Associating What and Where Using Temporal Cues

Nigel H. Goddard
Pittsburgh Supercomputing Center
4400 Fifth Avenue
Pittsburgh, PA 15213
ngoddard@psc.edu

Abstract

Johansson showed that people can recognize human gaits from brief presentation of only a few moving dots. A recently constructed connectionist model, MARS, is the first program of any type to model this phenomenon. One of the key ideas is that an association is formed between visual actions and spatial locations. Simulations show that in MARS the association mechanism is necessary for reliable recognition of multiple actions, and that the action-recognition process and the location association process act in consort to arrive at a stable interpretation of the image sequence. Association between location and action is performed in a spatiotopic network of cells that specialize in detecting temporal synchrony between visual events in the scene and predictions generated by active models of actions held in memory. The model suggests that such a mechanism may be used to build and maintain associations acquired sequentially.

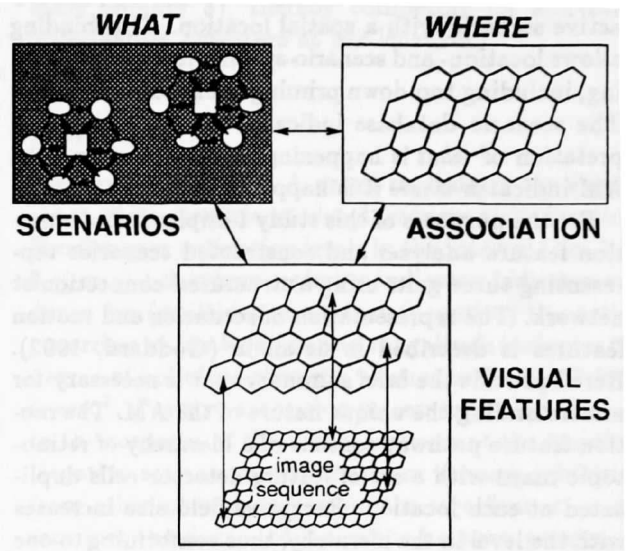

Figure 1: MARS Architecture

Action Recognition

Perception of articulated motion from impoverished image sequences has been repeatedly demonstrated by psychologists (e.g., (Johansson, 1973)). Common actions such as walking and running can be recognized with 500 msec presentation of these image sequences. Action recognition is a primary visual ability, yet it has received little attention in the modeling literature. In (Goddard, 1992) I presented a cognitive model of this recognition ability, called *MARS*. Here I describe in detail the one of the major connectionist[1] mechanisms used in MARS and demonstrate its function in the recognition process. This attention-like mechanism gives MARS the capability to identify multiple actions occurring simultaneously in a sequence of "retinal" images. Because of capacity limitations, MARS requires the mechanism to associate each action with a particular visual location. Neurological evidence suggests that the inability to recognize simultaneously presented objects (*simultanagnosia*) may be related to the impairment of such an association mechanism (Coslett and Saffran, 1991).

MARS's (Figure 1) consists of three modules: 1) a visual feature hierarchy that analyzes image sequences for static and motion parameters, 2) a high level representation of actions, the scenario hierarchy and 3) an association map (AM) which associates actions with spatial locations. Consider the example of presentation of a Johansson-like display of a person walking. These displays consist merely of a dot at the location of each limb joint (hip, knee, ankle, etc). The image sequence can be analyzed to recover trajectories of the dots (Olson, 1989), and the trajectories analyzed to recover the connected limb seg-

[1]Unlike many connectionist models, this one involves no learning, links between connectionist units are labeled, there are several classes of unit functions, and each unit has a significant amount of state information.

ments(Rashid, 1980). These transformations and further combinations in the feature hierarchy produce as output complex visual features representing uninterpreted features in the scene, for example a pair of connected line segments moving relative to each other. Their location is represented explicitly using an interpolation coding. The visual features index a hierarchical database of models of action, known as *scenarios*. Scenarios represent named actions that constitute a gait or other complex movement, for example "biped-walking". Recognition involves the scenario becoming synchronized with the action in the scene. The scenario representation does not encode location information. The AM uses input from the visual features that index the scenarios, together with the response of the scenarios, to bind each partially active scenario with a spatial location. This binding allows location- and scenario-specific focus of processing, including top-down priming of expected features. The scenario database indicates the system's interpretation of *what* is happening in the scene, and the AM indicates *where* it is happening.

