EMBEDDABLE RECONFIGURABLE NEUROPROCESSORS

Taher Daud, Tuan Duong, Harry Langenbacher, Mua Tran, and Anil Thakoor
Center for Space Microelectronics Technology
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, California 91109

Timothy Brown
Bell Communications Research
Morristown, New Jersey, 07960

ABSTRACT

Reconfigurable and cascadable building block neural network chips, fabricated using analog VLSI design tools, are interfaced to a PC. The building block chip designs, the cascadability and the hardware-in-the-loop supervised learning aspects of these chips are described. We also report on the computation-intensive problem of map-data classification. For this application, a 486-embeddable neuro-processor card is highlighted that shows the promise of system-level speeds (including I/O) exceeding that of CD-ROM.

INTRODUCTION

Artificial neural network paradigms are derived from the biological nervous system and are characterized by dense interconnections of simple processing elements. These processing elements called neurons are typically analog, have a multitude of signal-modulating weighted links from other similar neurons and perform a signal summation function followed by a nonlinear thresholding operation to the myriad of incoming signals. Such a structure provides massive parallelism in its information processing function, and is known to code and store intelligent information in a highly distributed manner via the weights of the interconnecting links[1].

To harness full power of neural net's parallelism for obtaining high-speed solutions to computationally intensive problems, these parallel architectures have been implemented both in digital and analog hardware. In general, analog implementations offer low power consumption and compactness, whereas digital implementations provide better precision[2-4].

JPL's approach of analog "building-block" chips is motivated by the requirements of deployable low power data processing systems. Design and fabrication of such chips that can be cascaded and reconfigured to cater to different architectures, sizes, and resolutions of synaptic weights in the analog domain has offered not only a flexible research tool, but also the ability to seek selected computation-intensive applications where conventional techniques are too slow or merely not capable.

We describe VLSI chip-implementations of synaptic matrices and arrays of neurons, cascadable to larger networks and reconfigurable with full connectivity[5]. Using these chips, embeddable 386/486 PC-compatible cards have been implemented. A feature classification application with map-data requiring a feedforward architecture and supervised training is presented here. The same card has also been reconfigured as a feedback net to perform such optimization tasks as resource allocation[4].

NEURAL NETWORK HARDWARE

Hardware implementation of neural nets involves design of two elements: synapses that arc the variable weight (conductance) links, and neurons that are the non-linear elements performing thresholding operations. We describe the VLSI implementation where the weights are stored on-chip using digital logic and incorporating the digital to analog converters (DACs) to obtain analog weights [6].

In effect, a synapse performs the function of multiplication of an input (current) with its stored weight, while the neuron maps the sum of several input current signals via a non-linear sigmoidal transfer function as a voltage output.

Synapse Chip Design

Our synapse chip design is a 32x32 synaptic crossbar matrix with 32 input and 32 output lines. A global voltage-to-current conversion as current input to synapses is provided for each row. At each node of the matrix is incorporated a 7-bit synapse. The basic synapse circuit consists of two functional blocks: (i) a 6-bit digital-to-analog conversion (DAC) with digital latches and associated current mirrors; and (ii) a current steering block for the sign bit. Additionally, ecoders for row/column select and address/data lines are included for random accessibility and programmability.

To increase the number of inputs and/or outputs beyond 32, additional chips with respective inputs and outputs provide a cascaded larger net. In addition, the resolution of the synapses can be increased by cascading in the third dimension. The relative strength of a synapse within a chip is determined by a global reference signal which provides the current-mirrored signals to individual synapses. By piggy-backing an additional chip with lower-valued reference signal (1/64th of its primary reference value) -- in effect paralleling each bit weight of the synapse in the array with all the weight bits of the respective synapse in the piggy-backed chip to act as a vernier-- the weight values have been resolved beyond the inherent 7 bits of accuracy to a nominal 13 bits (12 bits plus sign), with at least 11 most significant bits providing a linear and monotonic response.

