ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 95-100

Biologically-Inspired Human Motion Detection

Vijay Laxmi, J. N. Carter and R. I. Damper
Image, Speech and Intelligent Systems (ISIS) Research Group
Department of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, UK
Email: {vl99r|jnc|rid}@ecs.soton.ac.uk

Abstract.
A model of motion detection is described, inspired by the capability of humans to recognise biological motion even from minimal information systems such as moving light displays. The model, a feed-forward backpropagation neural network, uses labelled joint data, analogous to light points in such displays. In preliminary work, the model achieves 100% person classification on a set of 4 artificial subjects and another of 4 real subjects. Subsequently, 100% motion detection is achieved on a set of 21 subjects. In the latter case, the correspondence problem is also solved by the model, since the network is not 'told' which joint is which. Like human beings, the neural networks perform both tasks within a small fraction of the gait cycle.

1 Introduction

Human beings can routinely recognise biological motion, i.e., the motion of living things. This capability is not much affected by viewing distance or poor visibility conditions. Even at large distances or in poor quality video recordings, humans not only perceive motion but also the kind of motion, e.g., jumping, dancing, running or walking. We also use this information to identify people. Any familiarity cue as to clothes, appearance or hair style is obliterated at large distances. This suggests that it is the motion itself that is responsible for the identification. Psychologists have attempted to determine the processes underlying this perception mechanism by using moving light displays (MLDs) [2]. Such displays are obtained by filming moving subjects wearing reflective pads/light bulbs on their body joints. Filming is in almost dark conditions so that the display consists only of a configuration of light points with no information on the subjects' shape or identity. In spite of the paucity of information, human observers easily perceive not only motion but also the kind of motion—walking, running, dancing in couples, cycling, etc. Even with presentation time as small as 0.2 s, the human brain is able to organise the dynamic configuration of light points so as to detect a human figure walking. However, with MLDs of shorter duration, many observers fail to perceive the motion. So it seems that the human brain requires certain minimal temporal information for this perception to happen.

The observation that humans can perceive motion even from MLDs forms the basis of work presented in this paper. Here, we seek to investigate possible processes whereby the human brain might carry out motion recognition and associated person identification tasks. To this end, we propose a cognitive model based on artificial neural networks (ANNs). The ANN under consideration makes use of labelled joint positions in a video sequence, analogous to point lights in a moving light display. First, we investigate the applicability of such a model to human identification using gait as a basis. We view this as a simpler problem than motion detection as the classification task is closed-set and the model need not solve any correspondence problem. That is, the coordinate data for any particular joint are always fed to the same input node, so that the network is implicitly informed of which joint is which.

Encouraged by the success of this preliminary work, we have subsequently attempted to build a cognitive model of human motion detection. Ideally, this model should behave as humans do in the context of biological motion and our understanding of the human perceptual system. A good discussion on conventional motion recognition techniques can be found in [1].

The remainder of this paper is structured as follows. Section 2 highlights capabilities and limitations of the human perception. In Section 3, we demonstrate the applicability of a neural network for gait classification, i.e., recognising one subject from a closed set by gait alone. Section 4 presents a simple neural model for biological motion recognition. Conclusions and future work appear in Section 5.

2 Human Perception of Biological Motion

Most studies of human motion perception use moving light displays (MLDs). A moving light display contains only information about specific points of an object undergoing motion. In the case of moving humans, the specific points are generally body joints. Any change in motion perception due to transformation in position of the points may lead to an understanding of the perceptual mechanism itself. Johansson [2] was the first to apply MLDs to human perception of biological motion. His studies indicate that the human brain can perceive as well as categorise motion from MLDs. Although MLDs are minimal information systems, the perception of biological motion remains as vivid as a full video sequence if the display is dynamic. Another interesting aspect is that subjects can recognise human motion in approximately 0.2 s (about five frames of Johansson's displays and a small fraction of the gait cycle). This detection is relatively independent of the direction of motion with respect to the camera.

Subsequent studies [4] indicate that the human perceptual system can detect biological motion in the presence of static and dynamic masks, which consist of additional light points not attached to the walker. Spatial and phase scrambling of joints apparently have no appreciable effect on results. The perception of motion, however, breaks down when these MLDs are displayed upside-down [6].

