Proceedings of the 1993 IEEE/RSJ International Conference on Intelligent Robots and Systems Yokohama, Japan July 26-30, 1993

Selective Attention Mechanisms in a Vision System Based on Neural Networks

Michele Rucci and Paolo Dario

ARTS Lab
Scuola Superiore S. Anna
Pisa, Italy

Abstract

In this paper we propose a system for visual recognition derived from a recently developed theoretical framework on the overall organization of the human visual system. The system operates dynamically by analyzing different parts of the input scene at variable levels of resolution through an attentional spotlight. A constant amount of information is gathered from the scene and a fixed dimension icon is produced, so that a trade-off is set between the extension of the examined area and the level of resolution at which data are analyzed. The position of the spotlight and its dimensions are determined on the basis of the evolution of the recognition process. The icon is processed by a bottom-up path which is composed of a five-layer artificial neural network. The results of this net are analyzed by a planning module which determines if recognition has been achieved, or which action to undertake next. Finally, a top-down path, including a set of nets trained by the back-propagation algorithm, evaluates the parameters of the next sampling of information. The application of the system to the case of object recognition with varying view-point and range from the camera is investigated.

1. Introduction

Human vision works in a dynamic world characterized by large variability both in space and in time. As a consequence, an incredible amount of visual information falls at every time on our eyes. Nevertheless, we survive well in this world and we interact with the environment seemingly without any effort. A significant contribution to this result comes from selective visual attention, that is our capability of processing differentiately simultaneous sources of visual information [1]. By means of selective attention we can discriminate among input data so as to attend to the crucial information for the task at hand and ignore the irrelevant.

Selective attention can be seen as a mechanism for the proper allocation of a limited set of resources to the processing of a large amount of data [2]. In this respect attention mechanisms are extremely important in order to build a robotic system which is able to work in unknown environments. In fact, in spite of the great progresses in computer technologies, the computational power of current computers is still far lower than that of the mind. This implies that resources available to process perceptual data are very limited. Surprisingly, apart from some noticeable exception [3], researchers in computer vision have not dedicated much effort so far to the topic of selective visual attention [4].

Recent neurophysiological findings and psychological models have emphasized that Selective Attention (SA) can play a fundamental role in visual recognition [5], [6]. We believe that these results can be an important guidance to lead computer vision researchers towards the proper design of robotic vision systems.

Based on these considerations, in this paper we propose a neural network-based system for visual recognition which makes use of some mechanisms of selective attention. The system is derived from a theoretical framework about the organization of the human visual system developed recently by Nakayama [7]. Thanks to the adaptability and flexibility provided by SA we have observed that a scale-invariant and view-point independent visual recognition process is possible.

In the following section, the main properties of SA mechanisms in humans and the framework proposed by Nakayama are briefly reviewed. In section 3 the implementation of this architecture by means of Artificial Neural Network (ANN) techniques is described; finally, in section 4 the system application to the problem of view-point independent object recognition is analyzed.

2. Selective Attention and Visual Recognition

Many psychophysical experiments have contributed to spread the idea that SA in visual perception has the characteristic of a "limited extent" spotlight [8] [9]. The processing of the stimuli included in the spotlight beam seems to be facilitated with respect to external data [10]. Other experiments have shown that the spotlight can be moved independently on eye fixations [11] and it can vary in size [12].

A very important problem concerns the control of the attentional spotlight in time, that is which factors contribute to draw attention in a given situation. A fundamental distinction was already made in 1890 by William James [13], when he noticed that visual attention can be drawn by stimuli at least by two factors: from one side attention can be

0-7803-0823-9/93/$3.00 (C) 1993 IEEE

drawn automatically by the sensory characteristics of the stimuli, and from the other side attention can be drawn by the semantic characteristics of the stimuli. This opposition, which corresponds to a distinction between low-level and high-level mechanisms, seems to have recently received an experimental support by the finding of two separate temporal components contributing to the shifts of visual attention [14] [15].

It is an old controversy whether visual recognition is a parallel, one-step process or a serial, step-by-step one controlled by attention mechanisms [16]. The parallel process was maintained by the Gestalt school, with the assumption that visual recognition involves a single matching of the whole object with its internal representation. On the contrary, the serial process implies that many matches of the object parts and features are required to recognize and that the object internal representation is an assemblage composed of the representations of all the parts.

