Simulating Biological Motion Perception Using a Recurrent Neural Network

Roxanne L. Canosa

Department of Computer Science
Rochester Institute of Technology
Rochester , NY 14623
rlc@cs.rit.edu

Abstract

People have the ability to perceive biological motion under conditions of severely limited visual information. If the information is in the form of a point-light motion sequence of a human walker or jogger, perceptual discontinuities are barely noticeable. This phenomenon can be simulated for a machine perceptual system by using a recurrent artificial neural network. A feedback connection from the output of the hidden layer to the input of the hidden layer provides the network with information about past events that can be used to classify an entire sequence of events. In this case, the discrete events are the x,y-coordinates of the point-light displays during a specific motion sequence. Generalizations about temporal as well as spatial patterns can be made that enable the network to classify an input sequence as being either biological or non-biological in nature.

Introduction

Biological motion perception refers to the phenomenon that people are able to recognize biological motion from severely impoverished stimuli consisting of only a few points of light attached to a human walker's major joints and head. The perception one has is not of a sequence of disjoint and randomly moving points of light, as might be expected, but of a coherent sequence of plausible human motion. The perception of motion in this case relies upon the integration of spatial cues over time, and not just on the static display of the light points presented one after the other in succession. Thus, time plays an important role in human cognition.

Johansson (1973) was the first to show that severely impoverished stimuli in the form of points of light attached to the major joints and head of a human walker are sufficient for the recognition of biological human motion. The time course of the perceptual response was found to be less than one second. The explanation for the rapid response to such a restricted set of stimuli rests upon the recognition that the organization of the dot pattern follows strict mathematical laws of relative coherence among the dots. The ease of detectability is a result of our intuitive understanding of these laws.

In a more recent study, Neri, Morrone, and Burr (1998) found that biological motion detection requires the simultaneous analysis of the motion of more than one joint. Detection depends upon virtual links that exist between the joints. Increasing the number of illuminated joints causes a rapid increase in the sensitivity to biological motion, i.e., people can more readily detect biological motion when more points are presented. This sensitivity increases at a rate much greater than the rate of sensitivity to other types of non-biological rigid motion detection, suggesting that biological motion perception may represent a fundamental property of cortical function and an important evolutionary objective.

Pavlova et al. (2001) tested children, three to five years of age, and adults for sensitivity to biological motion from point-light displays. The displays consisted of either people walking on a treadmill, dogs walking or running, or cartoon birds moving. Static displays were also used in the study for comparison purposes. The recognition of biological motion was near perfect for the motion sequence, whereas recognition for the static displays was near chance. This suggests that the dynamic nature of presentation is a critical component in the detection process.

Vaina et al. (2001) used whole-brain functional MRI (fMRI) scans of individuals engaged in the detection and recognition of biological motion from point-light displays. The goal was to discover which parts of the brain are most active when an individual is presented with such a display. The method they used involved comparing the brain activity of subjects while they were engaged in a task that required them to distinguish between a walker and a non-walker from a point-light display. Brain activity involved both the dorsal and the ventral extrastriate areas, and the activity was distributed across both of these regions.

Grossman et al. (2000) also used fMRI to isolate the cortical areas associated with the perception of biological motion. Activation was located in the STS region, as well as the MT/MST region and the kinetic-occipital (KO) region. More activity was found in the right hemisphere

than in the left hemisphere, suggesting that spatial analysis may play a role in the detection process.

Giese and Poggio (1999) and Giese (2000) created an artificial neural network that is a plausible model of brain functioning during the recognition of biological motion. The model simulates processing along the ventral and dorsal pathways, which allows the simultaneous processing of both form and motion information. A dynamic recurrent network was used that is hierarchically organized into three layers each for the form (ventral) and motion (dorsal) pathways. Lateral recurrent connections between the two pathways allowed for the temporal association of information.

Model Description

The purpose of this study is to create a simulation of human detection and classification of point-light displays of motion patterns for a machine perception system. A recurrent neural network based on the Elman network was designed and used for this study. The purpose of the network is to decide whether or not a given recurrent input sequence corresponds to biological (human) motion.

