Modelling Pedestrians Using Artificial Neural Networks

Harsh Nanda & Larry Davis
Department of Computer Science
University of Maryland
College Park, MD, 20742

Abstract

There has been an increasing interest in pedestrian detection in the last decade to save the numerous deaths caused by accidents in which pedestrians are involved. Vehicle manufacturers are addressing these challenges by looking into extendable vehicle body structures , to be activated on first impact with a pedestrian. A complementary approach is to focus on sensor based solutions, which enable vehicles to "look ahead" and detect pedestrians in their surroundings.

Sensor based approaches require a model of pedestrians to validate the measurement against. Modelling pedestrians is especially hard because of the wide range of possible pedestrian appearances.

In this paper we have reported some preliminary experiments to demonstrate the feasibility and strength of artificial neural nets to learn and detect pedestrians. Our neural net based model achieves up to 97% classification accuracy.

1 Introduction

Humans have very flexible, not so well defined, complex shapes. They can appear in all colors and in varying contexts in front of different backgrounds. This makes the task of human modelling and detection in single images quite hard. One general approach involves shifting windows of various sizes over the image, extracting low level features, and then using pattern classification techniques to determine the presence of a pedestrian. For example [2] uses wavelet features in combination with Support Vector Machine (SVM) classifier. The system described in [4] uses contour features in a hierarchical template matching approach to efficiently "lock" onto candidate solutions using template matching. A powerful technique to establish regions of interest (ROIs) is stereo vision. It is used in combination with neural networks based classifier [3] or texture based pattern classifier [18].

Neural networks are known for their ability to express highly nonlinear decision surfaces, which makes them appropriate for classifying objects with high degree of shape variability such as humans. The driving hypothesis for our desire to use a Neural Net for this problem is that: *a trained neural net with single hidden layer with N input nodes will be able to learn the variations in shapes of the pedestrians and will then be able to successfully classify inputs as pedestrians and non-pedestrians*.

In this paper we present experimental results to support the above mentioned hypothesis. The outline of this paper is as follows. In Section 2 we discuss work done in the field of pedestrian detection and classification, Section 3 describes our neural net design. In Section 4 we discuss data collection and training of the neural net, Section 5 describes the test process. In Section 6 we discuss the performance of our system and Section 7 provides conclusion and future directions.

2 Related Work

A significant amount of progress has been made in the area of pedestrian modelling for detection from moving platforms in the past few years. Most of the vision-based pedestrian detection systems have taken a general learning-based approach, where the human appearance is described in terms of simple low-level features from a region of interest.

Most of the human tracking and motion analysis systems employ simple segmentation procedure such as background subtraction or temporal differencing to get the foreground region. Other than applications such as surveillance, where the camera is stationary, these techniques of extracting the foreground are not applicable. Some techniques such as Pfinder [5], W4 [6] and path clustering [7], have been developed to compensate for small, or gradual changes in the scene. Independent motion detection techniques can help [8, 9], but they are difficult to develop and are not feasible for non-rigid object extraction since different body parts move differently. In all these approaches the assumption is that all detected objects are pedestrians. This limits the generalization and application of these techniques.

More sophisticated pedestrian detection techniques have a two-step process: foreground detection followed by recognition step to verify if the target object is a pedestrian or not. The recognition step can be motion-based, shape-based or multi-cue based. Motion based approaches use periodicity of human walk or learned gait for pedestrian detection [10, 11, 12, 13, 14]. These approaches use temporal information for pedestrian detection and the procedure requires a sequence of frames, which delays the identification until several frames later and increases the processing time. Also such methods cannot detect pedestrians standing or doing something that does not contain the assumed periodic pattern.

Shape based approaches try to solve the harder problem of recognizing pedestrians in single images, hence taking care of both moving and stationary pedestrians. The biggest challenge that this problem offers is to model the huge amount of variations in the shapes, pose, size and appearance of humans and their backgrounds. [15, 16] use handcrafted human models for pedestrian detection. The main restriction of this approach is that it requires segmentation into body parts which itself is a very hard task. Lipton [17] uses an easy to calculate metric perimeter^2/area to classify human and vehicle. The metric is rather fragile to many cases where group of people are walking together.

Another line of approach involves shifting windows of various sizes over the image at different resolutions, extracting low-level features, and using standard pattern classification technique to determine the presence of a pedestrian. [2] extract wavelet features and then use SVM to classify them. [4] extracts edges and then uses chamfer distance measure to compare with an hierarchy of templates of human shapes.

A powerful technique to establish regions of interest is stereovision. It is used in [18, 3] in combination with texture-based pattern classification. [2] uses stereo vision, but prefer to combine it with a verification technique based on symmetry properties.