Recently, a theoretical speculative framework of the overall structure of the human visual system supporting the step-by-step hypothesis has been proposed by Nakayama [7]. This framework outlines the relationships between early vision and visual memory, and the role played by selective attention in visual perception. According to this model the visual system can be regarded as consisting of two parts, as shown in Fig.1: a feature pyramid and a visual memory. The feature pyramid is a massively parallel structure which receives afferent visual inputs and which works on different features at different levels of resolution. At the other extreme of the visual system there is the visual memory which consists of a large amount of tiny icons linked associatively. The linkage between these two systems is a narrow bandwidth channel having a capacity of the order of a single icon. In this framework selective attention moves the iconic channel in the feature pyramid, so as to sample subsequently the information from different parts of the scene and at variable level of resolution.

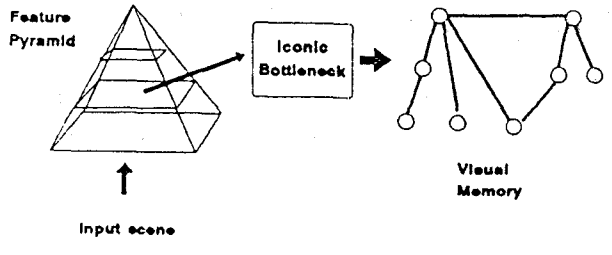

Fig.1: Nakayama's theoretical framework. The feature pyramid and the visual memory are linked by a narrow bandwidth channel: the iconic bottleneck.

The system described in this paper takes inspiration from Nakayama's framework: the feature pyramid, the iconic bottleneck and the associative memory, all have functional counterparts in the nets implementing the system we propose. The simulation of semantic, high level control processes of selective visual attention in visual recognition is the main focus of the system. Even if the underlying philosophy is inspired to the biological system, in the actual implementation of the system biological plausibility was not the main goal. Nevertheless, some neurophysiological results were used as design guide in the development of the system, as will be shown in the following section.

3. A neural network based architecture

We believe that Artificial Neural Networks (ANNs) are suitable tools for the implementation of artificial perception systems. Such peculiar features of many ANNs paradigms as their relatively high tolerance to noise, and to their own faults and defects [17] [18], are very important properties when building a computer vision system. Furthermore, ANNs have the intrinsic capability of processing simultaneously different sources of information such as edges and regions, so as to overcome the difficult problem of integrating the results produced by different algorithms, each one working on a different kind of information [19].

There are also some practical considerations that encouraged us to use ANNs, most of which are related to their speed and flexibility: for example, the application of the system described in this paper to different recognition problems requires only a few short additional training sessions.

The basic aim of the system described in this paper is to achieve a robust and fast recognition in a partially controlled environment, independently on the distance and the position of the analyzed object with respect to the camera. The system operates dynamically, so that the information processed at a given time drives successive fixations. In this way, the part of the scene examined at a given moment depends on the state of the system, that is on its past history during the recognition process. The examination of the scene or of a part of it can produce a series of hypotheses regarding object identity, which can be checked by looking for their salient parts and features.

A general scheme of the proposed architecture is shown in Fig.2. Basically, the architecture includes a bottom-up path, a planning module and a top-down path. The bottom-up path processes the information gathered through a variable-size moving attentional spotlight so as to activate decisions in the planning modules. Based on the results of the processing, the planning module determines whether recognition has been achieved or, if not, which part

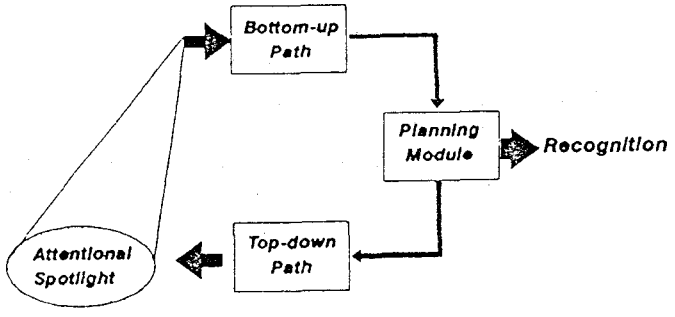

Fig.2: Scheme of the proposed system. The bottom-up path, the planning module and the top-down path are arranged in a loop so as to process dynamically the input scene.

of the input scene should be examined. The actual parameters of next attentional fixations (position and size of the selected area) are determined by means of the top-down path.

The representation of each object to recognize is distributed among the modules of the system. An example of object representation is depicted in Figure 3. In the figure an object is represented by means of a set of icons stored in the connections weights of the bottom-up path whose spatial structure is stored in the top-down path. A set of units acting as grand-mother cells for these representative icons are also present in the bottom-up path . These units, together with another unit sensitive to the object being considered, form a sort of high level representation, as can be observed in the Figure. The object representation is normalized with respect to scale factors. It will be shown later on that, if the spatial arrangement of the icons and their relative dimensions are also coded relatively to one or more reference dimensions of the attentional spotlight are properly set, objects differing only in size can produce similar icons. The icons, so that if the size of these icons has been determined the dimensions and positions of all the others can be singled out. The representation used here is similar to the Iconic Associative Memory of the theoretical framework described in Section 2. Also in this case, the use of a fragmentary representation gives a great contribute toward the goal of robust recognition.