An Elman network is a two-layer back-propagation network that has a recurrent connection from the output of the hidden layer to the input of the hidden layer. Neurons with tangent-sigmoidal activation functions are used in the hidden (recurrent) layer, and neurons with linear activation functions are used in the output layer. This configuration allows the network to model and detect non-linear as well as linear functions of time. The recurrent connection in the first layer allows for a delay, which provides the network with the capability of storing values from the previous time step and combining those values with the current input for prediction purposes.

Elman (1990) has suggested that time should be represented in a network by the effect that it has on processing. The goal of an Elman network is to enhance the ability of the network to process time-varying inputs by giving the network dynamic properties. In effect, the network gains the ability to remember past events, and uses that memory to predict future events. It is the recurrent connections in the network that allow the system to remember past events, and enables future behavior to be shaped by previous responses.

This type of network is ideal for the detection of biological motion from impoverished stimuli, because the detection is highly dependent upon the time-varying nature of the stimuli, as well as the spatial relationship between the various input points.

Method

Data Collection

Eight sequences of biological motion were captured using a SONY Digital Handycam® camcorder. All of the sequences were recorded in the NTSC video format at 30 frames per second.

The first four sequences were each of a (different) person walking on a treadmill at a rate of 3 miles per hour. Sequences 1 and 3 were of a (different) female person walking, and Sequences 2 and 4 were of a (different) male person walking. The camcorder was held stationary on a tripod and recorded a side-view of the person facing straight ahead while they walked. The direction of walking was to the left in the camera frame of reference. The second four sequences are sequences of non-walking motion tasks. Sequence 5 is of a person skipping rope, Sequence 6 is of a person performing jumping jacks, Sequence 7 is of a person jogging in place, and Sequence 8 is of a person swinging on an outdoor swing. From the eight sequences, there were a total of five different human motion tasks.

After the video sequences were recorded, a ten second segment (300 frames) was extracted from each sequence. For each sequence, the entire image set of 300 images was sub-sampled (every second and third image was deleted) to lower the temporal resolution to ten images per second. This was done to improve the processing speed for each sequence, and does not affect the perception of smooth motion (Palmer, 1999). Ten points of light were manually selected for each image to coincide with the major joints and head of a person in motion. Figure 1 shows a point light display taken from each of the five motion tasks, and an example of a non-biological motion sequence.

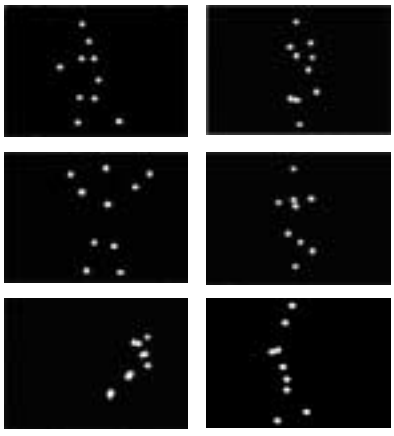

Figure 1. Point-light displays for five human motion tasks, and non-biological motion. Top row: walking (left), skiping rope (right). Middle row: jumping jacks (left), jogging (right). Bottom row: swinging on a swing (left), non-biological (right).

The points were selected in the same order for all of the images in a single sequence. For example, a given order might be: head, left shoulder, left elbow, left hand, right elbow, right hand, left knee, left foot, right foot, and left foot. As the points were selected, matrices of the x- and y-coordinates were created. Each row of the matrix is a vector of the x- or y-coordinates of the points selected from a single frame. Each column of the matrix is a vector containing the x- or y-coordinate of the same joint over an entire motion sequence. Since 10 points were selected, and each sequence consisted of 98 frames, the size of each matrix is 10 x 98. The two matrices, taken together, provide the information that is used for the input to the neural network.

Network Design

A two-layer recurrent network, also known as an Elman network, was created for the motion perception. The first layer (not including the input layer) is the hidden layer and has ten neurons, one for each joint point. The hidden layer is also the recurrent layer, and uses a tangent-sigmoidal activation function. The second layer is the output layer and has a single neuron with a linear activation function. The weights and biases were initialized using the Nguyen-Widrow layer initialization method (Nguyen and Widrow, 1990).