For the purposes of this study I implemented a motion feature analyzer and constructed scenarios representing three gaits using a structured connectionist network. The representation of scenarios and motion features is described in detail in (Goddard, 1992). Here I provide the brief summary that is necessary for understanding the unique nature of the AM. The motion feature pathway consists of a hierarchy of retinotopic maps with a set of feature detector cells duplicated at each location. Receptive field size increases with the level in the hierarchy, thus conforming to one of the fundamental structural aspects of the primate visual system: retinotopic maps of cells with local receptive fields that increase in size with distance from the retina.

Scenarios are *active* memory structures in the sense that they have internally varying activity that provides specific predictions of what will occur next, and when it will occur. A scenario is conceptualized as a list of labeled visual *events* (based on the work in (Rubin, 1986)), and the parameterized time *intervals* between consecutive events. For cyclic actions such as human gaits, the list is circular. A scenario is implemented using two types of connectionist processing units: 1) the event unit, which combines prior information with current evidence for the detection of a particular visual event and 2) the interval unit which uses an adaptive temporal delay and smoothing filter to represent the time between two events in an action.

Figure 2 shows a simple scenario representing the swinging of a pendulum. There are two event units

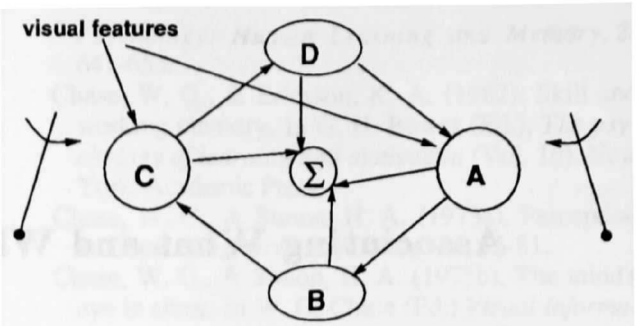

Figure 2: Scenario Network

(circular), which detect, respectively, the clockwise and anti-clockwise reversal of direction of rotation about the pivot, and two interval units (elliptical), arranged in a cyclic network. Suppose there is pendular motion in the input. Unit A detects the onset of clockwise rotation and responds with transient activation. The interval unit B detects this transient, initiates its internal clock, and passes the transient through a temporal delay and smoothing function, producing output at approximately the time the next event is expected to occur. Event unit C receives this *priming* activation at the same time as it detects the onset of anticlockwise rotation, the visual feature for which it codes. It combines the priming and feature activation to produce a new activation transient. This in turn initiates the clock in interval unit D, and the process continues with a wave of activation building up as it flows around the network. Each scenario network also contains a *summator* unit Σ, which monitors the activity of the event and interval units to produce an estimate of the overall activation level in the scenario.

The Association Mechanism

As high-level memory structures, scenarios are position-independent (despaced) representations. In connectionist vision systems, as in biological systems, low level feature detectors are duplicated, each copy having a limited receptive field. This allows parallel processing across the visual field and is an explicit representation of space. For position-independent recognition activation must be integrated across the visual field. Despaced high-level representations of objects and actions perform this integration. The alternative, duplication of complex object and action representations across the visual field, would require copius hardware (units), and the problem of learning new objects and actions would be made even harder.

In terms of the "what" and "where" distinction in the visual system (Mishkin, Ungerleider and Macko,

