A Road Sign Recognition System Based on Dynamic Visual Model

[1]C. Y. Fang, [2]C. S. Fuh, [3]S. W. Chen, and [1]P. S. Yen
[1]Department of Information and Computer Education
National Taiwan Normal University, Taipei, Taiwan, R. O. C.
[2]Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan, R. O. C.
[3]Department of Computer Science and Information Engineering
National Taiwan Normal University, Taipei, Taiwan, R. O. C.
E-mail:violet@ice.ntnu.edu.tw

Abstract

We propose a computational model motivated by human cognitive processes for detecting changes of driving environments. The model, call dynamic visual model, consists of three major components: sensory, perceptual, and conceptual components. The proposed model is used as the underlying framework in which a system for detecting and recognizing road signs is developed.

1. Introduction

The objective of road signs is to guide, warn, and regulate traffic. They supply information to help drivers operate their cars in such a way as to enhance traffic safety. However, when drivers get tired, they may not always notice road signs. A stable road sign detection and recognition system is thus desirable to alert the driver to presence of signs.

Road signs use particular colors and geometric shapes to attract drivers' attention. However, the difficulty in recognizing road signs is largely due to the following reasons: (1) Colors may fade after long exposure to the sun. Moreover, paint may even flake or peel off, and signs may get damaged. (2) Air pollution and weather conditions may decrease the visibility of road signs. (3) Outdoor lighting conditions vary from day to night and may affect the apparent colors of road signs. (4) Obstacles, such as trees, poles, buildings, and even vehicles and pedestrians, may occlude or partially occlude road signs. (5) Video images of road signs often suffer from blurring in view that the camcorder is mounted on a moving vehicle.

Many techniques have been developed to detect and recognize road signs. Pacheco, Batlle, and Cufi [7] proposed adding special color barcodes under road signs to help road sign identification for vision-based systems. However, much time and resources would be expended to replace road signs, making this solution uneconomical. Aoyagi and Asakura [1] used genetic algorithms to detect road signs from gray-level video imagery. Unfortunately, due to the discrete nature of crossover and mutation operators, optimal solutions are not guaranteed. Lalonde and Li [6] reported a color indexing approach to identify road signs, but the computation time will increase greatly in complex traffic scenes. In addition, many other studies on detecting and recognizing road signs by morphological methods, neural networks, and fuzzy reasoning have been reported.

Two potential problems with an automatic road sign detection system are that if it analyzes and reports a critical situation too slowly or if it makes errors, then the system would be of little use. Unfortunately, the above difficulties keep bothering researchers. We may appeal to the human visual system for a solution.

2. Computational Model

Figure 1 depicts the proposed dynamic visual model (DVM), which captures several aspects of the human visual process [2, 4, 5, 8]. The proposed model is comprised of three major components, the sensory, perceptual, and conceptual analyzers of the human visual system. The input to the model are video sequences. Video sequences can be subsampled to a degree appropriate to an application so as to reduce the quantity of the input data. In addition to data reduction, the format and structure of input data may be converted in order to increase the effectiveness of later processing. We refer to this stage of data reduction and conversion as the data transduction stage, since it corresponds to the transducer of the human visual system.

The transduced data are forwarded to the sensory component of the DVM to extract spatial and temporal information. Spatial information sketches the relations between objects in a single image, and temporal information describes the change of objects between successive images. All these kinds of information are important for correct detection and recognition. In the sensory component, we extract the temporal and spatial information of moving objects from the input video sequence.

In the perceptual component, a voluntary selectivity of attention is realized by introducing a module called the spatiotemporal attentional (STA) neural module, as well as a long term memory (LTM), which preserves the characteristics of the objects of interest. The information from the LTM will call the attention of the neural network

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

to the objects of interest when it is being innervated by the stimuli coming from the sensory component. Then the activations of the STA neurons are examined. If there is no focus of attention formed over the neurons, the system repeats the above process. Otherwise, the feature extraction step is evoked to detect categorical features of the objects within the image areas corresponding to the focuses of attention in the STA neural module.