3.1 The Bottom-up path

The input scene is analyzed through a moving attentional spotlight of variable size. Data included in the spotlight beam are preprocessed so as to replicate the iconic attentional bottleneck, whereas the data outside the beam are not considered. The amount of information gathered is limited to a constant value producing an icon whose dimensions are fixed and do not depend on the size of the attentional spotlight. An inverse proportionality is built between the attentional spotlight dimensions and the level of resolution at which the part of the scene is examined: an increment of the width of the analyzed area results in a corresponding decrement of the level of resolution at which the data are examined. In this way a multiresolution pyramid is dynamically simulated by producing at each moment its part of interest. The data sampled by means of the attentional spotlight are preprocessed in the bottom-up path: at first, the edges of the examined area of the input scene are extracted with a gradient operator. As shown in Figure 4, the resulting image is then re-sampled at a lower resolution by examining it through gaussian receptive fields, and the dynamics of the resulting icon is expanded so as to emphasize the parts of the examined area with stronger edges and/or higher concentration of edges.

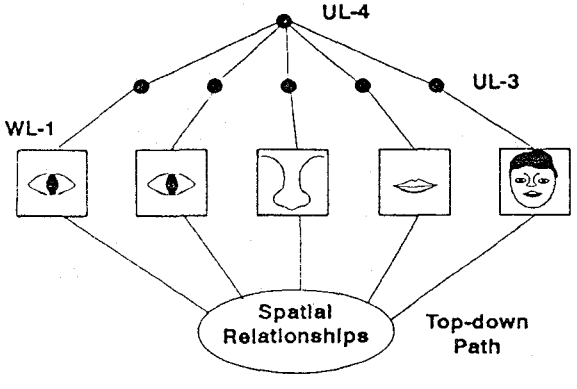

Fig.3: The object representation is composed of a set of icons stored in the bottom-up path and of their spatial relationships included in the top-down path. In addition, there is a high level representation of units sensitive to the representative icons.

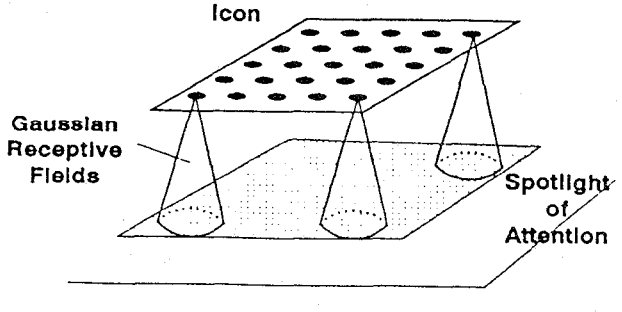

Fig.4: An icon is produced by re-sampling the image by means of gaussian receptive fields. The gaussian filtering reduce the aliasing due to the decimation.

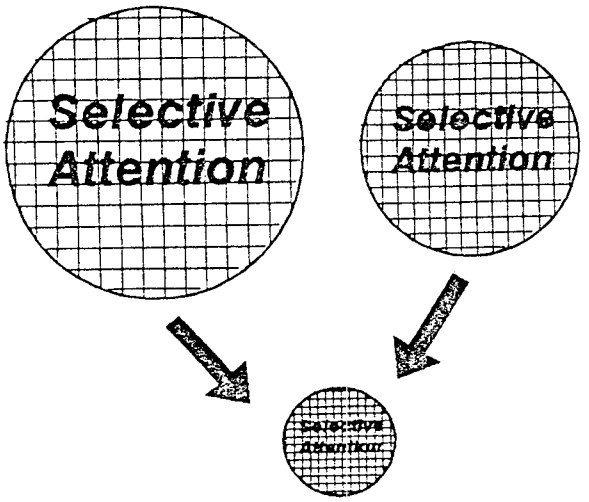

Fig.5: The basis for scale-independent recognition. When the size of the attentional spotlight is properly set, objects with different scale factors produce similar icons.