3 Experimental Design

An overview diagram of the method is shown in Figure 2. In the preprocessing stage, an input image sequence is first processed to segment the object from the background and track it in each frame (if it is moving). The obtained sequence of blobs are then properly aligned and scaled to a uniform height. These blobs of pedestrians are then divided into independent sets of training and test data. The negative training data is generated synthetically. the system is trained on the training data and then its performance measured on the test data.

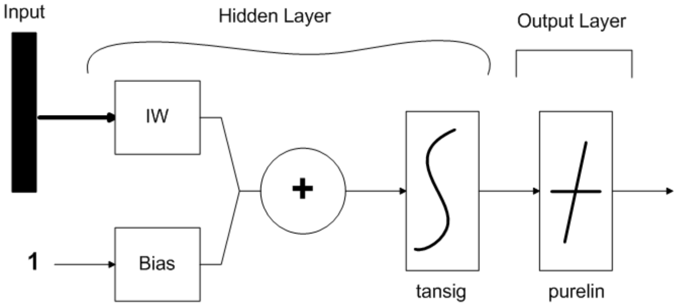

Figure 1: Neural Net Structure

3.1 Neural Net Design

3.1.1 Network Structure

Based on the fact that: *Neural networks with a bias, a sigmoid layer, and a liner output layer are capable of approximating any function with a finite number of discontinuities* [1]. We designed a network with one hidden layer of sigmoid neurons, followed by an output layer of a linear neuron. The input of the neural net are the pixels of the rectangle bounding the blobs of interest. The output of the neural net varies from -1 to 1.

3.1.2 Number of Hidden Nodes

The number of hidden nodes is a key parameter in structure of a neural network. Since there is no robust analytic method to find the number of hidden nodes, we resorted to experimental methodology. We train the neural net for different number of hidden nodes ranging from 5 to 38 and observe the classification accuracies obtained. The classification accuracy increases with the increase in the number of hidden nodes, as expected. The higher the number of hidden nodes, the more expressive is the neural net, and hence the higher is the classification rate. However, beyond a certain number, which turns out to be 19-21 (Table 1) for an input size of 1500 nodes (50x30 image), the classification accuracy saturates. As the number of hidden nodes increases further, the learning process becomes slower and it takes longer to learn. Thus using 20 hidden nodes gives us a neural net structure that is expressive enough for our problem domain of pedestrian modelling and takes the least learning time.

# Hidden Nodes	5	11	14	17	20	23	26
Error +ve Training Data	0	0	0	0	0	0	0
Error -ve Training Data	6	0	0	0	0	0	0
Error of Test Data	0	27	33	7	0	11	14
# Epochs	$\propto$	243	185	163	289	244	388

Table 1: Performance v/s hidden nodes

NOTE: Training method used for the above experiment is Scaled Conjugate Gradient

3.1.3 Training Method

Another crucial design parameter in the neural net is the training method to be used. There is no training method that works best for all domains. Also because of the size of the input vector and since we are using batch mode training, some methods have impractically high memory requirements. Such methods were not even considered for evaluation. The methods evaluated and corresponding observations are given below:

- Batch Gradient Descent: Converges in nearly 10000 Epochs (error rate similar to Scaled Conjugate Gradient)

- Batch Gradient Descent with Momentum: Converges nearly in 5000 (error rate similar to Scaled Conjugate Gradient)

- Variable Learning Rate: Does not converge

- Resilient Back-propagation: Does not converge

- Scaled Conjugate Gradient: Converges in less than 500 Epochs

- One Step Secant Algorithm: Converges in nearly 500 epochs but error rates much higher as compared to Scaled Conjugate Gradient

Thus, Scaled Conjugate Gradient gives us the best error rates and minimum learning time, hence is the method of choice.

4 Training and Test Data

A set of images containing pedestrians (positive data) and not containing pedestrians (negative data) were collected/generated for training and testing the neural network.