The categorical features obtained in the perceptual component serve as the input stimuli, represented as a supraliminal pattern, to a CART neural module in the conceptual component. The input supraliminal pattern first initializes the LTM of the CART neural module with the contents coming from a system memory, called the episodic memory. The configurations of the LTM and the associated components of the neural module have to be adapted in accordance with the contents. This adaptability of configuration is referred as the configurable capability of the neural module. Subliminal patterns to be matched with the input supraliminal pattern will be retrieved from the LTM, for which the search space of subliminal patterns is greatly reduced. The supraliminal pattern is compared with a subliminal pattern, and if they are similar enough, the class of the supraliminal pattern is regarded as that of the subliminal pattern under consideration. The CART module then performs a supervised learning through which the subliminal pattern in the LTM is updated under the guidance of the input supraliminal pattern. On the other hand, if no subliminal pattern is similar to any supraliminal pattern, an unsupervised learning, which represents the supraliminal pattern as a new subliminal pattern, is carried out.

After the classification stage, particular object features regarding the special category are extracted and fed into a CHAM neural module, which is the recognition stage in the conceptual component. Similar to the classification stage, the supraliminal object feature pattern first initializes the LTM of the CHAM module with the contents coming from the episodic memory. If the supraliminal pattern adequately matches a subliminal pattern, the supraliminal pattern is recognized successfully. Otherwise, our system is in a new situation and will attempt to learn and memorize the new experience for future recognition.

3. Road Sign Recognition System

3.1. Sensory Component

The data input to our system are color image sequences acquired using a camcorder mounted on a moving vehicle. In the sensory analyzer of our system, spatial and temporal information of dynamic scenes is extracted from the input video sequences, and noise is filtered out. The sensory analyzer is a primary analyzer which concerns itself only with local information. In road sign detection, color is a local feature which can be extracted from individual pixels. On the other hand, shape is global feature which must be decided by a neighborhood of pixels.

As mentioned previously, road sign detection is very difficult under poor weather conditions because of the influence of constantly varying outdoor illumination and optical distortion. Even though the actual colors of road signs are initially quite well controlled, the perceived colors are affected by illumination from light of various colors in their natural settings. Moreover, due to the effects of sunlight, the paint on signs often gradually fades. The hue component in the HSI model is invariant to brightness and shadows. Thus the hue component is suitable for extracting color features, given the uncertainty of weather and natural and artificial damage to road signs.

There are one-to-one mappings of sensory analyzers to the pixels of an input image, and a sensory analyzer processes only the information coming from a single pixel. First, the hue value, h, of each pixel is calculated. Then, the similarity between h and the stored hue values of particular colors in road signs is calculated. Let $\{h_1, h_2, \ldots, h_q\}$ be the set all the hue values of particular colors in road signs which are assumed to be Gaussianly distributed with variance σ^2. Then the output of the sensory analyzer is the degree of similarity

$$z = \max_{k=1,\ldots,q} (z_k), \text{ where } z_k = \frac{1}{\sqrt{2\pi}\sigma} \exp((h-h_k)^2 / 2\sigma^2).$$

Finally, the outputs of sensory analyzers are fed into the perceptual analyzer. Figure 2 gives an example showing the result of the sensory analyzers. The input image is shown in Fig. 2 (a). There are two road signs in the traffic scenes, one red and the other blue. The output of the sensory analyzers is shown in Fig. 2 (b) where the intensity of each pixel indicates its degree belonging to a road sign color.

3.2. Perceptual Component

We give only a brief description of the STA neural module in this subsection; more details can be found in our previous study. The STA neural module is structured as a two-layer network: one for input and one for output. The output layer is also referred to as the attentional layer. Neurons in this layer are arranged into a 2D array in which they are connected to one another. These connections are within-layer connections and are almost always inhibitory. There are no synaptic links among input neurons; they are, however, fully connected to the attentional neurons. These connections are called between-layer connections and are always excitatory.