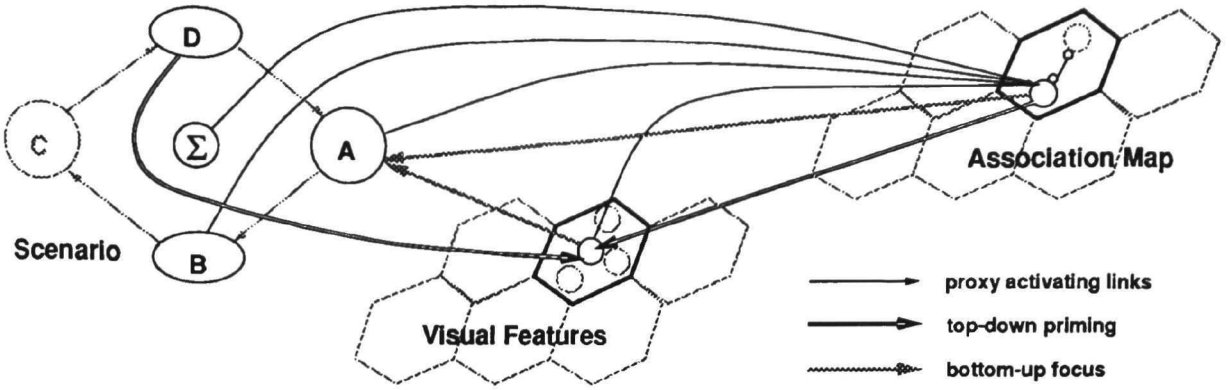

Figure 3: Association Map: Structure and Function

1983), I postulate that scenarios are analyzed primarily in the "what" pathway. Scenarios are an extension of the static concept of "what" to include the dynamic notion "what's happening". Despaced scenario representations are devoid of location information, but location information must be represented somewhere *and associated with the central representations.* In MARS the AM performs this representation and association function, encoding what is happening where. AM information is used to focus bottom-up flow of activation from particular locations in the visual feature maps to particular central scenario representations, and to focus priming activation in the reverse direction. The connection with selective attention is discussed further below, but first I describe the structure and function of the AM.

Structure

The AM is organized as a spatiotopic map encoding location and action. It is based on the understanding that the wave of activation flowing around an active scenario network is a mirror of the changes that are occurring in the scene. Therefore it is possible to use temporal synchrony between the visual changes at a location and the internal changes in the scenario network as the cue for determining an association between the location and the action (see (Goddard, 1988) for an early version of this idea).

At each location in the AM there is a set of *proxy* units (Figure 3 shows two proxies at one location). Each proxy unit represents a scenario. Its activation indicates the degree of belief that the action is occurring at that particular location (activation flowing around the scenario network indicates the belief that the action is occurring *somewhere* in the scene). Proxy dynamics are described below. The proxies at a given location in the spatiotopic map inhibit each other (Figure 3), thereby competing for activation from feature detectors at the location.

Function

Bottom-up Focus: A proxy modulates the significance to its scenario of visual features at its location. Proxies have activation levels in the interval $[-\alpha, 1.0]$, $0 \leq \alpha \leq 1.0$, where polarity indicates indicates evidence for (positive) or against (negative) the action occurring at the location and magnitude indicates the degree of belief, so that the neutral or "resting" level is zero[2]. Positive activation increases the significance to the scenario of the visual features at the location. Negative activation, which occurs through inhibition from other proxies, decreases the significance. α determines the degree of inhibition between proxies, as described below. Modulation is achieved with a link from the proxy to each of the event units in the scenario (shaded link in Figure 3). The link is labeled with the proxy's location, as are the links from the visual feature units to the event unit (shaded link). The activation $P_{S,L}$ from the proxy for scenario S at location L multiplicitavely modulates the activity from the feature units located at L by the factor $(1 + P_{S,L})$ which is in the range $[1 - \alpha, 2]$. In the simulations of gait recognition, α was set at 0.5, so that the modulation factor was always in the interval $[0.5, 2]$.

The modulation factor causes the scenario to "attend" more to locations where its proxy has an activation above resting level and to "neglect" locations

[2]Proxy activations are passed through a scaling function $f(x) = \frac{x+\alpha}{2}$ for transmission to other units, and its inverse upon reception, which ensures that values passed between units are always in the interval $[0, 1]$. The "resting" activation level is then $\frac{\alpha}{2}$. This (de)scaling is ignored here for simplicity.

where its proxy is below resting level. In a scene containing a small number of actions, the effect is that each location where there is action tends to be "owned" by one scenario, and other scenarios actively ignore that location. Simulations showed that this is important in reducing interference between actions that are occurring in different spatial locations and thus increasing the ability to recognize concurrent actions.

Top-down Priming: Recall that the scenario network contains clocked interval units that pass on priming activation to the succeeding event unit. This priming is sent when the visual change which the event codes for is expected to occur. This information is also used to enhance the response of the visual feature units representing the change that is expected, *prior* to the change occurring in the input (hollow link from D in Figure 3). The importance of the AM in this process is that the association that has been set up between an action and a location is used to direct the priming to the region in which the action is occurring (hollow link from proxy in Figure 3). This predictive priming acts as a multiplier on the unprimed response of the feature unit. The multiplicative factor at time t for a feature at location L that is is selected by one or more events in scenario S is

$$1 + \beta_1 \sqrt{\beta_2 P_{S,L}(t) I_S(t) + (1 - \beta_2) I_S(t)}$$

where I_S is the maximum level of priming from interval units in scenario S that predict the feature, β_1 is a parameter controlling the magnitude of the priming effect, and β_2 is a parameter that controls the modulating of priming by the proxy activation. In the simulations of gait recognition β_1 and β_2 were set to 0.25 and 0.7 respectively. The simulations showed that the predictive priming significantly increased the speed with which the correct scenario was activated.

AM Dynamics

A proxy unit has a set of receptive *sites*[3], one for each event its scenario. A site receives input from the event unit and succeeding interval unit in the scenario and from the visual feature units in the feature map that the event unit is selective for (solid links in Figure 3). However, unlike the event unit, the proxy receives visual feature input only from the location it represents. The site compares activity of the event

[3]A unit with sites can be thought of as representing a small network of cells, or a single cell and dendritic tree.

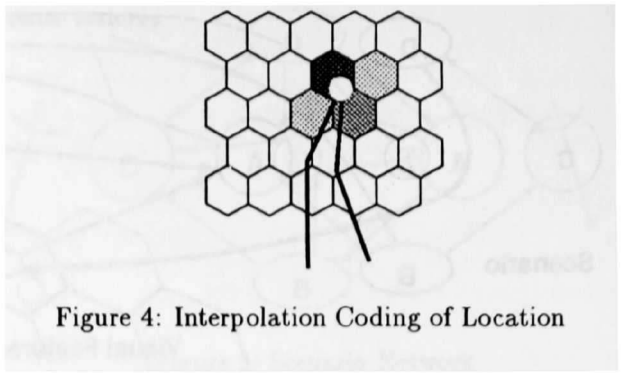

Figure 4: Interpolation Coding of Location

unit with activity of the feature units. It assigns a value which is dependent on simultaneous transients in both event unit and feature units and on the magnitude of those transients. When simultaneous transients are detected, the site value is the geometric mean of the two magnitudes. The site maintains this value during subsequent simulation cycles until the subsequent event has been primed, as indicated by the activation arriving along the link from the interval unit (e.g., unit B in Figure 3), at which time the site value decays to zero. This mechanism allows the proxy to set its activation from the relatively infrequent event transients but for the proxy activation to subside if the predicted events do not occur when expected. The values computed by the sites are combined in a scenario-dependent way to produce the synchrony cue $T(t)$ for the proxy. In the implementation that modeled the gait recognition data (Goddard, 1992) the two highest site values are averaged to produce $T(t)$.

$T(t)$ is combined with an overall estimate $S(t)$ of the scenario activity provided by a link from the scenario summator unit. The proxy activation function is given by:

$$P_{S,L}(t+1) = (1 - \gamma_1) P_{S,L}(t) + \gamma_1 \left[T(t) \left(\gamma_2 + (1 - \gamma_2) S(t) \right) - \alpha \max_{i \neq S} P_{i,L}(t) \right]$$

where γ_1 is a parameter controlling the attack and decay rates of the unit and γ_2 is a parameter controlling the extent to which the scenario activation modulates the synchrony cue. The latter is motivated by the observation that it makes no sense for $P_{S,L}(t)$ (S is happening at L) to be higher than $S(t)$ (S is happening somewhere). The final term in the activation function is the mutual inhibition between proxies at each location, controlled by the parameter α introduced above. In the simulations of gait recognition, γ_1 and γ_2 were set to 0.1 and 0.3 respectively and recall that α was set to 0.5.

