DETECTING PARTIAL OCCLUSION OF HUMANS USING SNAKES AND NEURAL NETWORKS

Ken Tabb, Neil Davey, Stella George & Rod Adams
e-mail: {K.J.Tabb, N.Davey, S.J.George, R.G.Adams}@herts.ac.uk
Department of Computer Science, Faculty of Engineering & Information Sciences,
University of Hertfordshire, England.

Abstract: This paper summarises the development of a computer system designed to detect moving humans in an image or series of images. The system combines the use of active contour models, 'snakes', which detect human objects in an image, with a 2 layer feedforward backpropagation neural network, to categorise the detected shape as human, or not. It was found that combining the neural network's output values with its confidence value provided a means of classifying unseen shapes into 'human' and 'non-human'. Moreover the confidence value can provide a measure of the degree of occlusion of a detected human.

Keywords: Occlusion, Pedestrian, Snake, Active Contour, Neural Network

I INTRODUCTION

This paper presents a technique for detecting human shapes in images, and for determining whether or not those human shapes are being partially occluded. An active contour [1], or 'snake', is used to detect and track objects in an image or sequence of images. When the snake has relaxed onto an object, that is, when its energy has been minimised, the contour's vector of (x,y) coordinates is re-represented using a novel encoding algorithm called the axis crossover representation. The resulting scale- and location-invariant vector can be used as an input pattern for neural networks.

A feedforward neural network has been found to classify 90% of unseen human shapes correctly, when trained with both human and non-human crossover vectors. Furthermore, when the network is presented with unseen, partially occluded, human shapes, a measure of occlusion can be obtained by analysing the neural network's output values. Results of these experiments are presented in section IV.

Whilst other techniques for pedestrian detection and tracking exist [2, 3], they are unable to detect or accommodate partial occlusion in the target human shape. The method presented in this paper is able to indicate a level of occlusion for a given pose and, as such, could be used as a 'higher level cognitive process' to control snakes, as suggested in the original active contour paper [1].

II DETECTING HUMAN SHAPES WITH SNAKES

A snake is an energy minimising spline whose energy function can be tailored to detect specific features in images, allowing for the detection of particular classes of object. Since their original design [1], several variants have been developed for specific scenarios or for computational efficiency [4, 5, 6]. Despite these improvements, snakes have no knowledge of what they are detecting or tracking, and thus cannot categorise their own shape. This makes the snake less robust in complex visual environments, where it is often unknown what objects may enter and divert the snake's attention away from the target outline and onto other edges or noise. Without a mechanism for identifying what is being tracked, snakes have limited appeal to artificial intelligence applications.

A snake is initialised around the target human shape in an image by the user (Figure 1). The snake is

then iteratively mapped onto the human outline in that frame by repeatedly minimising its energy function, resulting in a human-shaped contour. Once relaxed, the snake can be moved into the next frame of the movie and mapped onto the human's new position using its relaxed position from the previous frame as a starting point.

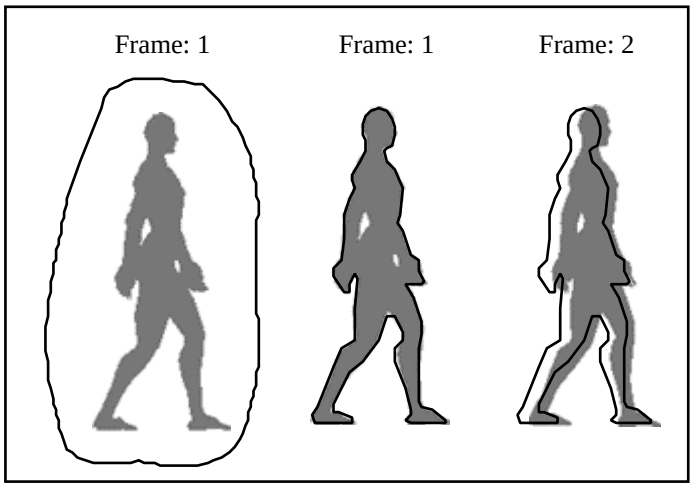

Figure 1: A snake relaxing on a human shape. [Left] The user initialises a contour around the target human. [Middle] Minimising the snake's energy function forces the snake to relax onto the human outline. [Right] Once relaxed, the snake is initialised in the next frame using its relaxed position as a starting point. Its energy is then minimised again to relax it onto the human's new position.