The input neurons are also organized into a 2D array as are the attentional neurons and the size of both the arrays be the same as that of the input images. Let w_{ij}

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

denote the weight of the link between attentional neuron n_i and input neuron n_j. The weight vector of attentional neuron n_i is written as $\mathbf{w}_i = (w_{i1}, w_{i2},\ldots,w_{im})$, where m is the number of input neurons. The activation of attentional neuron n_i due to input stimuli $\mathbf{z}$ (coming from sensory components) is

$$a_i(t) = \psi(a_i(t-1) + B(-\alpha a_i(t-1) + \beta A(I_i^v + I_i^l - \Gamma_n))),$$

where $I_i^v = \sum_{j=1}^{m} w_{ij} z_j$, $I_i^l = \sum_{k\in N_i, k\neq i} [u_{ik} M(\mathbf{r}_{ik}) a_k(t-1)]$,

$$A(v) = \begin{cases} v & \text{if } v > 0 \\ 0 & \text{if } v \le 0 \end{cases}, \quad B(v) = \begin{cases} v & \text{if } v > 0 \\ \gamma v & \text{if } v \le 0 \end{cases}, \quad \text{and}$$

$1 > \gamma > 0$.

In above equations, a_k is the activation of neuron n_k, threshold Γ_n prevents the effect due to noise, and α and β are positive parameters. Set N_i indicates the neighboring set of attentional neuron n_i; u_{ik} is the linking weight between neurons n_i and n_k; $M(\mathbf{r}_{ik})$ denotes a "Mexican-hat" function, and the parameter $\mathbf{r}_{ik}$ is the distance between neurons n_i and n_k.

Selective attention was realized by the STA neural module. Two important features of our module are that top-down expectations have been embedded beforehand in the input stimuli to the module so as to save processing time, and that both spatial and temporal information are managed in one construct so as to reduce the cost of the configuration.

The input stimuli of the STA neural module are the outputs of the sensory components, shown in Fig. 2 (b). Figure 2 (c) shows the corresponding attention map of the STA neural module. Once the focus of attention is developed, the following subsystems will pay attention to only the area of interest and ignore the rest of the input pattern.

Categorical features utilized to partition road signs into groups should represent common characteristics of the groups, not the specific ones of road signs. First, create an edge image E from the input Z of perceptual component and the attention map M of STA neural module. For each pixel (x, y), $E(x, y)$ is calculated by

$$E(x,y) = \begin{cases} |E'(x,y)| & \text{if } M(x,y) > 0 \text{ and } Z(x,y) > \Gamma_c \\ 0 & \text{otherwise} \end{cases},$$

where $|E'(x,y)|$ is the absolute edge magnitude of pixel (x, y), and Γ_c is the similarity threshold to determinate whether the colour of pixel (x, y) has a road sign colour. Second, by combining the colour and edge information, we can locate the candidate positions of road signs. Let Q contain the candidate positions of road signs. For each position (x, y),

$$Q(x,y) = \begin{cases} 1 & \text{if } E(x,y) > 0 \text{ and } Z(x,y) > 0 \\ 0 & \text{otherwise} \end{cases}.$$

A pre-attention map is used for modelling the expectation of the human brain. In the pre-attention map, the places where we expect road signs to be located have stronger stimuli than others. Now, we combine this prior information in Q:

$$P(x,y) = \begin{cases} 1 & \text{if } M^*(x,y) > 0 \text{ and } Q(x,y) = 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$M^*(x,y) = \min(M(x,y), M^p(x,y)),$$

where $M^p(x, y)$ is the pre-attention stimulus of pixel (x, y). Figure 3 gives an example to illustrate the effect of the pre-attention map. The original input stimulus is shown in Fig. 3 (a), and the pre-attention map is presented in Fig. 3 (d). Figures 3 (b) and (e) show the attention maps, M and M^*, respectively. Compared with M, the attention map given by M^* is more concentrated on the road signs, and noise near the ground is filtered out. Finally, patterns Q and P are shown in Fig. 3 (c) and (f), respectively. Next, we use the connected components technique to detect the road signs. In summary, road sign detection is accomplished by these steps.

After the road sign detection stage, the perceptual component extracts the categorical features input to the conceptual component. The categorical features indicate the colour horizontal projection of the road signs. In the colour horizontal projection, all gray pixels are treated as the same colour to eliminate the individual difference among road signs of the same class. Figure 4 shows the eight classes stored in CART.