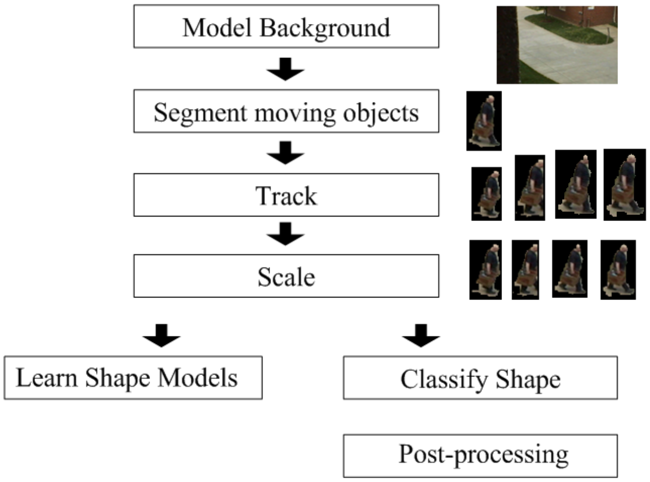

Figure 2: Three Stage Process of Classification

4.1 Positive Training Data

The training and test data for our problem domain was collected using a Sony Digital camera. People entering the building and exiting the building were captured over a day from morning till evening. Different people were captured with different zoom parameters depending on where in the scene were they captured. The orientation of the camera with respect to the pedestrians varied depending on the location of the target pedestrian on the ground. The data collected was not controlled at all and the pedestrians who were captured did not know about the setup at the time of taping. Figure 3 shows example shots of the captured data. The data collected was then processed as described in the Section 4.1.1 to get the positive training and test data.

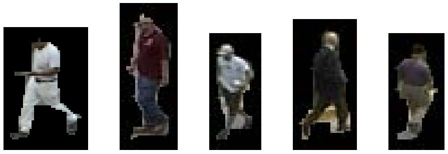

Figure 3: Example of Negative Training Data (Pedestrians)

4.1.1 Preprocessing

Foreground detection was achieved via background modelling and subtraction. We use the non-parametric background modelling technique that is essentially a generalization of the mixed-Gaussian background modelling approach[20], and is well suited for outdoor scenes in which the background is often not perfectly static (for e.g. occasional movement of tree leaves and grass). A number of standard morphological cleaning operations are applied to the detected blobs to correct for random noise. Frame-to-frame tracking of a moving object is done via simple overlap of its blob bounding boxes in the current and previous frames. These bounding boxes of the blobs representing the pedestrians are extracted. This data is normalized to the range -1 to 1. Without this normalization, with 20 hidden nodes, and using Scaled Conjugate Gradient method of Error Back Propagation (EBP), the system converges to the required performance level for our problem domain in more than 10000 epochs as compared to 200 epochs which it takes after normalization.

4.2 Negative Training Data

The negative training and test data was generated synthetically. We used an image editing software to manually create 600 images of random blobs. These images were then flipped, mirrored and both to make the total number of negative examples to be 2400. Examples of the negative data i.e. non- pedestrians are shown in Figure 3.

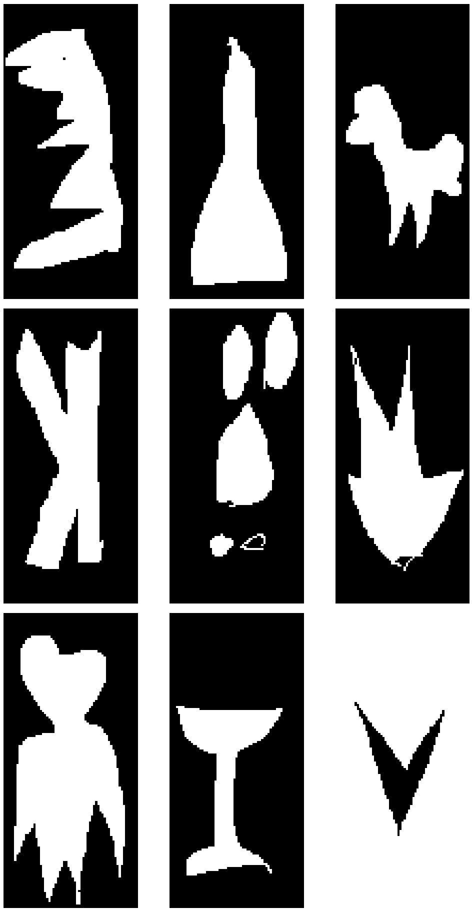

Figure 3: Example of Negative Training Data (Non-Pedestrians)

5 Classification

Once we have identified the neural net structure and trained it using the preprocessed data, we can design a three-stage system to evaluate the performance of our learned developed system (Figure 2). Stage one consists of extracting blobs of interest from the scene, tracking and resizing them. Stage two consists of feeding these blobs into the neural system which then classifies them as pedestrians or non-pedestrians. The system for each frame independently outputs a value in the range of -1 to 1. Finally, in stage three we threshold the output value to decide if the input was a pedestrian or not.