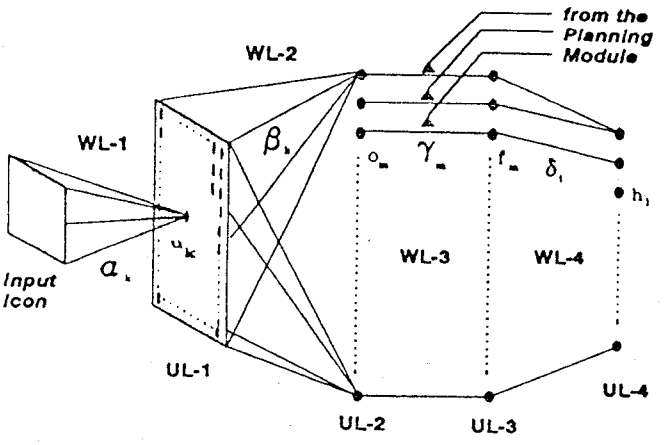

Fig.6: The bottom-up path is a feed-forward neural network composed of five layers.

The width of the gaussian receptive fields, which depends on the spotlight dimensions, is automatically adjusted by setting the proper value of σ.

The automatic adjustment of the receptive fields of the icon respect to the size of the examined area is the basis for scale-invariance in the recognition process. Figure 5 illustrates this concept by showing how objects having different size produce the same icon if the spotlight dimensions are properly set. According to the definitions introduced in Section 2, the initial width of the spotlight beam is established on the basis of low-level attention mechanisms; The dimensions of the spotlight in successive fixations are then evaluated according to this initial width and to the relative width of the examined part stored in the memory of the system. Although low-level attention mechanisms are not the main focus of this paper, an example of how the initial width of the attentional spotlight can be evaluated in the case of object recognition is shown in Section 4.

The processed icon is accepted as input by the five-layered neural network depicted in Figure 6 (the five layers of units include the input layer). As illustrated in the figure, the units in layer UL-1 are arranged in a square matrix whose number can change with the application (a 15x15 array was used in the experiments described in section 4). Layers UL-2 and UL-3 have the same number of units, and are connected in a 1-1 fashion. The number of units of layer UL-3 is equal to the number of objects to recognize. The net can be split in two main parts: the first three layers act as a classifier for the incoming icon, while the last two layers can be seen as the long term memory of the system. The first part of the net is similar to the counterpropagation network [20] with a self-organized topological feature map [21], [22] in the second layer followed by a sorting layer.

The topological feature map performs a matching operation between the processed icon and a set of generalized icons stored in the first layer connection weights (WL-1) during the training phase. The main difference between the first two layers of the bottom-up path and a self-organized topological feature map as proposed by Kohonen, is the lack of connections among the units of UL-1; this fact explains the absence of blob of activation in this layer.

The first layer of weights of the net WL-1 shares with a self-organizing map the same training procedure: an unsupervised learning process which requires the recurrent presentation of patterns of the training set to the net. During the training, the weight vectors of the maximally responding unit and of the units in a neighborhood of it are modified towards the input vector, whereas the size of the neighborhood and the parameters regulating weight changes decrease.

The icons used for training WL-1 have the attentional spotlight centered on the significant parts of the imaged object which are used for its representation. What is expected is that the map of UL-1 will learn to discriminate among different features by producing several disconnected areas where units are sensitive to icons of the same feature.

As in [23] the activation U_k of unit u_k in layer UL-1 is a measurement (comprised between 0 and 1) of the distance between its own weight vector w_k and the input icon $\mathbf{p}$.

$$U_k = \sigma(s_k) = \frac{1}{1+\exp(-\beta_u(\delta_u - s_k))} \tag{3.1}$$

$$s_k(p) = \begin{cases} 1 - \|p - a_k\| & \text{if } \|p - a_k\| < d_m \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

where d_m is an empirically chosen distance.

In the network of the bottom-up path, only the weights of the topological feature map are calculated by means of training. The weights in all the other layers are evaluated directly by algebraic calculations. The weights β_{km} (k=1,...K; m=1,...,M) of the layer WL-2 between the second and the third layer, are arranged in order to map the activation of the topological feature map so that each unit o_m in UL-3 is sensitive to a specific feature. For each weight β_{km}, the mean sensitivity of unit u_k to feature m is evaluated by averaging the output of the unit to all the icons representing the considered feature. In this way a vector η_k composed of the mean sensitivities of unit u_k to all the features is defined. The weight β_{km} between the units o_m in UL-2 and the u_k in UL-1 is evaluated as the normalized m-th component of vector η_k

$$\beta_{km} = \eta_{km} / \|\eta_k\| \qquad (3.3)$$

As in layer UL-1, also the units of UL-2 have a sigmoid activation function, so that the output O_m of unit o_m is given by

$$U_k = \sigma(s_k) = \frac{1}{1+\exp(-\beta_o(\delta_o - (\sum U_k\beta_{km} - \sigma(0)\sum\beta_{km})))} \qquad (3.4)$$

The first sum in the exponential of equation (3.4) is the net input to unit o_m provided by the topological feature map; the second sum is a unit-dependent threshold which has been introduced in order to eliminate the low resting activation of the units in UL-1.

