A Neural Network Approach to Pedestrian Detection

VICTOR-EMIL NEAGOE, CRISTIAN-TUDOR TUDORAN, AND MIHAI NEGHINA
Depart. Electronics, Telecommunications & Information Technology
Polytechnic University of Bucharest
Splaiul Independentei No. 313, Sector 6, Bucharest
ROMANIA
victoremil@gmail.com

Abstract: - The paper presents an original approach for pedestrian detection using the neural network classifier called Concurrent Self-Organizing Maps (CSOM), previously introduced by first author; it represents a winner-takes-all collection of neural modules. The algorithm has the following stages: (a) feature selection using one of the three candidate techniques Histogram of Oriented Gradients (HOG)/1D Haar transform/2D Haar transform; (b) classification using a CSOM classifier with two concurrent neural modules, where first module is trained with pedestrian images and the second one is trained with non-pedestrian images. We present the experimental results obtained by computer simulation of our model. For training and testing the neural classifier, we have used INRIA Person Dataset. One obtains the best Total Success Rate (TSR) of 99.7 %.

Key-Words: - pedestrian detection, neural network classifier, Concurrent Self-Organizing Maps, Histogram of Oriented Gradients, Haar transform.

1 Introduction
Pedestrian detection is about saving lives in traffic. Much has been done to improve internal safety for the passengers in the car-airbags, deformation zones, etc. However, for pedestrians, little work has actually been implemented when it comes to safety. Pedestrian detection is the act of automatically detecting humans in a traffic scene. It has its major application onboard a vehicle aiding the driver with perception. Pedestrian detection systems (PDS) are in most cases meant to be included in a driver assistance framework and they are connected with the term of “smart car” and “intelligent vehicle” that can contain anything that has to do with making the car “aware” of, and “reacting” to the environment. Detecting humans in images is a challenging task owing to their variable appearance and the wide range of poses that they can adopt. There exists an extensive literature on pedestrian detection [5], [6], and [9]. From the feature point of view, some of the techniques that have been investigated are: PCA descriptors, 1D and 2D Haar wavelets, Histogram of Oriented gradients (HOG). In terms of classifiers, a few of the techniques that have been considered are the following: the popular Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), feedforward neural networks.

In this paper we propose and evaluate an algorithm for pedestrian detection using a neural network classifier based on the system of Concurrent Self-Organizing Maps (CSOM) [10], [11]. For feature selection, we have compared the techniques of Histogram of Oriented Gradients (HOG), 1D Haar Transform (1D HT), and 2D Haar Transform (2D HT).

2 Algorithm Description
The proposed pedestrian detection algorithm has the following steps (see Fig. 1):
(a) feature selection using one of the three candidate techniques: (a1) Histogram of Oriented Gradients (HOG); (a2) 1D Haar Transform (1D HT); (a3) 2D Haar Transform (2D HT);
(b) classification using a Concurrent Self-Organizing Maps (CSOM) classifier with two concurrent neural modules, where first module is trained with pedestrian images and the second one is trained with non-pedestrian images.

In our study, we took a neural approach for pedestrian detection task, which is for classifying block-images to one of two distinct classes: pedestrian or non-pedestrian.
2.1 Feature Selection
A performant pedestrian detection requires a robust feature selection technique that allows the human shape to be cleanly discriminated, even in cluttered backgrounds under difficult illumination.

2.1.1 1-D Haar Transform
The 1D Haar Transform (1D HT) is applied to the image vector, whose space dimension is equal to the number of its pixels. For example, by taking a picture with 128 x 64 pixels, one obtains a vector with 8192 components. As a result of the trade-off between energy (information) preservation and the computational complexity, one retains a reduced number of components in the 1D HT space. For the considered example, we can retain 2700 1D HT descriptors to preserve 90% of the signal energy.

2.1.2 2-D Haar Transform
2-D Haar Transform (2D HT) is applied on the matrix representation of the input image. In this manner, the selected features can describe not only the vertical gradients (as it was the case of the 1D HT), but any direction gradient in the image.