3.3. Conceptual Component

The categorical feature extracted in the perceptual component serves as a supraliminal feature to be fed into the CART module in the conceptual component. The CART module is actually an ART2 neural network [3] with a configurable long term memory (CLTM). Figure 4 shows a classification result of the CART module. We scan sixteen road sign images and extract their categorical features to train and test our system. Sixteen categorical features of road signs are applied to the CART module. These features are classified into eight classes so that similar features will be classified into the same class. This result is memorized as the learned experience and used to classify the subsequent input features.

The CHAM neural module is structured as a two-layer network with one input layer and one output layer. The output layer is a winner-take-all competitive layer. In the input layer, neurons are arranged into a 2D array, and there are no within-layer synaptic links among these neurons. Suppose that the input layer of the neural network contains of m neurons and the output layer contains n neurons. Let w_{ij} denote the weight representing the strength of the link between output neuron i and input neuron j. The weight vector of neuron i is written as $\mathbf{w}_i = (w_{i1}, w_{i2},\ldots, w_{im})$. The net input to neuron i on the competitive layer due to innervation $\mathbf{z}$ is computed from

$$net_i = \mathbf{w}_i \cdot \mathbf{z} = \sum_{j=1}^{m} w_{ij} z_j ,$$

and the winner, n_c, after the competition can be found by $n_c = \arg(\max_i(net_i))$. Finally, only the winner on the competitive layer outputs a one while the rest output zeros.

$$v_i = \begin{cases} 1 & \text{if } i = n_c \\ 0 & \text{otherwise.} \end{cases}$$

Let $\{ \Theta_1, \Theta_2, ..., \Theta_p \}$ be the set of object feature patterns stored in the CHAM, where $\Theta_i = (\Theta_{ib}, \Theta_{iw})^T$, and $i = 1,2,...,p$. Patterns Θ_{ib} and Θ_{iw} represent the black and white feature patterns, respectively, of the *i*th road sign. If an input object feature pattern, Θ, is fed into the CHAM, then the output class corresponds to $\Lambda = \arg(\min_{i=1,...,p} |\Theta - \Theta_i|)$, where $|\Theta - \Theta_i|$ is the distance between Θ and Θ_i.

Figure 5 shows an example of the regulatory sign recognition. The CHAM is trained by the set of 31 regulatory signs, and one part of these signs are shown in column (a). The white and black object features extracted from the training set are shown in columns (b) and (c), respectively. Three test sets are prepared for the test stage. The first test set, shown in column (d), contains the smoothed images of those signs in column (a) with 5x5 neighborhood averaging. Their white and black object features extracted from the training set are represented in columns (e) and (f), respectively. The second and third test sets, shown in column (g) and (j), is comprised of the regulatory signs corrupted by 20 and 30 percent uniform noise, respectively. All 31 smoothed signs and 46 signs with noise are recognized correctly.

Figure 6 gives another example of regulatory sign recognition. The training set is the same as shown in Fig. 5 (a) but the test patterns are extracted from real images captured by camcorder. Their white and black object features are shown in columns (b) and (c), respectively. Column (d) shows the recognition results. Although these road signs are imperfect, they are still recognized.

The examples show how to recognize road signs in a single image. However, since the data input to our system are video sequences, we can collect more information during several successive images to make a better decision. Due to the road signs on the roadside getting closer to the vehicle, the visual sizes of road signs projected in the video images continuously increase in size and clarity. The road signs should still be very small when first detected, but our system may have difficulty recognizing these small signs. However, such signs still supply valuable information for eliminating the impossible candidates of road signs and reducing the search space. The more video images fed into our system, the more information can be used to strengthen our decision.

Suppose our system initially maintains all the candidates for road signs in the LTM of the CHAM network, then the candidate number p with the video input will be reduced to only one by the following procedure:

(1) Put all candidates into the LTM.

(2) Input the extracted object feature Θ of an image in a video sequence into the CHAM network.