Synapse-Neuron Composite Chip Design

A synapse-neuron composite architecture has 32 neurons replacing synapses along one of the diagonals in the synaptic matrix. The input to the neuron is a current signal that is sum of the current outputs by the connected synapses. The neuron input-output characteristics are a set of sigmoids with variable slope. Each neuron output feeds back to the respective input, thereby providing a fully feedback connected network, and by judicious programming of the synapses, a feedforward architecture on the chip is obtained. The synapse-neuron chips, as also the all-synapse chips, were fabricated in VLSI employing 2-μm feature geometry.

With the availability of these composite chips, two of them were cascaded with two synapse chips to act as a 64x64 matrix of 7-bit-resolution fully connected synaptic network with 64, wide-range, variable gain neurons. In addition, by cascading four additional synapse chips, thereby paralleling synapses of the respective two chips as shown schematically in Figure 1, and adjusting the chip gain levels accordingly, the effective dynamic range of weights was increased to 11 bits[5]. This circuit was characterized to obtain the synapse bit-resolution data in combination with a neuron transfer function. A set of typical curves shown in Figure 2 highlight the monotonicity of the input-output relationship along with the variable gain feature of the neuron design. Thus, a total of eight chips cascaded together in x-y-z plane formed a 64-neuron fully interconnected array with over 4000 synapses, each with an effective ≈11-bit resolution. This is the *first ever* embodiment with the synapse resolution of ≈11 bits, permitting hardware-in-the-loop learning.

Plug-in Card Development

Based on the development of building block neuron-synapse chips, a neuroprocessor card was designed and assembled with the required interface circuitry for direct insertion into a X86-compatible back plane (Figure 3). The neuroprocessor card, while affording great control and flexibility, only begins to tap the power of the neural network chips with their parallel processing and high speed capabilities. With analog processing, the interface with the host digital machine becomes an intricate challenge to design. To reduce the 1/0 bus processing time, a high-speed 2 MHz analog to digital converter that plugs directly into the PC-bus was incorporated.

LEARNING ALGORITHM

The Cascade Backpropagation algorithm that permits the starting perception architecture to evolve by allocating hidden neurons as needed is similar to the Cascade Correlation algorithm[6], and performs gradient-descent learning. The pseudo-inverse computation is used to directly calculate the perceptron weights based on input and output patterns (with initial compensation for the nonlinear sigmoidal function). A hidden neuron is then connected via synapses from inputs to outputs. Back propagation (gradient-descent) learning is used to set both the perception (and bias) and the newly added weights. Neurons are added as hidden units onc at a time to learn the required input to output mapping[7].

A key calculation for hardware-in-the-loop learning operation is the slope of the transfer characteristics of the neurons at their respective operating points. This was achieved by perturbing the bias weights by small amounts which provided the required change in the neuron operating points and hence allowed the calculation of the derivative to be performed. With the outputs of all neurons (input, hidden, and output) and their respective derivatives known, and the differences of actual and target outputs dctcrmincd, the weight change in hardware is effected through the software.

The iterative process repeats until the learning saturates (no change in output) or a limit, by way of a predetermined maximum number of iterations, is reached. The learning process is ended when the desired degree of tolerance between target and the network output is reached. The process of learning uscs 11 bits of synapse precision as available, and even though the weight updates might occasionally be in error in

magnitude or even in sign, the stochastic nature of analog VLSI would eventually cause the non-monotonicities to be bridged[7].

RESULTS

Paper maps contain a massive amount of important data in an unwieldy format. To increase its utility, copious amounts of these data have been scanned into digital map knowledge base where each pixel data is a 3 color, 8-bit pcr color (24 bits/pixel) representation. However, the user is more interested in, say, display of roads, or rivers, etc., rather than the shades of colors. Therefore, a further processing involves extraction of a few (6 to 7) features. This step of feature extraction not only makes the maps more useful, but also compresses the data from 24 to, say, less than 3 bits, and puts it in a format that can be easily manipulated by the analyst as required. Normally, the process involves statistical methods requiring assumptions of Gaussian data spread along pixels with lengthy manipulations. We chose to classify the data using a feed forward neural network for its speed and especially because of its capability to generate optimal decision surfaces without a priori assumptions of any relationships (Gaussian or otherwise) except the availability of representative ground truth data. A key requirement of the task was to demonstrate the speed potential exceeding the CD-ROM rate (≈60,000 pixels/sec)[7].

The map consisted of a 305X200 pixel fragment. A grcy scale version of the original is shown in Figure 4a. Each pixel was to be classified into one of seven classes (roads, rivers, forests, contour lines, symbols/names, man-made features, and open areas). A training set (3800 pixels) was generated by an analyst. To enable a pixel to be classified within its local context, a 3x3 map-window was considered as input, yielding 27, 8-bit inputs for each pixel classification.

To test the processing speed in hardware, a neural network was trained in software, and then the learned weights were downloaded into the plug-in card. To compensate for the discrepancies between the hardware and the software model, an abbreviated learning algorithm was applied to just the neuron bias (threshold) connections. Hardware-in-the-loop training adapted these weights to the hardware in less than 4 seconds. The complete processing of 61,000 pixels including a graphics display was performed in about 7 seconds; the neuroprocessing time was just a fraction of a second (144,000 pixels/s). This implies an overall feed-forward processing rate of over 8,700 pixels pcr second. A grey scale rendition of the output data is shown in Figure 4b.

This clearly showed that 1/0 still dominated overall processing time. Current efforts arc focusing on a next generation plug-in card that — based on experience gained so far — will generate another order of magnitude speed-up in communications through the use of parallel digital-to-analog converters and direct memory accessed (DMA) data transfers. The enhancement will demonstrate the processing speeds exceeding those of a CD-ROM.

CONCLUSIONS

Neurally inspired architectures with their massive parallelism when implemented in hardware offer near real time processing for certain ill-defined and/or computation-intensive applications. JPL's embeddable and reconfigurable ncuroprocessors arc

unique because of their use of analog device implementations that provide compactness and low power, essential for deployable hardware. Hardware-in-the-loop learning, obtained as a result of innovative high resolution synaptic designs, is an added feature required for selected time critical applications for such areas as autonomous guidance, chemical process control, vehicle health monitoring, focal plane image processing, resource allocation/target assignment, and other avionics applications.

ACKNOWLEDGMENT

The research described in this paper was performed by the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology, and was jointly sponsored by the All Source Analysis Systems Program Office, the Advanced Research Projects Agency, the Ballistic Missile Defense Organization/Innovative Science and Technology Office, the Office of Naval Research, and the National Aeronautics and Space Administration.

REFERENCES

1. P.K, Simpson, "Foundations of Neural Networks," in Artificial Neural Networks: Paradigms, Applications, and Hardware Implementations, Eds.: E. Sanchez-Sinencio and C. Lau, IEEE Press, New York, 1992, pp. 3-24.
2. M. Holler, S. Tam, H. Castro, and R. Benson, "An Electrically Trainable Artificial Neural network (ETANN) with 10240 "Floating Gate" synapses, " Proceedings of the IEEE Int. Joint Conf. Neural Networks, vol. II, June 18-22, 1989, Washington, DC, pp. 191-196.
3. C. Mead, Analog VLSI and Neural Systems, Addison-Wesley, Reading, MA, 1989.
4. S.P. Eberhardt, R, Tawel, T.X Brown, T, Daud, and A.P. Thakoor, "Analog VLSI Neural Networks: Implementation Issues and Examples in Optimization and Supervised Learning," IEEE Trans. Indust. Electron. vol. 39, no. 6, pp. S52-564, Dec.)992.
5. 1. Duong, S.P. Eberhardt, M, 'l'ran, T. Daud, and A, I>. Thakoor, "1.earning and Optimization with Cascaded VLSI Neural network Building-Block Chips," Proceedings of the IEEE/INNS International Joint Conference on Neural Networks, June 7-11, 1992, Baltimore, MD, vol. I, pp. 184-189.
6. S.E. Fahlmann, C. Lebiere, "The cascade correlation learning architecture," in Advances in Neural Information Processing Systems II, Ed: D. Touretzky, Morgan Kaufmann, San Mateo, CA, 1990, pp. 524-532.
7. T.X Brown, M.D. Tran, 1. Duong, T. Daud, and A.P. Thakoor, "Cascaded VLSI Neural Network Chips: Hardware Learning for Pattern Recognition and Classification," Simulation, vol. 58, no. 5, pp. 340-346, 1992.