The weights δ_k in layer WL-3 take binary values (0 or 1) and are controlled by the planning module according to the followed strategy. If the system is looking for a specific part of an object, in order to test a previously formulated hypothesis, only the unit sensitive to that feature is allowed to receive activation by the planning module; all the others δ_k are set to zero. This operation reduces the error probability by biasing the system toward the search of a particular stimulus. It should be pointed out that such a modulation effect on neuron response due to selective attention has been recently found by neurophysiologists in the visual cortex of the monkey [5].

Units in layers UL-3 and UL-4 act as Grand-Mother cells, i.e they are sensitive to specific input patterns. In particular, units in UL-3 are sensitive (as units in UL-2 to which they are connected in a 1-1 manner) to specific features of the object to recognize, and each unit in UL-3 responds to a specific object independently on the examined object feature. An abstraction process can be noticed in the bottom-up path, due to the fact that the units in successive layers are sensitive to increasingly generalized stimuli.

When put together, units in layers UL-3 and UL-4 build up a high-level representation composed of the Grand-Mother Cells of each object to recognize (see Figure 6); in this respect they constitute the long term memory of the system. In this high level representation every object to recognize is represented by a set of units in UL-3 (a sort of feature layer), each one standing for a feature or a part of the object, and by a single unit in UL-4 (a hypothesis layer) connected with the previous ones in an excitatory manner. As illustrated in Figure 6, each unit of UL-3 is activated by the corresponding unit of UL-2 when the spotlight of attention is centered on the feature which the unit represents and the size of the beam is such to completely enclose that feature. The structure of this high-level representation is similar to the scheme proposed by Burt [24].

Units in layer UL-3 act as accumulators, by storing and accumulating the activation provided by UL-2 units. The output of unit f_m is given by the sum of all the outputs of o_m multiplied by the connection weight γ_m on all previous attentional fixations.

$$F_m(t) = \Sigma_t\, \gamma_m(t)\, U(O_m(t)) \qquad (3.5)$$

where

$$U(x) = \begin{cases} x - S_m & \text{if } x > S_m \\ 0 & \text{otherwise} \end{cases} \qquad (3.6)$$

The function U(x) and its threshold S_m have been introduced in order to eliminate the resting activation of unit o_m. The activation of each UL-3 unit f_m can be set to zero by the planning module when the system fails to test a previously formulated hypothesis. As will be explained in the following, in this case the planning module resets all the units of UL-3 which are sensitive to features of the rejected object, by means of the returning inhibitory connections shown in Figure 7.

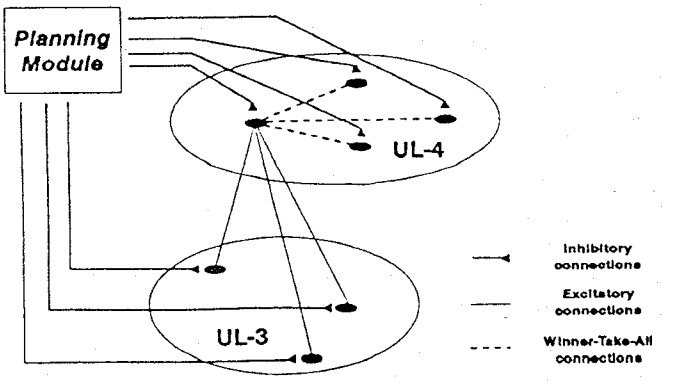

Fig.7: The inhibitory feedback net. By means of these connections the Planning Module can reset units in layers UL-3 and UL-4 of the bottom-up path.

The connections in layer WL-4 link those units sensitive to the same object, as displayed in Figure 6. The weights of these connections have positive fixed values whose aim is to bring excitation to the Grand-Mother cells in UL-4.

Units in layer UL-4 inhibit each other in a Winner-Take-All (WTA) fashion [25] [26]. Based on the WTA equations, the activation is allowed to diffuse among the units in layer UL-4 so that, after a short time, only one unit of the layer has a positive value of activation while all the others are inhibited (output equal to zero). This is equivalent to the formulation of an hypothesis on the identity of the observed object. From this point of view, recognition can be expressed as the eventual determination of a winning unit in layer UL-4.

It is important to notice that the network shows some phenomena peculiar to cognitive systems, such as generalization and abstraction processes. If significant, a part of an object can saturate the unit which represents that object in UL-4, thus producing recognition.

Furthermore, the use of Grand-Mother Cells is well suited for the interaction of neural and symbolic techniques, taking advantages of both methods. This allows the use of more traditional AI techniques, such as rule-based systems for high level analysis and for the interpretation of the results of neural modules.