2.1.3 Histogram of Oriented Gradients
The feature selection method based on Histogram of Oriented Gradients (HOG) recently introduced by Dalal et al. [3] consists of the following steps:
- Computation of the gradient magnitudes and orientations in each point of the input image.
- Division of the input image in segments. For example, we can consider cells of 8 x 8 pixels. An input image of 128 x 64 pixels can be divided into 128 cells.
- Computation of the Histogram of Gradient Orientations (HOG) for each cell. For example, one can choose 9 orientation bins. The "weighted votes" of gradients corresponding to these bins are proportional to gradient magnitudes. After this step, each cell can be expressed using these 9 coefficients that make the histogram of oriented gradients. An input image of 128 x 64 pixels can be represented by this method by a vector of 1152 = 128 x 9 features.
- Normalization of HOG features by one of several variants; the chosen strategy significantly impacts system performances:
  a) Global normalization involving the division of the feature vector $v$ by the maximum value of its components, $v_{\text{MAX}}$:
    $$ v \rightarrow v / v_{\text{MAX}} $$
  b) Cell normalization representing L2 type normalization performed on each cell:
    $$ v \rightarrow v / \sqrt{\|v\|^2 + \varepsilon^2}, \text{ with } \varepsilon = 0.05 $$
    to prevent division by zero
  c) Block normalization, characterized by the fact that normalization step computed according to one of the above mentioned variants is performed over partially overlapping blocks, each containing 2 x 2 cells. For the considered input picture of 128 x 64 pixels, a feature vector of 3780 components is obtained.
2.2 Neural Classification

The considered classifier is based on the model of Concurrent Self-Organizing Maps (CSOM) [10], [11]. It is a collection of small SOM modules, whose number is equal to the number of classes. Each module is trained to correctly classify the patterns of one class. Since the goal of our system is to recognise the belonging of an unknown image to one of the two classes (pedestrian / non-pedestrian), the CSOM consists of two identical SOM modules. Each of the two modules is trained using only the sample images for its corresponding class. (Figs. 2 and 3).

2.2.1 Training of each SOM\(^{(k)}\) module (k=1, 2)

Assume that the module SOM\(^{(k)}\) has \(J^{(k)}\) neurons; particularly, one can choose

\[
J^{(1)} = J^{(2)} = \frac{J}{2}
\]

where \(J\) is the number of CSOM neurons.

For each \(SOM^{(k)}\) module, a specific training data subset is prepared containing all the training vectors having the label “\(k\)”, as shown in Fig. 2.

Assume also that the number of vectors having the class label “\(k\)” is \(N^{(k)}\), so that

\[
\sum_{k=1}^{2} N^{(k)} = N
\]

where \(N\) is the total number of training vectors. Usually, \(J^{(k)} >> N^{(k)}\), to use the interpolation capacity of CSOM.

![Fig. 2. The training phase of the CSOM model.](image)

![Fig. 3. The classification phase of the CSOM model.](image)
2.2.2 Recognition Phase

After the training phase, the system should be able to correctly classify an unknown feature vector as belonging to either the pedestrian class or to the non-pedestrian class, using the information stored in the neural network weights.

For the recognition, the test image feature vector has been applied in parallel to both previously trained SOM modules. The distances between all the neuron weight vectors of both modules and the input vector is computed. The neural module providing the minimum distance neuron is decided to be the winner and its index becomes the class index that the pattern belongs to (see Fig. 3).

In fact, CSOM is a system of systems having improved performances over a single big SOM with the same number of neurons, both from the point of view of recognition accuracy and for reducing the training time as well [10].

For comparison we have considered the classical statistical classifiers of nearest neighbour (NN) and K-Means (the nearest mean).

3 Experimental Results

3.1 Database

For training and testing the proposed system, we have selected pictures from INRIA Person Dataset, developed at the French National Institute for Research in Computer Science and Control [3].

In the training phase, we have used up to 1772 “positive samples” – block-images of 128 x 64 pixels containing persons, and therefore assigned to the PEDESTRIAN-class (see Fig. 4). We have also chosen up to 4458 “negative samples” – identical sized block-images that do not contain people - assigned to the NON-PEDESTRIAN-class (see Fig. 5).