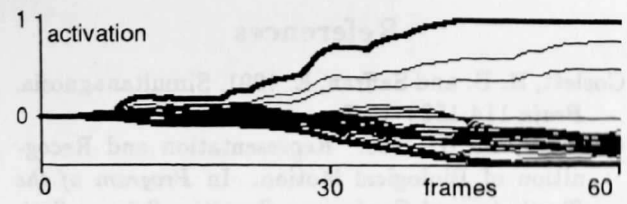

Figure 5: Proxies' Activation

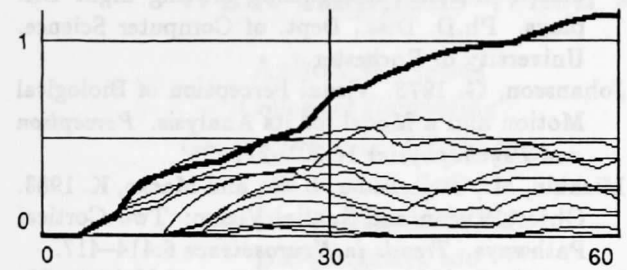

Figure 6: Scenarios' Activation

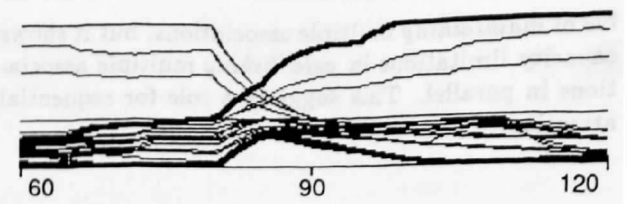

Figure 7: Plasticity

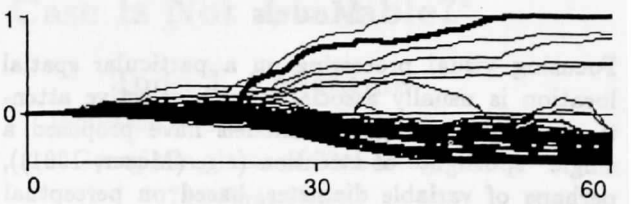

Figure 8: Multiple Actions

Simulations

For the simulations I acquired human gait data from three males and three females walking, running and skipping, using a high resolution imaging system (Scholz, 1989). The data were analyzed to construct scenario network models of the three gaits, modeling the movements of arms and legs as distinct actions. Full details are contained in (Goddard, 1992), here I describe more recent simulations focusing on the AM.

Location Coding in the AM: Simulations in which single actions (e.g., "legs-walking-right") were presented showed that the AM codes location information more finely than the resolution of a single cell. In Figure 4, shading illustrates the activation levels of the proxy units for "legs-walking-right" when the legs were in the location shown (leg actions were modeled to be "located" at the hip). Figure 5 shows a trace of the activation of all the proxies at all locations over time (60 frames/sec simulated). The thick line in Figure 5 indicates the proxy activation for "leg-walking-right" at the heavily shaded location in Figure 4. The next highest activation trace (0.75) is at the moderately shaded location and the third highest (0.5) is at the two lightly-shaded locations. The location of the action can be recovered by interpolating the active locations using their activation as a weighting. Other proxies end up at or below resting level (0).

What-Where Interaction: Figure 6 plots the time-course of activation in the scenarios (summator outputs are shown). Note that by frame 30 in Figures 5 and 6 the leading trace shows high levels of activation and the two continue to rise together. This demonstrates the location binding in the association mechanism ("where") and the activation of scenarios ("what") occuring in parallel. The two processes act cooperatively to settle on a consistent solution.

Plasticity and Phase Insensitivity: A simulation was run to show that the binding in the AM is plastic. 60 frames of "legs-walking-right" were presented, recognized, and an association formed in the AM (Figure 5). Without resetting the network, 60 frames of "legs-running-right" were presented (Figure 7). The previous association dies away after about 30 frames (0.5 sec) of the new action, and the correct new association to "legs-running-right" (thick trace) is formed soon thereafter. The AM is a plastic mechanism.