We used a modification of the Fast Snake model [4], which proved more suitable than the original model for detecting human shapes without user guidance. A summary of these features follows, whilst a more exhaustive list can be found in [7]. Fast snakes do not require the user to provide corrective guidance to the snake; indeed no external energy is used at all. Fast snakes space their control points equally along the snake, without explicitly expanding or contracting the snake. The original model's 'shrink-wrapping' [4] can continue to shrink a snake even when it is already situated on the target contour, pulling it back off the target object. Fast snakes allow corners to form at certain points by providing 'personalised' energy functions for control points. In the original model, the user-defined parameters are identical for all control points, resulting in a tendency to cut the corners off the target contour, as they are aiming to achieve global smoothness. To move a control point, fast snakes determine whether the control point would have lower energy if it were located elsewhere within its neighbourhood. By considering every location in the immediate neighbourhood, rather than jumping the control point from one location in the image to another fast snake overcome the original model's limitation in so much as strong local image features cannot be overlooked.

III THE AXIS CROSSOVER REPRESENTATION

Snakes are stored as a vector of (x,y) coordinates, from which a spline is constructed. This native representation is neither scale- nor location-invariant, so that similarly shaped contours may not have similar vectors. A representation of the snakes is presented which, in addition to being both scale- and location-invariant, can be customised so that it encapsulates salient features of the object class being detected.

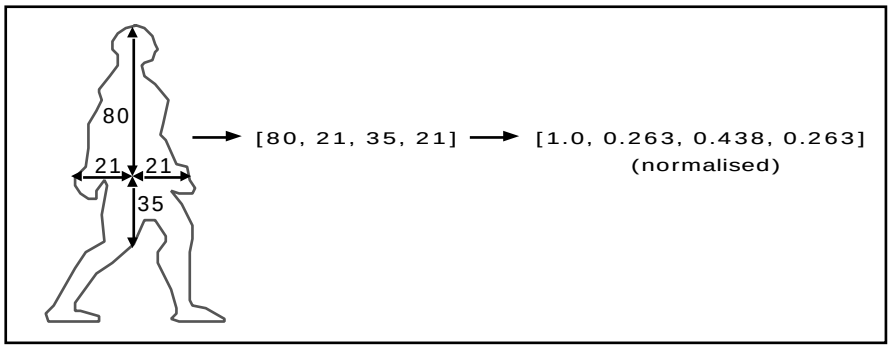

Figure 2: Encoding a human contour as a 4-axis crossover representation. With equi-spaced axes, the distance between the centre of a snake and its edge is measured at regular angles, in this case 0°, 90°, 180° and 270°, and stored in a vector. The vector is then normalised, making it scale-invariant.

To obtain an axis crossover representation of a snake, the centre of the snake is calculated, using the mean of the snake's control point (x,y) coordinates. Axes are projected from the snake's centrepoint to its edges at specified angles. For example an equi-spaced 4-axis representation would grow axes at 0°, 90°, 180° and 270° within the snake (Figure 2). In this study, all axes are equi-spaced for simplicity, although axes could in theory be projected at irregularly-spaced angles. In tasks where only certain parts of an object need to be inspected by the neural network, for example production line assembly, projecting axes at particular angles during encoding may generate a more compact representation for the object class being detected (Figure 3). Once the axes have been projected, the distance along each axis, from its origin at the snake's centre to the snake's edge, is stored in a vector. The resulting vector, whose length equals the number of axes being used in the representation, is then normalised. This normalised vector can then be used as a training or test pattern to the neural network.

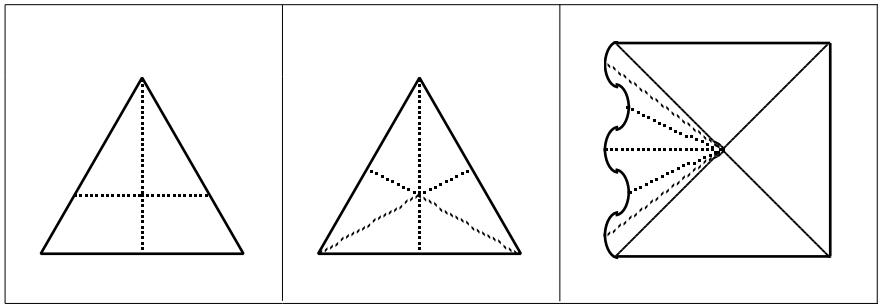

Figure 3: The axis crossover representation can be customised for encoding particular object classes. [Left and Middle] Encoding contours using 6 axes may generate a more robust representation of an object's triangular properties than if only 4 axes are used. [Right] Axes can be projected at irregularly spaced angles if necessary, allowing areas of interest to be encoded in detail whilst omitting irrelevant areas from the representation.

The number of axes used can be varied so that differing degrees of detail can be captured. However, using more axes in a representation results in a more complex categorisation task for the neural network.