The weights and biases were updated using back-propagation, along with a gradient descent learning function that used momentum and an adaptive learning rate. Both the training function and the adaptation used an approximation of the gradient, not the actual gradient, to determine the weight and bias updates. An approximation is used because the error from the recurrent connections is ignored when calculating the gradient, which allows the network to learn more efficiently. This has the disadvantage of making the network less reliable than it would be if there were no recurrent connections, however the disadvantage can be overcome if more neurons are used in the hidden layer (Demuth and Beale, 2000).

Network Training

The input to the network is a sequence of numbers that represents the x- and y-coordinates of the joint points from a motion task. The x- and y-coordinates are scaled to range from 0 to 1 before they are input into the network. It was found empirically that an input range from 0 to 1 allows the network to learn most quickly.

After scaling, the Euclidean distance is found from each point of light to the upper left corner of the display, which provides a reference coordinate. The distances taken together provide a frame-of-reference for the purpose of binding together all points from a single sequence into a coherent display of motion. The binding of the coordinates used in conjunction with the spatial relationship between the joint points describes coherent biological motion. This provides the information used by the recurrence relation and defines the network behavior.

The network was trained on two examples of valid biological motion (Sequences 1 and 2) as well as on two examples of non-biological motion. The non-biological sequences were created by randomly permuting the x- and y-coordinates of the joint points. That is, the x-coordinates were shuffled and the y-coordinates were shuffled, and then they were randomly matched together. For example, the x-coordinate of the head point might be matched to the y-coordinate of the knee joint. This procedure destroys the binding of the coordinates and thus the coherence of the biological motion pattern, yet contains the same absolute information as the valid sequence. The perception one has when viewing such a sequence is that the motion is not biological in nature, despite a distinct perception of some other type of coherent motion.

The network was trained on two examples of the walking sequence and two examples of the non-biological (randomized) sequences. The non-biological sequences used for training were created from the walking sequences used for training.

The network trained in a reasonable amount of time (70 to 85 seconds) and the summed squared error was stabilized to approximately 0.1 after 110 training epochs. Figure 2 shows the summed squared error of the network as training progressed.

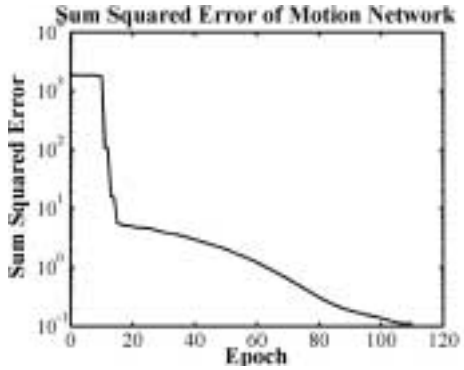

Figure 2. The summed squared error during training of the network designed to perceive biological motion. SSE stabilizes to 0.1 after 110 training epochs.

Results

Testing the Network

The network was first tested on the training data to ensure that a specific input would give the correct classification. Sequences 1 and 2 (both walking) were used as the valid training data and a randomized version of Sequences 1 and 2 were used as the invalid (non-biological) training data. The network was trained to output +1 if the input data is

biological in nature and –1 if the input data is non-biological in nature. During the training stage, the network was sequentially presented with the four training sets, alternating a valid sequence with a non-valid sequence. Each sequence was input into the network sequentially, beginning with the first frame of the motion sequence, and ending with the 98th frame of the sequence. Figure 3 shows the output of the network after testing the network on the trained data. Time steps 1 – 98 and 196 – 293 correspond to a valid motion sequence, and time steps 99 – 195 and 294 – 391 correspond to the invalid, randomized sequences. As expected, the training data is correctly classified.

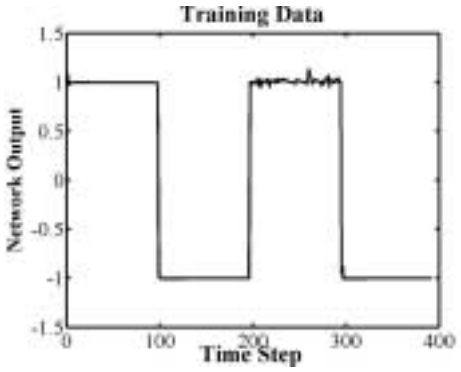

Figure 3. Testing the network on the training data.