3.2 The Planning Module

The results of the bottom-up path are examined by the Planning Module, which determines whether recognition has been achieved or, otherwise, which part of the hypothesized object the system should look for. As shown in Figure 6 and in Figure 7, the Planning Module interacts actively with the bottom-up path by means of a set of inhibitory connections.

The adopted control strategy assumes that the first sampling of information (driven by low-level attention mechanisms) is such to analyze the whole object. This implies that the attentional spotlight is centered on the object centroid and the beam size includes the object completely. The reason for this constraint is that the spatial structure of the object is coded according to reference icons that are created with the previous spotlight conditions. The first attentional icon and the corresponding spotlight parameters (position and dimensions) are stored in a short-term memory by the planning module when the corresponding unit in UL-4 is activated.

By means of the first attentional fixations a set of hypotheses on the identity of the observed object is formulated, i.e several units in UL-4 receive activation greater than zero. Due to the WTA connections in this layer, after a transient all the hypotheses other than the winning one are suppressed. The basic strategy followed by the Planning Module requires that the features of the hypothesized object are sequentially examined: recognition is achieved if the net input to the UL-4 unit is larger than a predetermined threshold. If a hypothesis cannot be tested by the successive attentional fixations, the Planning Module resets all the units of layer UL-3 which are sensitive to the features of the rejected object, and the WTA layer UL-4. In this way the second most probable hypothesis wins the competition and its features are then analyzed. The serial examination-inhibition cycle is repeated until recognition is achieved or all hypotheses are sequentially examined.

When a feature is chosen, the Planning Module performs two operations: a) it sets the weights γ_m in layer WL-3 of the Bottom-up path so that only those units corresponding to the expected feature are allowed to receive activation; and b) it activates the modules of the Top-Down path corresponding to the considered feature.

The introduction of a threshold on recognition accuracy allows a trade-off between recognition time and precision: if the threshold is high, more attentional fixations are required and the system is more reliable; on the contrary if the threshold is low, recognition can be achieved with a shorter number of attentional fixations, but the error probability is higher.

3.3 The Top-Down Path

As seen in paragraph 3.2, during the recognition process the Planning Module determines which part of the object to examine next. In order to move the spotlight in the corresponding part of the scene and to set the proper level of resolution a what-where conversion is required, i.e. the feature identity should be transformed in the corresponding spatial position. Two factors contribute to this transformation: an *apriori* knowledge of the spatial structure of the object (a top-down information which is included in the three-dimensional object representation), and the scene information (bottom-up information) which is included in short-term memory of the system with the first attentional fixation.

This transformation is carried out by the Planning Module which, on the basis of the feature selection carried out by the Planning Module, determines the location where to move the attentional spotlight in the scene, and the level of resolution at which data should be analyzed. As illustrated in Figure 8, the Top-down Path is composed of two modules: a Position Module (PM) and a Dimension Module (DM). The PM includes a four-layer full-connected feed-forward ANN trained by means of the back-propagation algorithm [27] for each feature of each object to recognize. The DM has the same number of storage cells. Only one network and one cell, those corresponding to the feature currently examined, are activated by the Planning Module and operate at a given time.

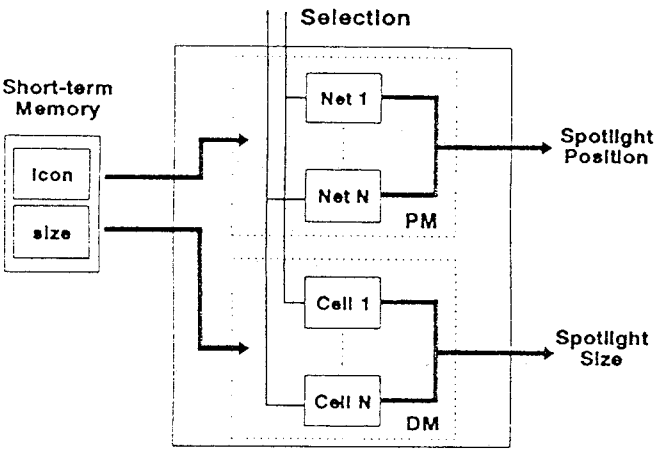

Fig.8: The top-down path is composed of a Position Module and of a Dimension Module which provides the spotlight coordinates and size, respectively.

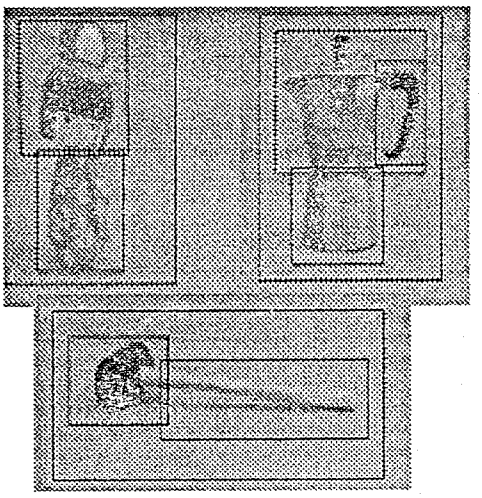

Fig.9: Three objects used for testing the system. The squares correspond to representative icons stored by the system.