IV NEURAL NETWORK EXPERIMENTS

Categorisation

It was necessary to determine whether or not a neural network could distinguish one group of crossover vectors (humans) from other groups of crossover vectors (non-humans). The vectors were evaluated with simple hidden layer backpropagation networks, as the task, at this stage at least, was a categorisation of the vectors.

The axis crossover representation allows for different numbers of axes to be used in the contour representation. Having fewer axes simplifies the neural network's task, however enough axes need to be used for the human qualities of the contours to be encapsulated in the vectors, so that they can be distinguished from the non-human vectors. It was decided to test several different numbers of axes used in the representations, which in turn meant testing several different neural networks, each with as many input units as there were axes in the representations. It was hoped that these experiments would identify the optimal number of axes to use in representing the particular object class relevant to this project.

Double output unit networks were used so that their categorisation confidence values, based upon the two output units' values, could later be analysed. The networks were trained with a range of different hidden layers, to allow the network with the optimal generalisation skills to be identified. The training set contained 150 human and 150 non-human shapes. The non-human shapes were of outdoor objects, for example cars, streetlights and traffic lights, as the project involves outdoor scenes. In all experiments, network output was allowed some lenience; an output of 0 - 0.2 was classed '0', and an output of 0.8 - 1 was classed '1'. Training was stopped when the network had reached 15% or less error. It was then tested with 10 unseen human and 10 unseen non-human shapes.

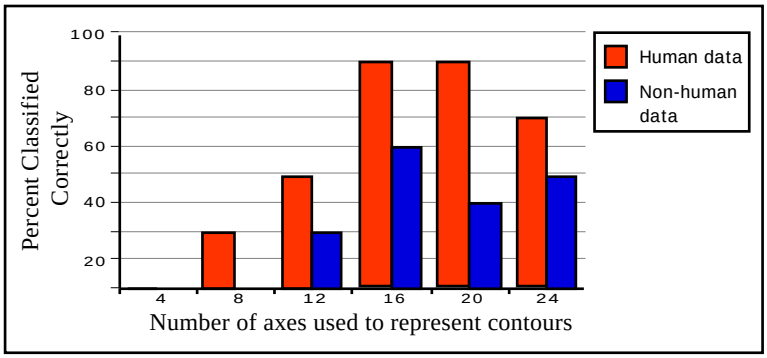

Figure 4: Results from experiments on double output unit neural networks trained using human and non-human shapes. Each score represents the averaged score of ten identical neural networks, each trained with a different initial weight matrix. Experiments were run using 4-, 8-, 12-, 16-, 20- and 24-axis crossover representations. Classifications were said to be correct if the output values were between 0.8 - 1.0 and 0.0 - 0.2 respectively for a desired output of '1 0', and vice versa.

Figure 4 shows the average results of the various double output unit networks when tested with the 10 unseen human and 10 unseen non-human shapes; each network had been trained and tested 10 times using different initial weight matrices.

It can be seen from the graph that networks trained on 16-axis vectors were most accurate at classification, with 90% of unseen human shapes and 60% of unseen non-human shapes being categorised correctly.

At this stage in the experimentation, the axis crossover representation had been shown to be sufficiently descriptive of human shapes; the 16-axis representation was adopted as a standard, as it had outperformed the other configurations of the representation, and was judged to be most able to encapsulate the human qualities of a given contour. Neural networks with 16 input units were therefore used henceforth.

Confidence values

Since the network has two output units, a confidence value of its classification can be obtained by differencing the two values. The average confidence value when classifying a human shape correctly is 0.81 (the difference between the average values of output unit 1 (0.97) and output unit 2 (0.16)), with 1.0 representing complete confidence that the shape is human, whereas the average confidence value when classifying a non-human shape correctly is -0.65, with -1.0 representing complete confidence that the shape is non-human. A confidence value of 0 represents the least confident classification. These values reflect the trend seen in Figure 4, that is, the network is more accurate at classifying human shapes than non-human shapes. In particular, the average output values when given non-human shapes only just fall within the 'classified correctly' zones of 0 - 0.2 and 0.8 - 1.

Partial Occlusion

Having identified that the axis crossover representation was a suitable means of encoding snakes for neural networks, it was interesting to test how robustly the networks behaved when presented with partially occluded human shapes. The partially occluded human shapes were generated by removing percentages of complete human shapes, so that the human could be envisaged as walking behind an invisible screen; all occlusion took place on the leading edge of the human. The network was not re-trained with partial human shapes, training only included complete human and complete non-human shapes. In order to identify humans as partially occluded, the network needed to classify them as non-human, or at least not human.