(3) For each candidate Θ_j in the LTM, apply the following rule:

If $|\Theta - \Theta_j| > \rho(t)$, then remove the *j*th candidate from the LTM, where $j = 1,2,...,p$ and $\rho(t)$ is a constant depending on the size of the road sign projected on the image.

(4) Repeat Steps (2) to (4) until only one candidate is left, and output the candidate as the recognition result.

This procedure illustrates the recognition process for a road sign. If there are two or more different signs in the image sequence, the recognition process is the same except that the contents of the LTM should be modified for different signs.

4. Experimental Results

The input data to our system were acquired using a video camcorder mounted in the front windshield of a vehicle while driving on expressways. In our experiments, each video sequence was down-sampled to a frame rate of 5 Hz before being submitted to the system. Furthermore, each 720 x 480 pixel input image was reduced to 180 x 120 pixels. We downsample input video sequences for the purpose of reducing the processing load on the computer. Likewise, we subsample video images for reducing the processing time.

A number of video sequences were collected for experiments with one or two signs included in each sequence. One example of the experimental results is presented in Fig. 7. It shows only part of a video sequence (the seven images in column (a)). In this sequence, two road signs should be recognized, one is a speed limit sign and the other is a warning sign. The corresponding attention maps of the input images, column (a), are shown in column (b). Column (c) represents the detection results of the candidate road signs. We frame these road signs with white boxes. Column (d) shows the contents of the white boxes. They are extracted from the input image and normalized to 60 x 60 pixels. The recognized results, shown in column (e), are output only when one candidate is left in each LTM. In this example, the warning sign is recognized first, and the speed sign is recognized later.

Figure 8 shows another six examples of experimental results. In each sequence we select only one image to represent the whole sequence in this figure. Columns (a) and (d) are the selected images of the input video

sequences, and columns (b) and (e) give their detection results. The recognition results are presented in columns (c) and (f).

The CART and CHAM networks should be well trained before being tested. Some training patterns of CART are shown in Fig. 4. These patterns are first normalized to 60 x 60 pixels, and then the categorical feature vectors are extracted to train the CART network. Although the patterns in the training set were made in a computer, they can be used to classify real patterns captured by a camcorder. In our experiments additional patterns are not need training for the CART network.

For each category stored in the CART networks, a corresponding LTM of the CHAM network should be trained. The weights in these LTMs are recorded in episodic memory, and will be moved to the LTM as they are needed.

Although most road signs in input images are detected at the detection stage, there are some misdetections. Misdetection usually occurs when the road signs are small in the image. Since our system integrates the results of several successive images to make a decision, a few misdetections in a sequence does not affect the decision making of our system.

If the speed of a vehicle is very high, then vehicle and camcorder vibration cannot be avoided, and the quality of input video sequences is reduced. Some patterns extracted from these sequences appear jerky, though they are detected and classified correctly. However, the recognition of these patterns (about 1/10) may be incorrect. The incorrect results do not affect the correctness of system's decision since the decision is made by integrating several images.

5. Conclusion and Future Work

In this paper, a computational model motivated by human visual cognitive and recognition processing was presented. The road sign recognition system is not the only subsystem in vision-based driver assistance systems. There are several other subsystems, which perform obstacle recognition, and environmental change detection, etc. Developing and integrating these subsystems to collect significant information in driving environments is very important for improving traffic safety. We hope the proposed DVM is helpful for designing various subsystems for functions. Moreover, other applications to event detection and recognition can also be accomplished with this model by extracting different kinds of features.

Acknowledgements

This work was supported by the National Science Council, Republic of China, under Contract NSC-91-2213-E-003-003. The authors gratefully acknowledge the assistance of Prof. Robert R. Bailey of National Taiwan Normal University for his many helpful suggestions in writing this paper and for editing the English.

References

[1] Y. Aoyagi and T. Asakura, "A Study on Traffic Sign Recognition in Scene Image using Genetic Algorithms and Neural Networks," *Proceedings of the IEEE IECON International Conference on Industrial Electronics, Control, and Instrumentation*, Taipei, Taiwan, pp. 1838-1843, 1996.

[2] J. J. Clark and N. J. Ferrier, Model Control of an Attentive Vision System, *Proceedings of International Conference on Computer Vision*, ampa, pp. 514-523, 1988.

[3] J. A. Freeman and D. M. Skapura, *Neural Networks---Algorithms, Applications, and Programming Techniques*, Addison-Wesley, Readings, Massachusetts, 1992.

[4] R. L. Gregory, *Eye and Brain*, 3th Ed., McGraw-Hill, New York, 1978.

[5] A. Hurlbert and T. Poggio, "Do Computers Need Attention," *Nature*, Vol. 321, pp. 651-652, 1986.

[6] M. Lalonde and Y. Li, "Detection of Road Signs Using Color Indexing," Technical Report CRIM-IT-95/12-49, Centre de Recherche Informatique de Montreal, *http://www.crim.ca/sbc/english/cime/ publications.html*, 1995.

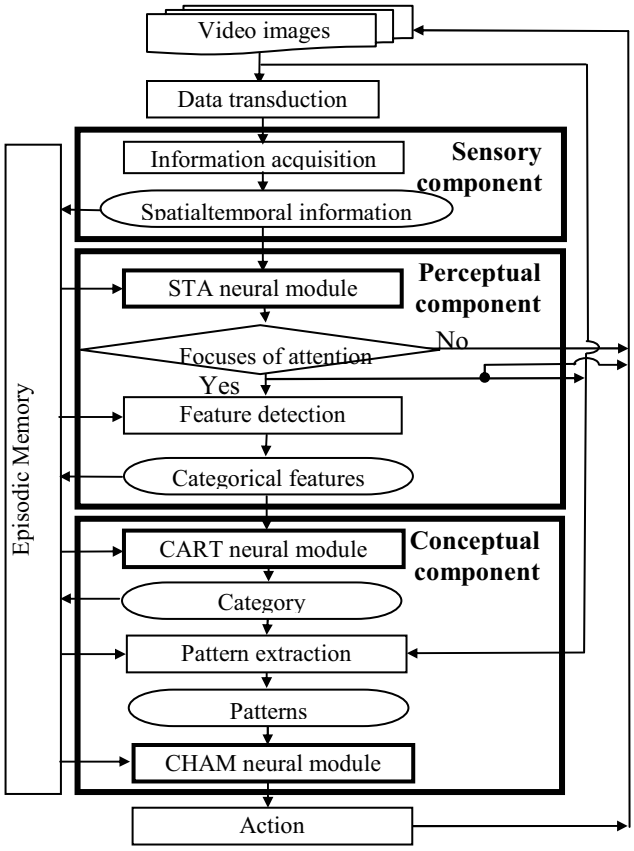
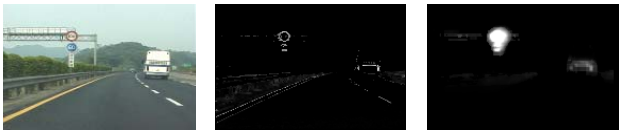
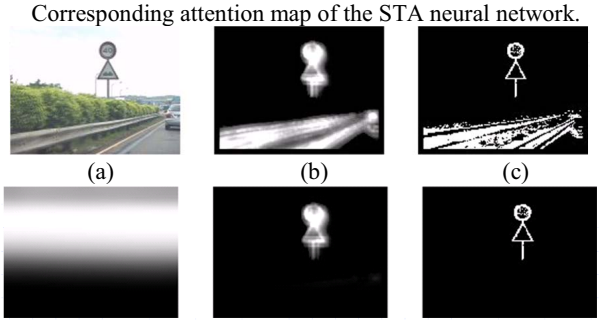


Fig. 1. The proposed DVM.



(a) (b) (c)

Fig. 2. An example of the attention map of the STA neural network. (a) One image of an input video sequence. (b) Corresponding outputs of the sensory components. (c) Corresponding attention map of the STA neural network.



(a) (b) (c)

(d) (e) (f)

Fig. 3. Experimental results with and without pre-attention. (a) Original image. (b) Attention map without pre-attention. (c) Regions containing road sign candidates without pre-attention. (d) Pre-attention map. (e) Attention map with pre-attention. (f) Regions containing road sign candidates with pre-attention.

(a) (b) (c) (d) (e) (f) (g) (h)

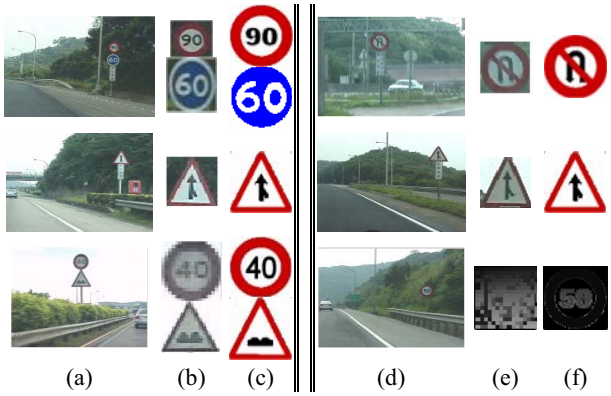
Fig. 4. Eight classes are stored in CART neural module. (a) Warning signs. (b) Regulatory signs. (c) Construction signs. (d) Construction signs. (e) Guide signs. (f) Guide signs in highway. (g) Information signs in highway. (h) Guide signs in highway.

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 5. A partial experimental result of regulatory signs recognition. (a) Training set of the regulatory signs. (b) White object feature of column (a). (c) Black object feature of column (a).(d) Blurred test set. (e) White object feature of column (d). (f) Black object feature of column (d). (g) Noisy test set. (h) White object feature of column (g). (i) Black object feature of column (g).(j) Noisy test set. (k) White object feature of column (j).(l) Black object feature of column (j).

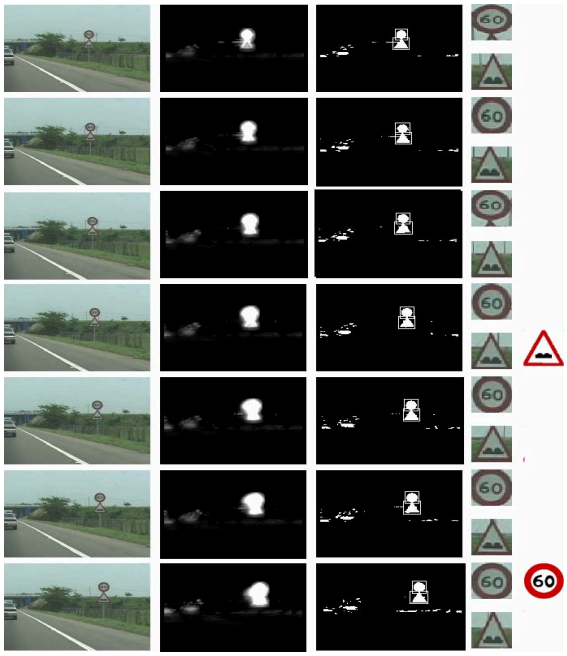
(a) (b) (c) (d) (a) (b) (c) (d)

Fig. 6. Recognition results of real road sign patterns. (a) Test images captured by camcorder. (b) Corresponding white object feature of column (a). (c) Corresponding black object feature of column (a). (d) Their recognition results.



(a) (b) (c) (d) (e) (f)

Fig. 8. Some experimental results of road sign detection and recognition. Columns (a) and (d) are the input video sequences. Columns (b) and (e) show their detection results. Columns (c) and (f) represent the recognition results.



(a) (b) (c) (d) (e)

Fig. 7. Result of road sign detection and recognition. (a) The input video sequence. (b) The corresponding attention map. (c) The results of road sign detection. (d) The road signs extracted after categorical analyzer. (e) The recognition result.

[7] L. Pacheco, J. Batlle, and X. Cufi, "A New Approach to Real Time Traffic Sign Recognition Based on Colour Information," *Proceedings of the Intelligent Vehicles Symposium*, Paris, pp. 339-344, 1994.

[8] K. Toyama and G. D. Hager, Incremental Focus of Attention for Robust Visual Tracking, *Proceedings of Computer Vision and Pattern Recognition*, San Francisco, California, pp. 189-195, 1996.