recognition Under Varying Pose (unpublished). Center for Biological and Computational Learning, 1993.

[8] Brooks, R. A. "A Robust Layered Control System for a Mobile Robot," IEEE Journal of Robotics and Automation, Vol. 2, No. 1, March 1986, pp. 14--23; also MIT AI Memo 864, September 1985.

[9] Brooks, R. A., Intelligence Without Representation, Artificial Intelligence Journal (47), 1991, pp. 139--159.

[10] Brooks, R.A., C. Breazeal (Ferrell), R. Irie, C. Kemp, M. Marjanovic, B. Scassellat and M. Williamson, Alternate Essences of Intelligence, AAAI-98.

[11] Brooks, R. A., Intelligence Without Reason, Proceedings of 12th Int. Joint Conf. on Artificial Intelligence, Sydney, Australia, August 1991, pp. 569--595.

[12] Brooks, R.A. and L.A. Stein, Building Brains for Bodies, Autonomous Robots, Vol. 1, No. 1, November 1994, pp. 7-25.

[13] Brunelli R. and Poggio T. Face Recognition: Features versus Templates. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1993.

[14] Brunelli R. and Poggio T. Face Recognition Through Geometrical Features. ECCV92, 1992.

[15] Coombs. Real-time Gaze Holding in Binocular Robot Vision , PhD Thesis, University of Rochester, June 1992.

[16] Craig. Introduction to Robotics. Addison-Wesley, 1986.

[17] Damasio, A. Descartes' error: Emotion, reason and the human brain. *New York: Grosset/Putnam Book*, 1994.

[18] Faugeras O. 3D Computer Vision: A geometric viewpoint. MIT Press, 1992.

Craig J. Introduction to Robotics. Addison-Wesley, 1986.

[19] Ferrell, C. Orientation Behavior Using Registered Topographic Maps. Proceedings From the 4th International Conference on Simulation of Adaptive Behavior Conference, Cape Cod, MA, 94 103, 1996.

[20] Fitzpatrick P., Space-variant image sampling for foveation (unpublished). 1998.

[21] Goldstein E. Sensation & Perception. Brooks/Cole, 1996, 4th Ed.

[22] Hashimoto S. et al. Humanoid Robots in Waseda University — Hadaly-2 and WABIAN, IARP First International Workshop on Humanoid and Human Friendly Robotics, October 26-27, 1998 Tsukuba, Japan.

[23] Horn B. Robot Vision. MIT Press, 1986.

[24] Inoue H. A Platform-based Humanoid Robotics Project, IARP 1st International Workshop on Humanoid and Human Friendly Robotics, October 26-27, 1998 Tsukuba, Japan.

[25] Kandel, E., Schwartz, J. & Thomas, J. Essentials of Neuroscience and Behavior. In Appleton & Lange (Ed.) Norwalk: Connecticut, 1995.

[26] Konen W. Comparing facial line-drawings with gray-level images: A case study on PHANTOMAS. ICANN, 1996.

[27] Kwon Y. and Da Vitoria Lobo N. Face Detection Using Templates.

[28] Lorigo L., Brooks R., Grimson W. Visually-Guided Obstacle Avoidance in Unstructured Environments. IEEE International Conference on Intelligent Robots and Systems, 1997.

[29] Murray, Li, Sastry. Robotic Manipulation , CRC Press, 1994.

[30] Nastar C., Moghaddam B. and Pentland A. Generalized Image Matching: Statistical Learning of Physically-Based Deformations. Fourth European Conference on Computer Vision, 1996.

[31] Oren M, Papageorgiou C., Sinha P., Osuna E. and Poggio T. Pedestrian Detection Using Wavelet Templates, Proceedings of CVPR'97, June 17-19, 1997, Puerto Rico.

[32] Osuna E., Freund R. and Girosi F. Training Support Vector Machines: an Application to Face Detection, Proceedings of CVPR'97, June 17-19, 1997, Puerto Rico.

[33] Osuna E., Freund R. and Girosi F. Support Vector Machines: Training and Applications, AI Memo 1602, March, 1997.

[34] Pinker S. How the Mind Works. W.W. Norton & Company, Inc., NY. 1997.

[35] Reid I. and Murray Q. Active tracking of foveated feature clusters using affine structure. IJCV, 1996.

[36] Reisfeld D., Wolfson H. and Yeshurun Y. Detection of Interest Points Using Symmetry. ICCV90, 1990.

[37] Santos-Victor. Visual Perception for Mobile Robots : from Percepts to Behaviours , IST Ph.D. Thesis. 1994

[38] Brian Scassellati. Eye Finding via Face Detection for a Foveated, Active Vision System. AAAI-1998.

[39] Brian Scassellati. A Binocular, Foveated, Active Vision System. MIT-AI Memo 1628, January 1998.

[40] Scassellati B. Imitation and Mechanisms of Joint Attention: A Developmental Structure for Building Social Skills on a Humanoid Robot. Springer-Verlag, 1998.

[41] Schyns P. and Bulthoff H. Conditions for Viewpoint Dependent Face Recognition.

[42] Sinha P. Object Recognition Via Image Invariants: A Case Study. In Investigate Ophthalmology and Visual Science, Florida, 1994.

[43] Sinha P. Perceiving and Recognizing Three-Dimensional Forms. PhD Thesis, MIT, 1995.

[44] Tyler C. and Miller R., Computational Approaches to Face Recognition.

[45] Yuille A., Cohen D. and Hallinam P. Feature extraction from faces using deformable templates. International Journal of Computer Vision, 1992.