3.2 Parameters for Performance Evaluation

In order to evaluate the performances of the proposed pedestrian detection algorithm, we have chosen the following six parameters:

- Correct Identification Rate [%]:
  \[ CIR = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \times 100 \ % \]

- Correct Rejection Rate [%]:
  \[ CRR = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \times 100 \ % \]

- False Positives Per Window:
  \[ FPPW = \frac{\text{False Positives}}{\text{True Negatives} + \text{False Positives}} \]

- Miss Rate:
  \[ MR = \frac{\text{False Negatives}}{\text{True Positives} + \text{False Negatives}} \]

- Total Success Rate [%]:
  \[ TSR = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \times 100 \ % \]

- Total Error Rate [%]:
  \[ TER = \frac{\text{False Positives} + \text{False Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \times 100 \ % \]

  \[ = 100 - \text{TSR} \ % \]

3.3 Pedestrian detection performances of the neural classifier

3.3.1 Influence of CSOM module architecture

In Table 1 and Fig. 6 one can see the experimental results concerning the influence of CSOM module architecture and size on the detection performances, by considering the algorithm variant without feature selection (Fig. 1) for input color images.
Table 1. Pedestrian detection performances as a function of CSOM module architecture (color images, without feature selection; number of training epochs = 30).

<table>
<thead>
<tr>
<th>CSOM structure</th>
<th>Number of training epochs</th>
<th>CIR [%]</th>
<th>CRR [%]</th>
<th>FPPW</th>
<th>MR</th>
<th>TSR [%]</th>
<th>TER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 x 10</td>
<td>30</td>
<td>94.1</td>
<td>92.8</td>
<td>0.118</td>
<td>0.08</td>
<td>98.1</td>
<td>6.9</td>
</tr>
<tr>
<td>10 x 20</td>
<td>30</td>
<td>94.2</td>
<td>94.1</td>
<td>0.058</td>
<td>0.08</td>
<td>94.1</td>
<td>5.9</td>
</tr>
<tr>
<td>10 x 20</td>
<td>cylindrical</td>
<td>92.8</td>
<td>91.6</td>
<td>0.084</td>
<td>0.072</td>
<td>92.2</td>
<td>7.8</td>
</tr>
<tr>
<td>10 x 20</td>
<td>toroidal</td>
<td>95.2</td>
<td>94.1</td>
<td>0.044</td>
<td>0.068</td>
<td>94.4</td>
<td>4.9</td>
</tr>
<tr>
<td>1 x 50</td>
<td>circular</td>
<td>94.4</td>
<td>84.0</td>
<td>0.18</td>
<td>0.056</td>
<td>98.5</td>
<td>10.5</td>
</tr>
</tbody>
</table>

Fig. 6. Total Success Rate (TSR) as a function of CSOM module architecture (color images, without feature selection).

3.3.2 Performances of CSOM versus K-Means for grayscale images

In Table 2 and Fig. 7 there are shown the results of the comparison experiments between a neural classifier (CSOM) and a simple statistical classifier (K-Means) by considering only the luminance component of the input pictures.

Table 2. Detection Performances of CSOM versus K-Means using only the luminance component (CSOM module architecture: square, 10 x 10 neurons per module; 10 training epochs).

<table>
<thead>
<tr>
<th>Feature Selection Method</th>
<th>Classifier</th>
<th>CIR [%]</th>
<th>CRR [%]</th>
<th>FPPW</th>
<th>MR</th>
<th>TSR [%]</th>
<th>TER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D HT</td>
<td>CSOM</td>
<td>92.5</td>
<td>99.6</td>
<td>0.004</td>
<td>0.075</td>
<td>96.05</td>
<td>3.95</td>
</tr>
<tr>
<td></td>
<td>K-Means</td>
<td>83.4</td>
<td>91.3</td>
<td>0.087</td>
<td>0.166</td>
<td>87.35</td>
<td>12.65</td>
</tr>
<tr>
<td>2D HT</td>
<td>CSOM</td>
<td>97.1</td>
<td>99.6</td>
<td>0.004</td>
<td>0.029</td>
<td>98.35</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>K-Means</td>
<td>87.6</td>
<td>90.7</td>
<td>0.083</td>
<td>0.124</td>
<td>89.15</td>
<td>19.86</td>
</tr>
<tr>
<td>HOG (a)</td>
<td>CSOM</td>
<td>94.2</td>
<td>99.5</td>
<td>0.005</td>
<td>0.058</td>
<td>96.85</td>
<td>3.15</td>
</tr>
<tr>
<td></td>
<td>K-Means</td>
<td>89.1</td>
<td>96.1</td>
<td>0.084</td>
<td>0.141</td>
<td>87.75</td>
<td>11.25</td>
</tr>
<tr>
<td>HOG (b)</td>
<td>CSOM</td>
<td>99.5</td>
<td>99.7</td>
<td>0.003</td>
<td>0.005</td>
<td>99.6</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>K-Means</td>
<td>93.1</td>
<td>95.3</td>
<td>0.041</td>
<td>0.099</td>
<td>94.2</td>
<td>5.8</td>
</tr>
</tbody>
</table>

Fig. 7. Detection performances of CSOM versus K-Means using only the luminance component.