Each ANN is trained to produce the coordinates of the centroid of the corresponding feature in a scale-independent reference system: the actual coordinates of the feature centroid are evaluated by multiplying the values produced by the nets by the scale factor. The nets accept as input the first sampled attentional icon stored in the short-term memory of the system and give the two values of the coordinate of the feature centroid. A sparse code based on M+1 output units (M is the number of pixels of the icon side) optimized experimentally has been used. All the nets have the same structure with MxM units in the input layer and M+1 units in the second layer, and two branches, each composed of M+1 units in both the third and the forth layers.

Each storage cell stores the relative dimensions (in the size-independent reference system) of the attentional spotlight when centered on the considered feature. Once again the true dimension of the spotlight is obtained by multiplying the relative dimension by the scale factor stored in the short-term memory during the first attentional fixation.

4. Object Recognition

The system described before has been applied to the case of view-point independent object recognition among a pre-established set of objects. The object could appear with different orientations and distance with respect to the camera, so that different features with different scale factors are visible in each presentation.

For each object a number of features have been manually selected in order to create the object representation. A set of objects used in this experiment (a small Japanese doll, an Italian coffee-maker and a hammer) are shown in Figure 9, along with their representative features The objects were located on a table in front of a homogeneous background and the illumination was arranged so as to avoid sharp shades and reflexes. Several images, 256x256 pixel wide were acquired by a black and white camera with 64 grey-levels.

The icons were squares of 20 pixels side and were processed by a bottom-up path composed of 400 units in UL-1 (a topological feature map with 20 units along each axis), 10 units in UL-2 and UL-3 (the same number of the features), and 3 units in UL-4 (the same number of the objects). The top-down path included 7 networks and 7 storage cells, one for each feature to locate); as discussed in section 3.3, each net was composed of 400 (20x20) units in the input layer, 21 and 21 units in the hidden layers, and 42 (21+21) units in the output layer.

A number of images for each object were used to train the nets of the system: windows located on the representative features were extracted from the images and their positions and dimensions were stored. The topological feature map in layer UL-1 of the bottom-up path was trained by means of the unsupervised learning procedure described in paragraph 3.1, where the icons corresponded to the selected windows. For the training of the nets included in the top-down path, the lowest resolution icons (that is the icons corresponding to windows covering the whole object) and the feature relative positions were used as input and output signals, respectively.

Typical values of unit parameters in the bottom-up path were β_u=0.89 and δ_u=0.5 which corresponds to a sigmoid function with s(0)=0.05 and s(1)=0.95.

The first attentional fixation was driven by a very easy low-level attention mechanism which was aimed to

produce the lowest resolution icons shown in Figure 10. This mechanism was implemented by subtracting the object image to the background so as to eliminate all the pixels that did not belong to the object. The width and the height of the window were then easily calculated.

The system exhibited good performances: percentages of correct recognition were 94.7% for the doll, 92.3% for the coffee-maker and 91.7% for the hammer. Recognition was robust due to the fact that object identity could be assessed also if some features did not provided activation. All the errors were due to the lack of activation of the correct hypothesis during the first attentional fixation.

This situation can be further improved by integrating other sensor modalities. An example of this approach, based on the integration of visual and tactile perception, is described in [28]

5. Conclusions

The replication of selective attention mechanisms, as evidenced by humans, is extremely important in the field Computer Vision. By means of selective attention a proper allocation of the available resources for the task at hand can be achieved. In this paper we have shown how a system for visual recognition based on attentive processes can be implemented by means of ANNs techniques.

Several research directions are being investigated. In particular, we are evaluating the limits of the system when the number of objects to recognize increases. Furthermore, the possibility of updating the system during the recognition process and of including also low level mechanisms for the control of attention is being considered. Finally, the integration of visual attention mechanisms with manipulation and touch-based active exploration [29] [30] is investigated as means to increase the understanding and the usefulness of artificial perception in advanced robotics.

Acknowledgements

This work has been supported in part by the Special Project on Robotics of the Italian National Research Council (CNR) and by the Italian Ministry of University and Research (MURST 40% and 60%). One of the authors (M. Rucci) has been supported by a fellowship from Istituto per la Ricostruzione Industriale.

References

[1] W.A.Johnston and V.J.Dark "Selective Attention", Ann. Rev. Psychol. 37, 43-75, 1986.