The scenarios are capable of aligning themselves with the input, independent of initial phase of an action, as described in (Goddard, 1992). The AM receives all its timing expectations from the scenarios and is is therefore also insensitive to phase,

Multiple Actions: Two actions were were presented simultaneously. When the two were presented in approximately the same location, there was usually too much cross talk for the AM to establish any scenario/location association. When the actions were spatially separated, the AM formed the correct association with each location (Figure 8). The thick and thin lines that asymptote at 1.0 are the activation of the proxies of the two actions at the closest AM location. The other pairs at about 0.75 and 0.5 are the corresponding proxies at the other AM locations used in the interpolation-coding. It takes the AM longer to establish the associations when two actions are presented simultaneously. I presented three spatially-separated actions, and the AM took much longer to establish the associations. The AM is capa-

ble of *maintaining* multiple associations, but it shows capacity limitations in *establishing* multiple associations in parallel. This suggests a role for sequential attention.

Selective Attention and Spotlight Models

Focusing visual processing on a particular spatial location is usually associated with selective attention. Previous cognitive models have proposed a single "spotlight" of attention (e.g, (Mozer, 1991)), perhaps of variable diameter, based on perceptual data (e.g., (Posner, Snyder and Davidson, 1980)). The association mechanism outlined here is capable of forming, in parallel, multiple associations between simultaneously-presented spatially-separated actions and their locations. Thus it can be seen as a multiple-spotlight model (see (Shiffrin, 1988) for a review of the data). It would be a relatively simple matter to add inhibition between locations to restrict the model to a single spotlight, as in (Mozer, 1991). However the simulation results suggest another interpretation. A mechanism such as the AM may be used to build up *and maintain* a set of action/location bindings sequentially. As more actions are added to the presentation, it becomes more difficult for the indexing process to reliably activate any scenario model due to crosstalk. The association mechanism cannot focus processing on a particular location until a scenario is at least partially active. If a separate attentional spotlight were added, it would be possible for the AM to make associations between location and action one pair at a time using a sequential spotlight cued by motion or other parameters.

Conclusions

The AM forms a crucial part of MARS, the first program to model the Johansson biological motion data. Modeling arm- and leg-movements separately, I found that the AM was required to enable recognition of full-body human gait. Using as a cue temporal synchrony between scene-action and internal active memory structures representing actions, it associates *what's happening* with *where* it is happening. The AM displays an ability to maintain multiple associations in parallel but cannot necessarily form those associations in parallel, suggesting a role complementary to that of sequential attention.

References

Coslett, H. B. and Saffran, E. 1991. Simultanagnosia. *Brain* 114:1523–1545.

Goddard, N. H. 1988. Representation and Recognition of Biological Motion. In *Program of the Tenth Annual Conference Cognitive Science Society*, 230–236. Hillsdale, NJ: Lawrence Erlbaum.

Goddard, N. H. 1992. The Perception of Articulated Motion: Recognizing Moving Light Displays. Ph.D. Diss., Dept. of Computer Science, University of Rochester.

Johansson, G. 1973. Visual Perception of Biological Motion and a Model for its Analysis. *Perception and Psychophysics* 14:201–211.

Mishkin, M., Ungerleider, L. G., and Macko, K. 1983. Object Vision and Spatial Vision: Two Cortical Pathways. *Trends in Neuroscience* 6:414–417.

Mozer, M. C. 1991. *The Perception of Multiple Objects: a Connectionist Approach.* Cambridge, MA: MIT Press.

Olson, T. J. 1989. An Architectural Model of Visual Motion Understanding. Ph.D. Diss., University of Rochester.

Posner, M. I., Snyder, C. R. R., and Davidson, B. J. 1980. Attention and the Detection of Signals. *Journal of Experimental Psychology: General* 109:160–174.

Rashid, R. F. 1980. Lights: A System for the Interpretation of Moving Light Displays. Ph.D. Diss., University of Rochester.

Rubin, J. 1986. Categories of Visual Motion. Ph.D. Diss., Department of Psychology, Massachussetts Institute of Technology.

Scholz, J. P. 1989. Reliability and Validity of the WATSMART Three-Dimensional Optoelectronic Motion Analysis System. *Physical Therapy* 69(8):679–689.

Shiffrin, R. M. 1988. Attention. In Atkinson, R. C.*et al.*(Eds.), *Steven's Handbook of Experimental Psychology, Volume 2: Learning and Cognition*, 739–811. New York: John Wiley & Sons.