This means that a cognitive model of human detection based on how humans perceive biological motion should display noise immunity and the ability to recognise motion from short data sequences. One way to develop such a model would be to base it on a thorough understanding of the perceptual mechanism. As yet, no satisfactory account for the capabilities of human perception has been put forward. Even attempts to determine a subset of joints crucial to the effect have been unsuccessful, as deletion of any particular joint seems to have no adverse effect. Although the perceptual system is sensitive to many factors, such as the dynamic symmetry along the limbs, the position of limbs with respect to the torso, and phase relations between limbs, significant perturbations to any of these is not enough to disorganise the perception. Hence, we can say that no single factor or group of factors crucial to the perception process has yet been isolated.

3 Human Gait Classification

In this section, an ANN-based model is applied to human gait classification. We start with this problem on the grounds that we believe it is easier than the motion-detection problem. The model works on joint data, manually labelled by author VL, analogous to point lights in an MLD. We assume that the scene contains only one person walking fronto-parallel to the camera and the ordering of joint positions in all frames is the same. This implies that the model need not solve the so-called correspondence problem (i.e., determining which joint is which). A further assumption is the availability of the coordinates of all joints under consideration implying no necessity to handle occluded or missing data.

For each frame in a sequence, the positions of seven joints—shoulder, elbow, wrist, hip, knee, ankle, toe—and a time tag representative of the frame's position in the gait cycle were recorded. Features presented to the ANN were derived by grouping frame data, consisting of 15 parameters—time tag plus $\langle x, y \rangle$ coordinates of the 7 joints, through a 'context' window moving over the entire image sequence. This is the classical way of allowing a simple feed-forward network (without feedback) to handle sequence data [5]. For an image sequence of N frames, the number of features is $N - F + 1$, each feature being a $15F$ tuple. So, for a window size $F = 5$ for example, the first data point (feature) was generated by combining frames 1 to 5, the second data point by combining frames 2 to 6, the third data point by combining frames 3 to 7 and so on. For practical reasons, the data were normalised to a unity hypercube and subject to mean removal before being fed to the ANN.

The addition of the time tag rendered the model dependent on the natural order of walking. To prevent it being influenced by variations other than those in walking patterns, the number of frames per walking sequence was kept the same for each experiment while each walker entered the first frame at the same phase in the gait cycle. Only one leg and one arm were considered as human gait is bilaterally symmetrical. The model was tested on walking sequences of 4 synthetic and 4 manually-labelled human subjects. For the synthetic walkers, the angles of rotation of leg joints (hip, knee and ankle) were derived from the mean angles of rotation from [3]. Using the standard deviation in [3] as a guideline, the angle patterns for different synthetic walkers were generated. To account for small variations in an individual's walk, sequences were generated by varying angles of rotation by a small amount.

The number of input units to the network depended on the size, F, of the sliding

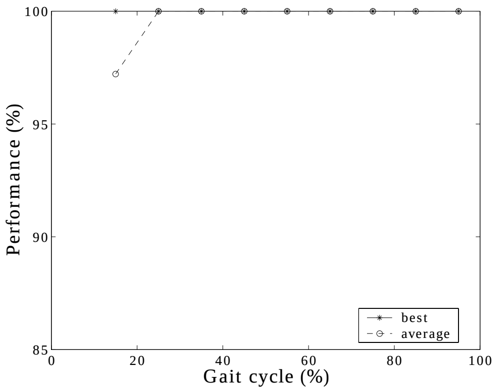

Figure 1: Identification performance of an ANN classifier on four synthetic walkers as a function of window size, F, expressed as a percentage of the gait cycle.

window, i.e., the number of frames per data point. As only 7 joints were used, the number of input units was $15F$. The number of hidden units was 16 in all cases. The number of output units was 4 as each subject was allocated a binary code with only one bit set to 1. (Note that as F increases, we go from a situation in which there is little context information but a relatively large number of data samples to the opposite situation where there is considerable context but few data samples.) For both synthetic and manually-labelled data, backpropagation training used one walking sequence per subject, and testing used unseen walking sequences.

Figure 1 shows the results obtained for gait classification of four synthetic subjects and Figure 2 displays the results of four manually-labelled real subjects. Both the best and the average results over five repetitions of training with different initialisation points are shown. The neural model achieves 100% classification at about 35% of the gait cycle. For manually-labelled data, the performance degrades slightly when 'context' window approaches a complete gait cycle. The reason for this is unknown; it has not been investigated further. A possible reason could be decrease in number of data points as F increases. These results demonstrates that, in the restricted circumstances described, the ANN is a suitable classifier. It is, therefore, a suitable candidate for a cognitive model for human motion detection.