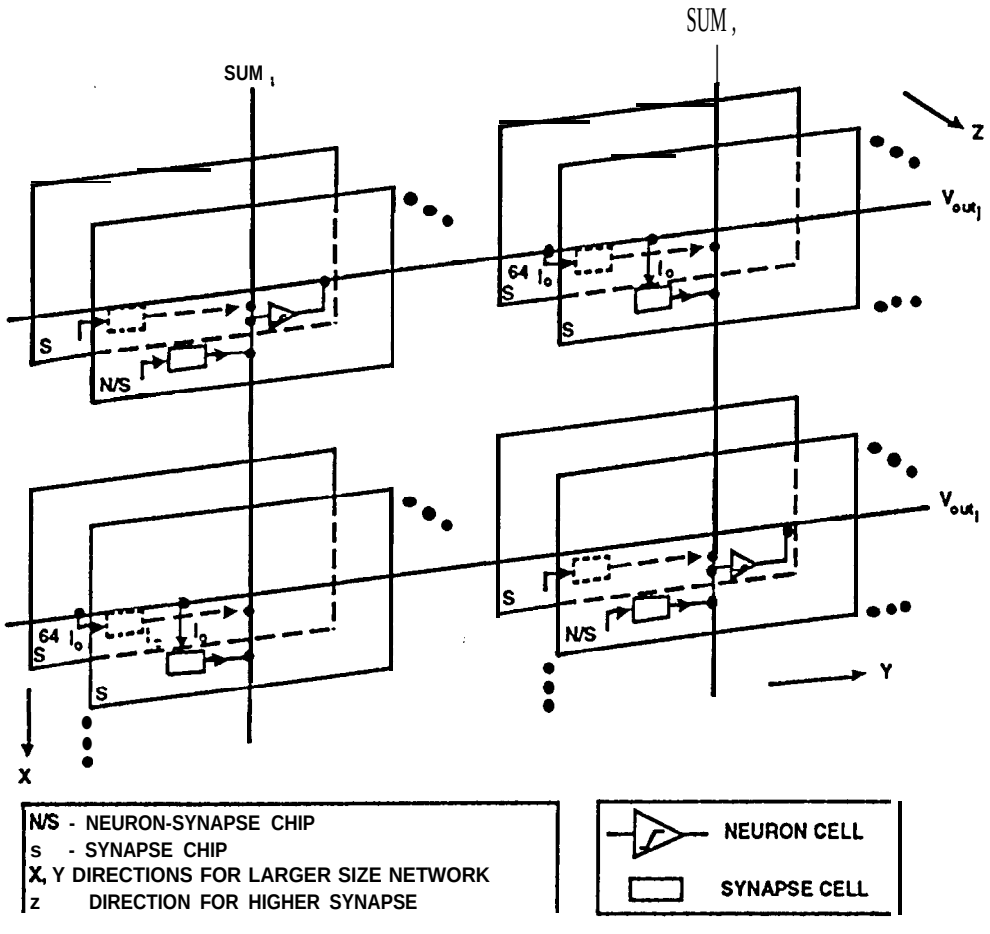

Figure 1. A schematic of an eight-chip, cascaded 64x64 neuron-synapse circuit with 13-bit resolution.