[2] R.A.Kinchla "Attention", Ann. Rev. Psychol. 43, 711-742, 1992.

[3] P.J.Burt "Smart Sensing within a Pyramid Vision Machine", IEEE Proceedings, 76, pp 1006-1015, 1988.

[4] R.D.Rymey, C.M.Brown "Selective Attention as Sequential Behavior: Modelling Eye Movements with an AHMM", Tech. Rep327, University of Rochester, New York, 1990.

[5] J.Moran, and R.Desimone "Selective Attention Gates Visual Processing in the Extrastriate Cortex", Science 229, 782-785, 1985.

[6] B.Olshausen, C.Anderson, D.Van Essen "A Neural Model of Visual Attention and Invariant Pattern Recognition:, CNS Memo 18, California Institute of Technology, September 1992.

[7] K.Nakayama "The iconic bottleneck and the tenuous link between early visual processing and perception", in Vision: Coding and Efficiency. C.Blakemore Ed., University Press, 1991.

[8] M.I.Posner "Orienting of Attention", Quaterly Journal of Experimental Psychology, 32, 3-25, 1980.

[9] M.I.Posner, C.R.R. Snyder, B.J. Davidson "Attention and the detection of signals", Journal of Experimental Psychology: General, 109, 160-174, 1980.

[10] S.Yantis "On analog movements of visual attention", Perception and Psychophysics 43, 203-206, 1988.

[11] H. von Helmholtz, *Psychological Optics*, J.P.C. Sothall (Ed.), New York:Dover, 1866/1925.

[12] C.W.Eriksen and J.D.ST.James, "Visual attention within and around the field of focal attention: A zoom lens model", Perception and Psychophyusics 40 (4), 225-240, 1986.

[13] W.James, *The Principles of Psychology*, Harvard University Press, Cambridge 1890/1983.

[14] K.Nakayama, and M.Mackben, "Sustained and Transient Components of Focal Visual Attention", Vis.Res. 11, 1631-1647, 1989.

[15] E.Weichselgartner, and G.Sperling "Dynamics of Automatic and Controlled Visual Attention", Science 238, 778-780, 1987.

[16] D.Noton, L.Stark "Eye Movements and Visual Perception", Scientific American, 224 (6), pp 34-43, 1971.

[17] P.D. Wasserman *Neural Computing: Theory and Practice*, Van Nostrand Reinhold, New York 1989.

[18] P.K. Simpson *Artificial Neural Systems*, Pergamon Press, New York, 1990.

[19] S.Grossberg, E.Mingolla and D.Todorovic "A Neural Network Architecture for Preattentive Vision", IEEE Trans. on Biomedical Engineering, 36, 1, pp 65-83, 1989.

[20] R.Hect-Nielsen "Applications of the Counter-propagation Networks", Neural Network 2, 1, 1988.

[21] T.Kohonen "Self-Organized Formation of Topologically Correct Feature Maps", Biol. Cybern. 43,59-69,1982

[22] T.Kohonen "Analysis of a Simple Self-Organizing Process", Biol.Cybern., 44, 135-140, 1982.

[23] R.Miikkulainen "Trace feature map: a model of episodic associative memory", Biol.Cybern. 66, 273-282, 1992.

[24] P.J.Burt "Attention Mechanisms for Vision in a Dynamic World", IEEE Int. conf. on , 1988.

[25] C.Koch and S.Ullman "Shifts in Selective Visual Attention: Toward the Underlying Neural Circuitry", in L.M.Vaina (Ed.) Matters of Intelligence, 115-141, Reidel Pub. Comp., 1987.

[26] J.K.Tsotsos "Localizing Stimuli in a Sensory Field using an Inhibitory Attentional Beam", Vision Research (in press), 1992.

[27] D.E. Rumelhart, G.E. Hinton and R.J. Williams "Learning internal representations by error propagation", in *Parallel Distributed Processing*, vol.1, pp 318-362, Cambridge, MA: MIT Press, 1986.

[28] P.Dario, and M.Rucci "An Approach to Disassembly Problems in Robotics", Proc. of IROS '93, Yokohama, Japan 1993.

[29] P.Dario, A.M.Sabatini, B.Allotta, M.Bergamasco, and G.Buttazzo "Object Characterization and Sorting by Active Touch", Proc. of IROS '91, pp.1353-1356, Osaka, Japan, 1991.

[30] P.Dario, A.M.Sabatini, B.Allotta, M.Bergamasco, and G.Buttazzo "A Fingertip Sensor with Proximity, Tactile and Force Sensing Capabilities", Proc. of IROS '90, pp.883-889, Tsuchiura, Japan, 1990.