4 Towards a Cognitive Model for Motion Detection

Perception studies, outlined above, indicate that although human beings perceive biological motion from fronto-parallel MLD sequences, or a 'normal' view, of an upright walker, an upside-down presentation of the same sequence is not recognised as biological motion. This suggests that a cognitive model for human motion detection should also have this property. We propose that such a model, an ANN, can be trained to discriminate between the normal view and a non-normal view. The proposed ANN

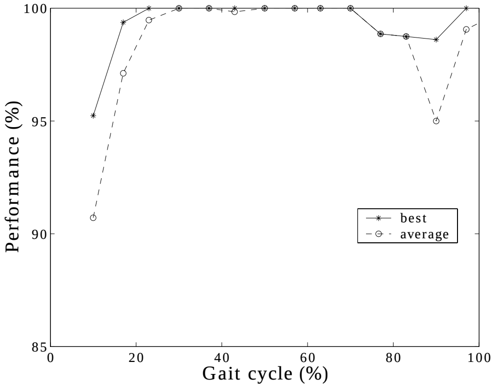

Figure 2: Identification performance of an ANN classifier on four manually-labelled walkers as a function of window size, F, expressed as a percentage of the gait cycle.

is similar to that described in Section 3 but with some differences as now detailed.

The human data are from a high-resolution motion capture system, utilising infrared markers on each joint. The coordinates of 15 different joints—shoulder, elbow, wrist, hip, knee, ankle, toe (for both right and left limbs) and the head—were recorded with a video system as the subject walked. The three-dimensional coordinates were projected onto a suitable plane thus simulating a camera. No manual labelling was performed, rather the projected joints coordinates were first sorted in ascending order of y coordinates. If the y coordinates were same, further sorting was done on x coordinates. This method is equivalent to scanning the joints row-by-row from left to right. As joints may appear in a different order in each frame using such line-scan technique, the ANN now needs to solve the correspondence problem. No time tag was added to each frame data as the model needs to detect the presence of human motion irrespective of how and when a person enters the scene. As the model should be able to detect human motion irrespective of how fast a person walks, no resampling is done. Otherwise the data were presented to the ANN exactly as before.

The ANN was trained using the normal view as a positive instance. In addition to upside-down, a top view has also been taken as a negative instance. (Informally, the authors could not perceive this as a biological motion.) Training was done with data obtained from one of the subjects selected at random and testing was done with normal and non-normal views of 20 other subjects. Although all these views can be derived from one another by appropriate geometric transformations, the human perceptual perceives only the normal view as ‘biological motion’ and so does our cognitive model with 100% accuracy. The results for the ANN motion detector are shown in Table 4. The first column of the table is the window size used for grouping frames. The second column depicts this in terms of fraction of average training gait cycle. Columns 3 to 7 indicate the performance of the cognitive model for five experiments with different initialisation points of the ANN. As the model performs well even for small values of F, i.e., small fractions of the gait cycle, testing for larger values has not been carried

Window size (F)	Average %age of gait cycle	Run 1 (%)	Run 2 (%)	Run 3 (%)	Run 4 (%)	Run 5 (%)
5	6.9	99.95	99.95	99.95	99.95	100.0
7	9.7	100.0	100.0	100.0	100.0	100.0
9	12.5	100.0	100.0	100.0	100.0	100.0
11	15.3	100.0	100.0	100.0	100.0	100.0

Table 1: Performance of ANN motion detector on human motion data using infra-red emitters as body joint labels.

out. The model successfully learned to discriminate between normal and upside-down motion, generalising from one example of one subject to 20 different subjects. The ability to make correct classification within a small fraction ($\sim$10%) of the gait cycle is in accordance with human behaviour.

5 Conclusions and Future Work

A neural-network-based cognitive model of human motion recognition and classification has been demonstrated to reproduce some of the main results of human perceptual experimentation. The model needs to be developed and tested for robustness to noise and incomplete data, and under the conditions where human perception succeeds and fails. The structure of the resultant ANN may provide useful insight into the mechanism of human perception and may also be useful as the basis for human identification and classification on real video sequences (e.g., in security applications).

References

[1] C. Cédras and M. Shah. Motion-based recognition: A survey. *Image and Vision Computing*, 13(2):129–155, 1995.

[2] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14(2):201–211, 1973.

[3] M. P. Murray. Gait as a total pattern of movement. *American Journal of Physical Medicine*, 46(1):290–329, 1967.

[4] J. Pinto and M. Shiffrar. Subconfigurations of the human form in the perception of biological motion displays. *Acta Psychologica*, 102(2–3):293–318, 1999.

[5] T. J. Sejnowski and C. R. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1(1):145–168, 1987.

[6] S. Sumi. Upside-down presentation of the Johansson moving light-spot pattern. *Perception*, 13(3):283–286, 1984.