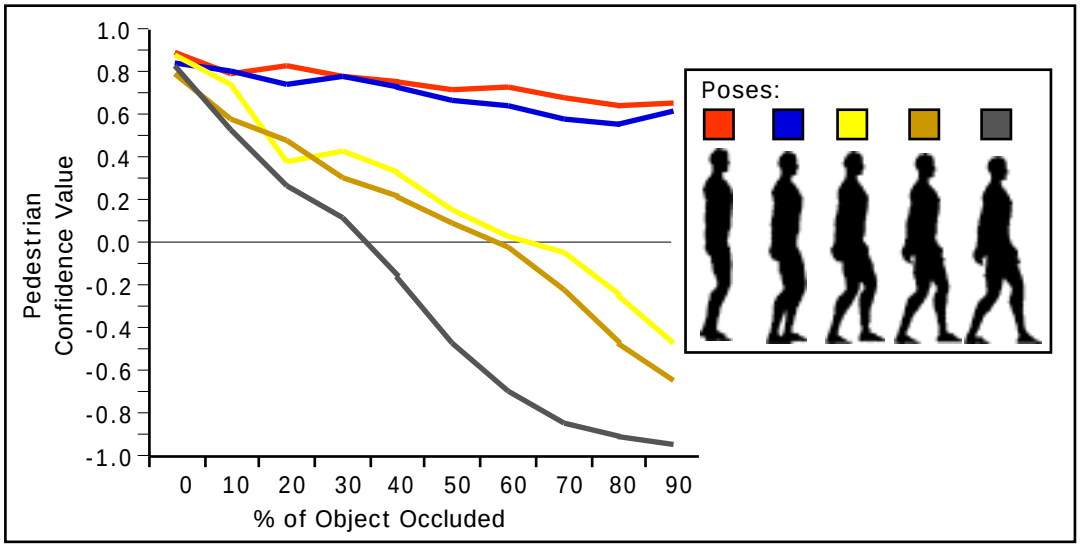

Figure 5: Confidence values for partially occluded human shapes. Scores shown are the averaged scores of ten identical networks, each of which had been trained with different initial weight matrices. The training set did not contain any partially occluded human shapes.

Figure 5 shows the results of the networks when presented with partially occluded human shapes, of

varying degrees of occlusion. Five different unseen poses were used, with 10 degrees of occlusion on each. When each of the five poses were 0% occluded, they were all categorised as ‘human’ with a high confidence value. The confidence of the network decreased with occlusion for each pose, although was found to decline at different rates for different poses. The network is able to handle reasonably high levels of occlusion in upright poses, but only low levels of occlusion in outstretched poses, before it is uncertain of its classification.

V CONCLUSION

This paper describes a technique for representing human shapes so that a feedforward neural network can be trained to categorise them as either ‘human’ or ‘non-human’. The axis crossover representation forms a scale- and location-invariant representation of the shape. The representation can be customised in terms of the number of axes used, allowing more detailed representations to be encoded. Furthermore the axes can be projected from the centre of the contour at user-defined angles instead of being equi-spaced around the contour, which may allow more task-specific representations to be generated.

Sixteen equi-spaced axes were found to encapsulate sufficiently the human qualities of a given shape. Networks which had been trained with these representations were particularly successful at correctly categorising unseen human shapes, with an average score of 90%. When presented with unseen partially occluded human shapes, the network’s output values were able to indicate the level of occlusion of a given human pose. The networks were found to be more resilient with upright human poses than with outstretched human poses.

We are currently investigating the possible applications of this system to track moving humans in an environment where the humans may become temporarily partially occluded.

REFERENCES

[1] M. Kass, A. Witkin and D. Terzopoulos. Snakes: Active Contour Models. International Journal of Computer Vision (1988) 321-331

[2] J. J. Little and J. E. Boyd. Recognizing people by their gait: The shape of motion. Videre: Journal of Computer Vision Research Vol 1 No 2 (MIT Press 1988) 2-32

[3] A. M. Baumberg and D. C. Hogg. An efficient method for contour tracking using active shape models. University of Leeds School of Computer Studies Research Report Series, Report 94.11 (1994)

[4] D. J. Williams and M. Shah. A fast algorithm for active contours and curvature estimation. CVGIP - Image Understanding 55 (1992), 14-26

[5] T. F. Cootes and C. J. Taylor. Active shape models - ‘Smart snakes’. British Machine Vision Conference Sept 1992, 276-285

[6] R. Curwen and A. Blake. Dynamic contours: Real-time active splines. In A. Blake and A. Yuille (Eds). Active Vision. (MIT Press 1992) 39-58

[7] K. Tabb and S. George. Snakes and their influence on visual processing. University of Hertfordshire Department of Computer Science Technical Report No 309 (Feb 1998)