6 Results

We did foreground extraction for a total of 15 different pedestrian videos with 300 frames each (total 4500 positive examples). We then performed leave-one-out on the set of these 15 videos. For each pedestrian video, we created a neural net using all positive videos except that video and leaving out randomly chosen 300 negative frames as the training data and then tested the system on the video and the 300 negative frames which had been left out. 97% of the positive test data i.e. pedestrians and 100% of the negative test data i.e. non-pedestrians was correctly classified. All the pedestrian frames that were misclassified belonged to 2 videos, one of which was of a person who was carrying a bag-pack and another one was wearing a hat. As our dataset consisted of only 15 videos, when we left these videos out, none of the training videos contained anything similar. Thus the training data was not representative enough of the test data in those cases.

7 Summary and Future Work

More experimentation needs to be done with a much larger data set to verify and benchmark the performance of the system. There are no false alarms as the negative training data was randomly (manually generated). Experiments need to be performed where the negative training data is also collected from real scenes.

In this paper we have described a robust neural net based technique for modelling the variability in pedestrian shapes. The success of this approach lies in the fact that neural nets are able to learn non-parametric functions and hence can capture the variations in human shape. This provides us an effective method to model pedestrian shapes. This verification module combined with a detection module can robustly do pedestrian detection.

References

[1] Tom Mitchell. ”Machine Learning”. McGraw Hill, 1997.

[2] C. Papageorgiou, T. Evgeniou, T. Poggio. A trainable pedestrian detection system. IEEE Int. Conf. on Intelligent Vehicles, pp. 241-246, Germany, Oct 1998.

[3] L. Zhao, C. Thorpe. Stereo- and neural network based pedestrian detection. In Proceedings ITSC, Tokyo, Japan, 1999.

[4] D. M. Gavrila, V. Philomin. Real-Time Object Detection for ”Smart” Vehicles, ICCV, Greece, 1999

[5] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland. Pfinder: Real-time Tracking of the Human Body, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 780-785, July 1997.

[6] I. Haritaoglu, D. Harwood, L. Davis. W4-Real Time Detection and Tracking of People and their Parts. Technical Report, University opf Maryland, Aug. 1997.

[7] J. Segen, S. Pingali. A camera-Based System for Tracking People in Real Time. Proc. of the 13th Int. Conf. on Pattern Recognition, pp. 63-67, 1996.

[8] P. J. Burt, J. R. Bergen, et al. Object Tracking with a Moving Camera: An Application of Dynamic Motion Analysis. Proc. of IEEE Workshop on Visual Motion, pp. 2-12, 1989.

[9] R. Polana, R. Nelson. Low level recognition of human motion. Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects, pg. 77-82, Austin, 1994.

[10] R. Cutler, L. Davis. Real-time periodic motion detection, analysis and applications. Proc. of IEEE Conference on Computer and Pattern Recognition, pg. 326-331, Fort Collins, USA, 1999.

[11] C. Wohler, J. K. Aulanf, T. Portner, U. Franke. A Time Delay Neural Netowrk Algorithm for Real-time Pedestrian Recognition. International Conference on Intelligent Vehicle, Germany 1998.

[12] H. Mori, N. M. Charkari, T. Matsushita. On Line Vehicle and Pedestrian Detection Based on Sign Pattern. IEEE Trans. on Industrial Electronics, Vol. 41, No. 4, pp. 384-391, Aug, 1994.

[13] A. A. Niyogi, E. H. Adelson. Analysing Gait with Spatiotemporal Surfaces. IEEE Workshop on Motion of Non-Rigid and Articulated Objects. pp. 64-69, Austin, 1994.

[14] S. A. Niyogi, E. H. Adelson. Analysing and Recognizing Walking Figures in xyt. IEEE Conference on Computer Vision and Pattern Recognition, pp. 469-474, 1994.

[15] D. Hogg. Model-based Vision: a Program to See a Walking Person. Imae and Vision computing, Vol. 1, No. 1, pp. 5-20, 1983.

[16] K. Rohr. Towards Model-Based Recognition of Human Movement in Image Sequences. CVGIP: Image Understanding, Vol. 59, No. 1, pp. 94-115, Jan, 1994.

[17] J. Lipton, H. Fujioshi, R. S. Patil. Moving Target Classification and Tracking from Real-Time Video. Workshop on Applications of Computer Vision, Princeton, NJ, Oct. 1998.

[18] U. Franke, D. M. Gavrila, et al. Autonomous Driving goes downtown. IEEE Intelligent Systems, 13(6):40-48, 1998.

[19] T. Tsuji, H. Hattori, N. Nagaoka, M. Watanabe. Development of night vision system. Proc. IEEE International Conference on Intelligent Vehicles, pg: 133-140, Tokyo, Japan, 2001.

[20] A. Elgammal, D. Harwood, L. Davis. Non-parametric model for background subtraction, ECCV, 2000.