Sequences 3 and 4 (also walking) and the randomized versions of these sequences were used to test the network on newly presented data. Figure 4 shows the output when the network was presented with these sequences. The network correctly classifies these sequences as either valid biological walking motion or invalid, non-biological walking motion.

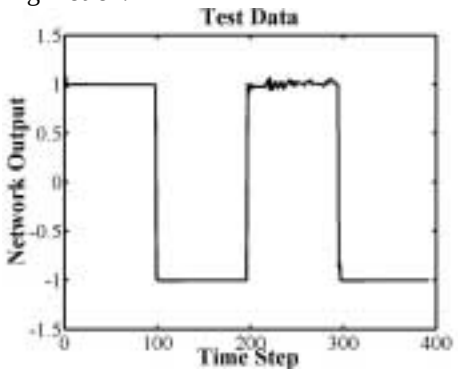

Figure 4. Testing the network on unknown data.

Generalization of the Network

It is apparent from figures 3 and 4 that the network has correctly learned the temporal and spatial patterns of the joint movements associated with a person walking on a treadmill at a speed of 3 miles per hour. It is possible that the network has over-learned this one specific task and will not generalize well to other kinds of human motion tasks. Therefore, the second set of non-walking motion sequences was input to the network to test the generalization capabilities. These sequences consist of a person jogging, a person skipping rope, a person performing jumping jacks, and a person swinging on a swing. Randomized versions of these motion sequences were not used as input to the network for either training or testing.

The results are shown in figure 5, and suggest that the network has learned some generalization of the training data. For the jogging sequence the motion was classified as being definitely biological in origin. This is likely because the movement of joints during running is similar to the movement of joints during walking. For the skipping rope sequence, the classification is also biological. For the jumping jacks sequence the classification is not entirely biological, and the network seems to have moments of "doubt" during portions of the sequence when the arms are in the upward position. Despite this, and despite the fact that the viewpoint is frontal rather than sideways as in the walking sequence, the network still classifies the motion as being mostly biological in nature. Finally, in the swinging sequence, the network alternates between classifying the motion as being biological when the person on the swing is in the downward position, and not biological when the person is swinging in the air. This is probably because the network was trained on data where the center of gravity of the joint points was close to the center of the frame. During swinging, the center of gravity moves upward and to the far left or far right of the frame. This is a positive generalization feature of the network because human biological motion does not usually include flying in the air.

Conclusions

A recurrent neural network in the form of an Elman network is able to correctly identify and classify human biological motion that is similar to walking. An application of this type of network for the solution of a real-world problem would be to use the network in conjunction with a video camera to monitor a geographic area for the presence of people walking by, where there may be other kinds of motion in the background. For example, it may be desirable to distinguish between a person walking by and a truck or bicycle crossing the path. A surveillance camera used with this type of neural network could successfully determine the presence or absence of people walking by when other types of motion are present.