3.3.3 CSOM Performances for color images

In Table 3 and Fig. 8 there are given the detection performances obtained by experimenting the proposed model for RGB representation of the input pictures.

Table 3. Detection Performances of CSOM for color images (RGB) versus grayscale input images (Y) (CSOM module architecture: square, 10 x 10 neurons per module; 10 training epochs).

<table>
<thead>
<tr>
<th>Feature Selection Method</th>
<th>Image Information</th>
<th>CIR [%]</th>
<th>CRR [%]</th>
<th>FPPW</th>
<th>MR</th>
<th>TSR [%]</th>
<th>TER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D HT</td>
<td>Y</td>
<td>92.5</td>
<td>99.6</td>
<td>0.004</td>
<td>0.075</td>
<td>96.05</td>
<td>3.95</td>
</tr>
<tr>
<td></td>
<td>RGB</td>
<td>95.6</td>
<td>99.3</td>
<td>0.007</td>
<td>0.064</td>
<td>96.45</td>
<td>3.55</td>
</tr>
<tr>
<td>2D HT</td>
<td>Y</td>
<td>97.1</td>
<td>99.6</td>
<td>0.004</td>
<td>0.029</td>
<td>98.35</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>RGB</td>
<td>97.7</td>
<td>99.6</td>
<td>0.004</td>
<td>0.023</td>
<td>98.65</td>
<td>1.35</td>
</tr>
<tr>
<td>HOG (a)</td>
<td>Y</td>
<td>94.2</td>
<td>99.5</td>
<td>0.005</td>
<td>0.058</td>
<td>96.85</td>
<td>3.15</td>
</tr>
<tr>
<td></td>
<td>RGB</td>
<td>93.8</td>
<td>99.6</td>
<td>0.004</td>
<td>0.062</td>
<td>96.7</td>
<td>3.3</td>
</tr>
<tr>
<td>HOG (b)</td>
<td>Y</td>
<td>96.7</td>
<td>99.1</td>
<td>0.009</td>
<td>0.013</td>
<td>98.9</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>RGB</td>
<td>98.6</td>
<td>99.4</td>
<td>0.006</td>
<td>0.014</td>
<td>99</td>
<td>1</td>
</tr>
<tr>
<td>HOG (c)</td>
<td>Y</td>
<td>99.5</td>
<td>99.7</td>
<td>0.003</td>
<td>0.005</td>
<td>99.6</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>RGB</td>
<td>99.9</td>
<td>99.7</td>
<td>0.003</td>
<td>0.003</td>
<td>99.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Fig. 8. Detection Performances of CSOM in the variants color (RGB) versus grayscale (Y).

4 Concluding Remarks

1) The paper presents a neural network approach for pedestrian detection using the classifier called Concurrent Self-Organizing Maps (CSOM), representing a winner-takes-all collection of two neural modules. First module is trained with pedestrian images and the second one is trained with non-pedestrian images. For feature selection one chooses the following three candidate techniques (a) Histogram of Oriented Gradients (HOG) /1D Haar transform/2D Haar transform.

2) By considering the algorithm variant without feature selection for input color images, from the results given in Table 1, one deduces that the best performances correspond to the choice of a rectangular neural module with 10 x 20 neurons; this implies a maximum TSR = 94.1 % and a minimum FPPW = 0.058.
3) The results of performance evaluation taking into account CSOM vs. K-Means for several feature selection techniques (considering only the luminance component), one obtains the best TSR = 99.6 % for CSOM cascaded with HOG (c) variant, by comparison to TSR = 95.25 % for K-Means cascaded with the same feature selection variant (see Table 2).

4) For RGB representation of color images, the best results correspond to the cascade CSOM-HOG (c), leading to TSR = 99.7 %; this represents a slight increasing over the luminance representation of input pictures.

References:


