MVA '92 IAPR Workshop on Machine Vision Applications Dec. 7-9,1992, Tokyo

NEURAL NETWORK APPROACHES FOR ATTRACTIVE AREA EXTRACTION FROM VIDEO IMAGES

Jun Ogata, Mikiya Sase and Yukio Kosugi

Tokyo Institute of Technology
Interdisciplinary Graduate School of Science and Engineering
4259, Nagatsuta, Midori-ku,
Yokohama 227, Japan

ABSTRACT

On-line machine vision algorithms are potentially applicable for various systems, including robot vision, security and guarding systems, clinical image diagnostic systems, and automatic video editting systems. However, in processing tremendous amount of image data in a stream of video signal, software burden required for image understanding sometimes restricts the realization of the systems.

In our biological system, we first extract some essential scenes out of continuously in-coming stream of visual data given to the eyes both in temporarily and spatially, before through analysis. Furthermore, in understanding an image, our attention is focused on some narrow area of the image, e.g., the face of a person.

We realized the above functions in neural network approaches. In finding out the important sub-area,we used three indices; color and brightness index, space frequency index, and symmetry index. Using hierarchically arranged BP networks, we realized such image processing with high speed. At the training stage, we gave the correct answers, obtained from human subjects through visuo-psychological experiments. After finishing with the training, the network system revealed reasonable results even to untrained images. This system will be helpful for data compression required for high-speed analysis in automatic image recognition systems as well as for editting video programs.

INTRODUCTION

When we think of our visual processing system in the brain, we do not analyze nor memorize whole part of a long sequence of images. We unconsciously concentrate our attention to some significant parts, e.g., to discontinuous part, of images in a stream of video data. Furthermore, in understanding an image, our attention is focused on some narrow area of the image, e.g., the face of a person. Most of the background area of the image will not significantly contribute to the understanding of the image. That is, in our biological system, we first extract some essential scenes out of continuously in-coming stream of visual data given to the eyes both in temporarily and spatially, before through analysis.

The above consideration motivated us to built up a pre-processing system capable of 1.recognizeing the sequential image discontinuity due to the change of environment, to cut the image sequence to a set of scenes, 2.extracting the most significant sub-area of an image to condense the individual image's features.

The above functions are realized in neural networks. Especially, in finding out the important sub-area, we used three indices; color and brightness index, space frequency index, and symmetry index. We arranged hierarchical BP network, and tried to realize such image processing with high speed. Through visuo-psycological experiments, we obtained correct answers from human subjects and gave them to the network for it's training. After finishing with the training, the network system is expected to give right answer even to untrained images.

EXTRACTION OF ATTRACTIVE AREA

Network Model

The neural network system that we propose in this paper is shown in Fig.1. This network system has two parts, one is the image-feature-processing networks (we call it as the pre-processing network) that treates feature indicies of color images, and the other is the feature-integration network that fuses the information processed at the pre-processing network. To make the learning data set for the network, we made experiments for volunteers. To show an image on the video display, and ask volunteers to select the most inportant area on the image (each image has 25 areas). The training of the pre-processing networks stop when the network converged to some level, then we start training of the feature-integration network. The feature-integration network consisits of a layer of 9 input units, a hidden layer of 5 units, and a layer of 3 output units. Input for the network is a divided area of image, and output of the network is the judgement that current processing area is adequate to the candidate for the most

important area of the image or not. In training the feature-integration network, the teaching signals are given as same as the training of the pre-processing network.

In the following we will explaine the function of the pre-processing network more in detail.

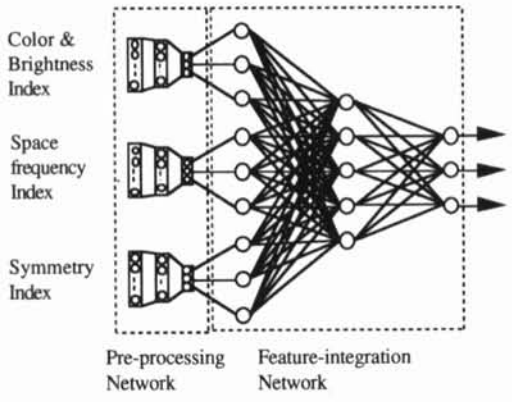


Fig.1 The Neural Network system

Pre-processing Network

For image processing, we considered three indicies of image; color and brightness index, space frequency index, and symmetry index. And we prepared three pre-processing networks to process each index. In the processing, we assumed a window (this window covers 25% area of the image) on the image and the processing was performed to the data within the window area. This idea is based on the following:

[Restricted area image recognition on the visuo-motor]

When a human subject tried to recognize the image, it needs 30~50% size of image area for recognition and understanding of the image[1]. On the other hand, for the image processing point of view using indicies of image, a small data size is desirable. So we used 25% size of image area for the present study.

Color-and-brightness index processing network

Fig.2 shows the network that treats color and brightness index of image.

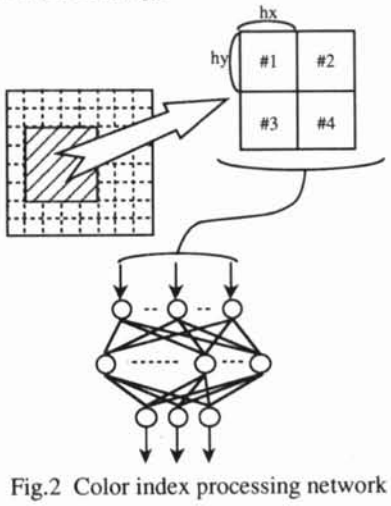


Fig.2 Color index processing network

We used the red and green indecies and brigtness index of the images. To make the input data for network, we first divide the window area into four parts, and calcurate each index value using (1)~(2).

$$\text{AreaColor}[i] = \frac{\sum_{y=0}^{hy-1}\sum_{x=0}^{hx-1} \text{color}(x,y)}{hx \times hy} \quad (i=1\sim4) \tag{1}$$

$$\text{AreaBright}[i] = \frac{\sum_{y=0}^{hy-1}\sum_{x=0}^{hx-1} \text{bright}(x,y)}{hx \times hy} \quad (i=1\sim4) \tag{2}$$

From the four parts of the window, we have twelve input data to be applied to the network. The network consisits of a layer of 12 input units, a hidden layer of 8 units, and a layer of 3 output units.

Space-frequency index processing network

This network treates power spectrum in the spatial frequency domain obtained from FFT analysis of horizontally scanned image data. As shown in Fig.3, we set up three detecting lines in the window, and calcurate FFT on each line. From the results, we combined power spectrum data on the three lines according to (3) and finally 16 data are given to the network, where $fft^j[i]$ is the power spectral density at spatial frequency i, of j th scanning.

The network consisits of a layer of 16 input units, a hidden layer of 8 units, and a layer of 3 output units.

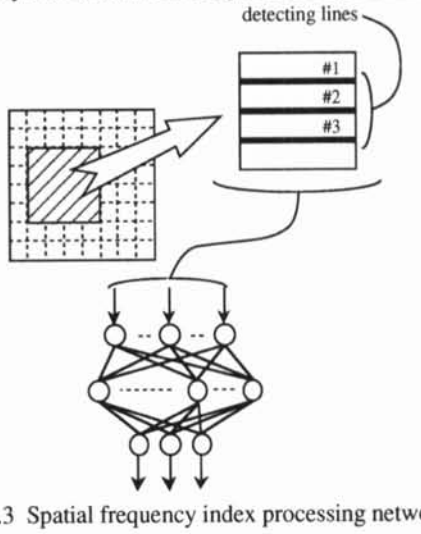


Fig.3 Spatial frequency index processing network

$$\text{in}[i] = \frac{1}{3}\sum_{j=1}^{3} (fft^j[i] + fft^j[i\pm1] + fft^j[i\pm2]) \quad (i=1\sim3) \tag{3}$$

Symmetry processing index network

This network examines whether there is a highly symmetrical structure in the window or not, by using the brightness signal of the image. As shown in Fig.4, we divided the window into 25 areas and each area has a tentative vertical axis of symmetry in it's centre. Calcurating symmetrisity index on each axis using (4) and put the resutlts to the network.

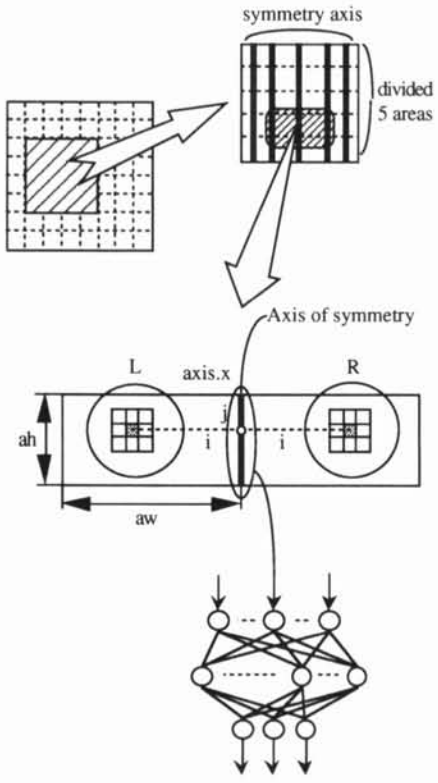


Fig.4 Symmetry index processing network

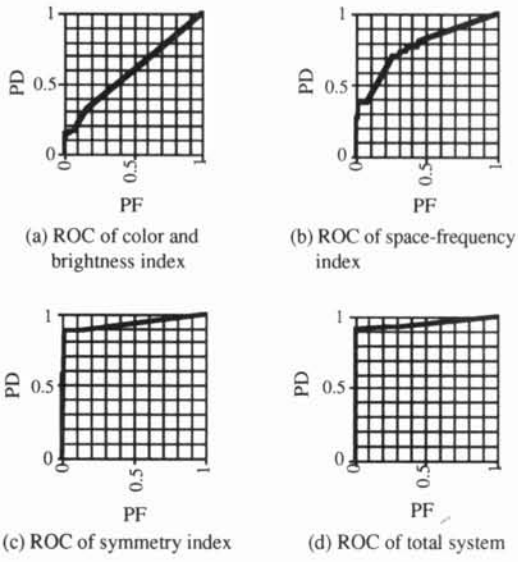
$$sym[n] = \sum_{j=1}^{ah-1} \sum_{i=1}^{aw-1} \left| \frac{\sum_{v=j-1}^{j+1} \sum_{u=i-1}^{i+1} L(axis.x-u, v)}{9} - \frac{\sum_{v=j-1}^{j+1} \sum_{u=i-1}^{i+1} R(axis.x+u, v)}{9} \right| \quad (4)$$

This network only examines the horizontal symmetrisity of image. The network consisits of a layer of 25 input units, a hidden layer of 12 units, and a layer of 3 output units.

Each network mentioned above is trained by using Back Propagation Method[2].

RESULTS

The results of these networks are evaluated by using the Receiver Operating Characteristics (ROC)[3]. Fig.5(a) ~ (c) are each pre-processing networks' ROC. To see them, each networks' recognition results are not sufficient. Especially in the case of color and brightness index(a), the probability of successful detection (PD) does not approach unity unless we lower the threshold level of the network's output unit at the sacrifice of increased false detection probability(PF). Among them, the ROC of the symmetry index network is better. And when fusing them, the recognition result of the feature-integration network (Fig.5(d)) is better than that of the symmetry index alone. So, it can be concluded that, though each pre-processing networks' abilities are not sufficient, by fusing these networks the total ability becomes better.



(a) ROC of color and brightness index

(b) ROC of space-frequency index

(c) ROC of symmetry index

(d) ROC of total system

Fig.5 ROC of each network

TEMPORAL SEPARATION OF THE IMAGE SEQUENCE

To edit video image sequences that consist of many scenes, it needs to find out the connecting points, i.e., the semantic discontinuity of image scenes.

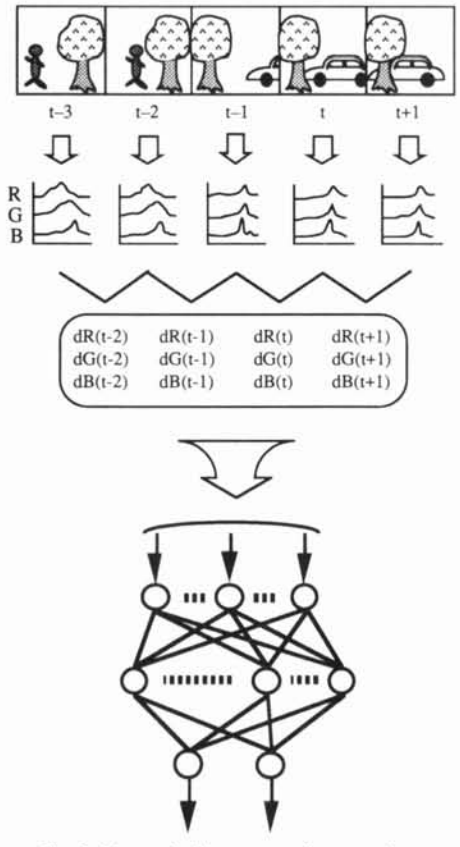


Fig.6 Network for temporal separation

To find them out automatically, we examined RGB histogram of a color image and difference of two successive images' histograms. For the two continuous image frames, difference of two histograms is very small, while for the pair of discontinuous image frames, it shows high value. So, it is usefull to take this index for finding the connecting points out of image scene sequences. To realize it, we used a simple neural network shown in Fig.6. This network has three layers. The input for the network is a differential information of four succesive images' histograms. These histograms are obtained from a current image(T=t) and two backword images(T=t-1, t-2) and one forword image(T=t+1). The network yields the judgement whether the connecting point of images(T=t, t-1) are continuous or not.

Network was trained by using BP method. Fig.7 shows the network's training process, where the learning coefficient of α is taken as a parameter. According to this, network shows rapid convergence irrespective of it's learning coefficient, α. After training, this network can recognized the connecting points effectively.

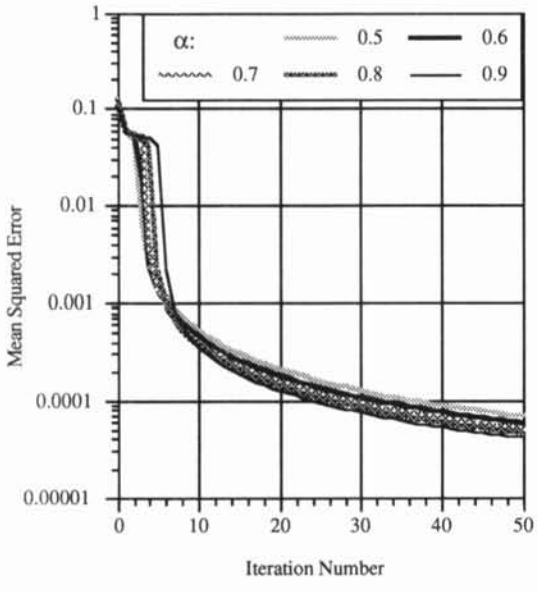


Fig.7 Convergence of discontinuity detection network

CONCLUSIONS

In this paper, we suggested the possibility of attractive area extraction from video images by using neural network approaches. The processing system is implemented with three pre-processing networks and an image-feature-fusing network consisting of a simple three layers neural network. Fig.8 shows the results of network's cumulative matching score by using manhattan distance. This figure shows the differences between the sub-areas that selected by human subjects and the network system. Though there remains some differences, most of the network outputs were acceptable for human observers.

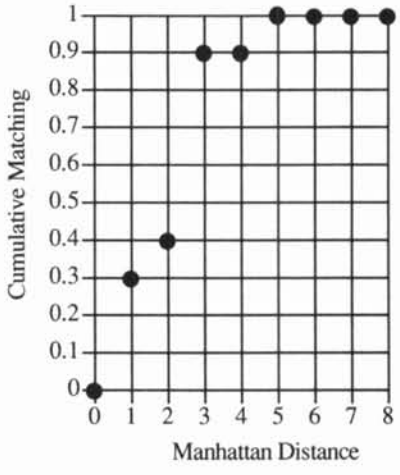


Fig.8 Cumulative Matching Score

We also suggested the network which can separate the image sequences temporally.

The video editting is a time consuming and hard work for the editting person because it needs to see all of its contents before editting. Under the aid of those network systems mentioned in the above, on-screen editting may be easily realized. In the near future, computation cost will became more reasonable and computation power will be enough to calcurate neural network system with personally available computers. So, it will be possible to use WS or PC for editting video data easily.

ACKNOWLEDGMENTS

We thank to Mr.Maekawa and Mr.Sogou of MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD., for offering their analogue-digital converted image data and technical assistance.

REFERENCES

[1] Shioiri S., Ikeda M. and Uchikawa K., *Visual and tactile pattern perception.*, Conference digest of the 13th Congress of the International Commission for Optics., 74-75(1984)

[2] D.E.Rumelhart and J.L.McClelland, eds., *Parallel distributed processing: Exploration is the microstructure of cognition. Vol. 1, Foundations.* Cambridge, MA: Bradford Books/MIT Press.(1986)

[3] HARRY L.VAN TREES : "Detection, Estimation, and Moduration Theory", WILEY(1968)