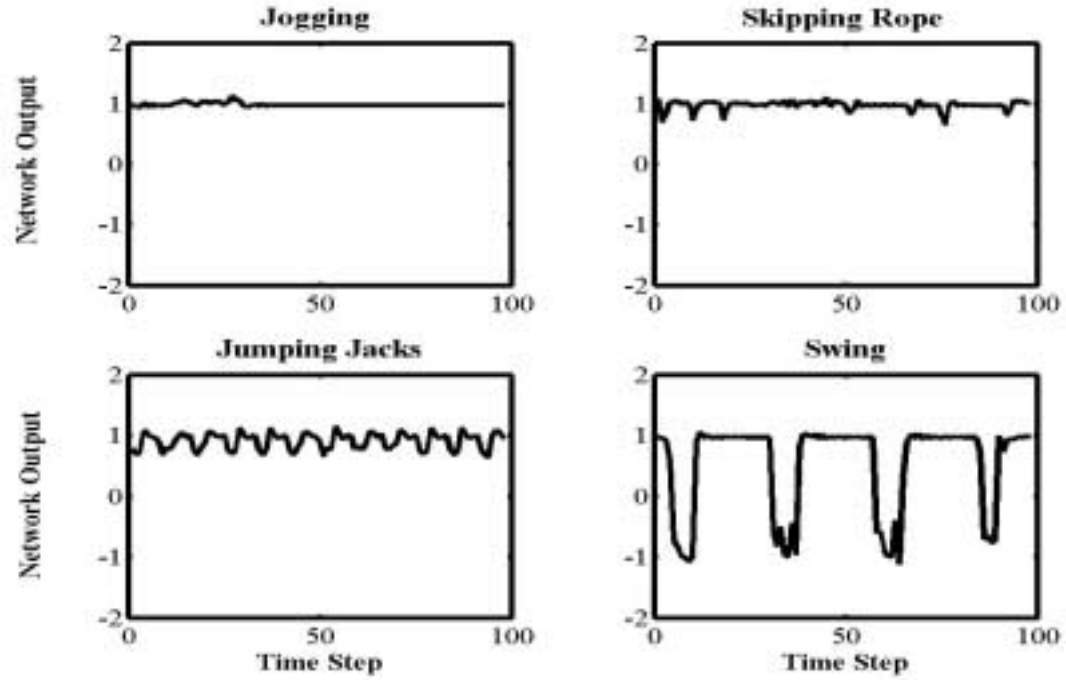

Figure 5. Testing the generalization of the network on other human motion sequences.

Additional work on the design could include an algorithm to automatically extract the joint-points from the video images, and a means of determining the fewest number of points that can correctly classify the input. Other areas of investigation could include presenting the joint-points in a field of other randomly moving points of light to determine if the network is able to segment the biological motion from the random motion. Also, changing the direction of the motion may affect the network generalization capabilities, and occlusions might result in occasionally missing points, effecting the classification. These issues need to be considered.

Other types of biological motion, such as a cat running or a bird flying, could be included in the model. Another interesting application would be to combine a recurrent neural network with a Bayesian network to determine certain properties of the motion sequence, based on evidence gathered from the temporal and spatial relationship between the joint-points. For example, using this type of network it may be possible to correctly guess the weight of a box being lifted, solely from the motion of the joints of the lifter.

It may be advantageous to incorporate top-down influences, such as an attentional mechanism, into the design. This would reduce the number of joint-points to only those that are necessary for a specific action-perception task. This constrains the amount of information that the network requires to properly classify the motion. Fewer, yet more relevant, input coordinates may enable the network to learn a broader class of motion inputs, and therefore improve generalization capabilities.

References

Demuth, H., and Beale, M. 2000. *Neural Network Toolbox User's Guide*, Natick, MA: Mathworks, Inc.

Elman, J.L. 1990. Finding structure in time. *Cognitive Science* 14:179-211.

Giese, M.A. 2000. Neural model for the recognition of biological motion. From G. Baratoff and H. Neumann, eds. *Dynamische Perzeption* Infix Verlag, Berlin, 105-110.

Giese, M.A., and Poggio, T. 1999. Synthesis and recognition of biological motion patterns based on linear superposition of prototypical motion sequences. Paper for the IEEE Workshop on Multi-View Modeling and Analysis of Visual Scene, Fort Collins, CO.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., and Blake, R. 2000. Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience* 12(5):711-720.

Johansson, G. 1973. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics* 14:201-211.

Neri, P., Morrone, C., and Burr, D.C. 1998. Seeing biological motion. *Nature* 395:894-896.

Nguyen, D, and Widrow, B. 1990. Improving the learning speed of two-layer neural networks by choosing initial values of the adaptive weights. *International Joint Conference on Neural Networks*, San Diego, CA, III:21-26.

Palmer, S. 1999. *Vision Science: Photons to Phenomenology*, Cambridge, MA: MIT Press.

Pavlova, M., Krägeloh-Mann, I., Sokolov, A., and Birbaumer, N. 2001. Recognition of point-light biological motion displays by young children. *Perception* 30:925-933.

Vaina, L.M., Solomon, J., Chowdhury, S., Sinha, P., and Belliveau, J.W. 2001. Functional neuroanatomy of biological motion perception in humans. *Proceedings of the National Academy of Sciences* 98(20):11656-11661.