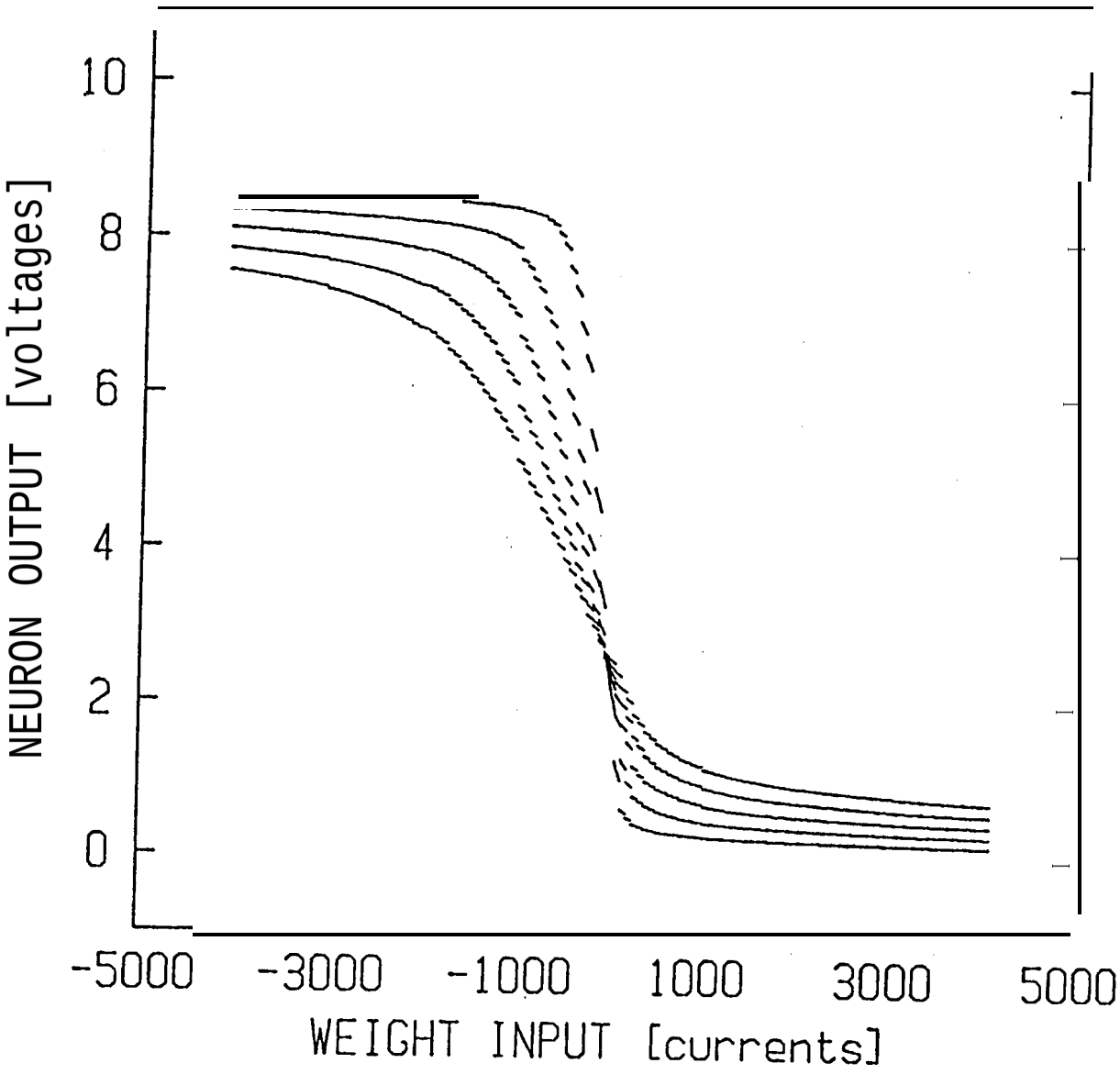

Figure 2. Transfer curves of the synapse -neuron, depicting the 13-bit resolution of synapses and variable gain, sigmoidal, & monotonic characteristics of neurons.

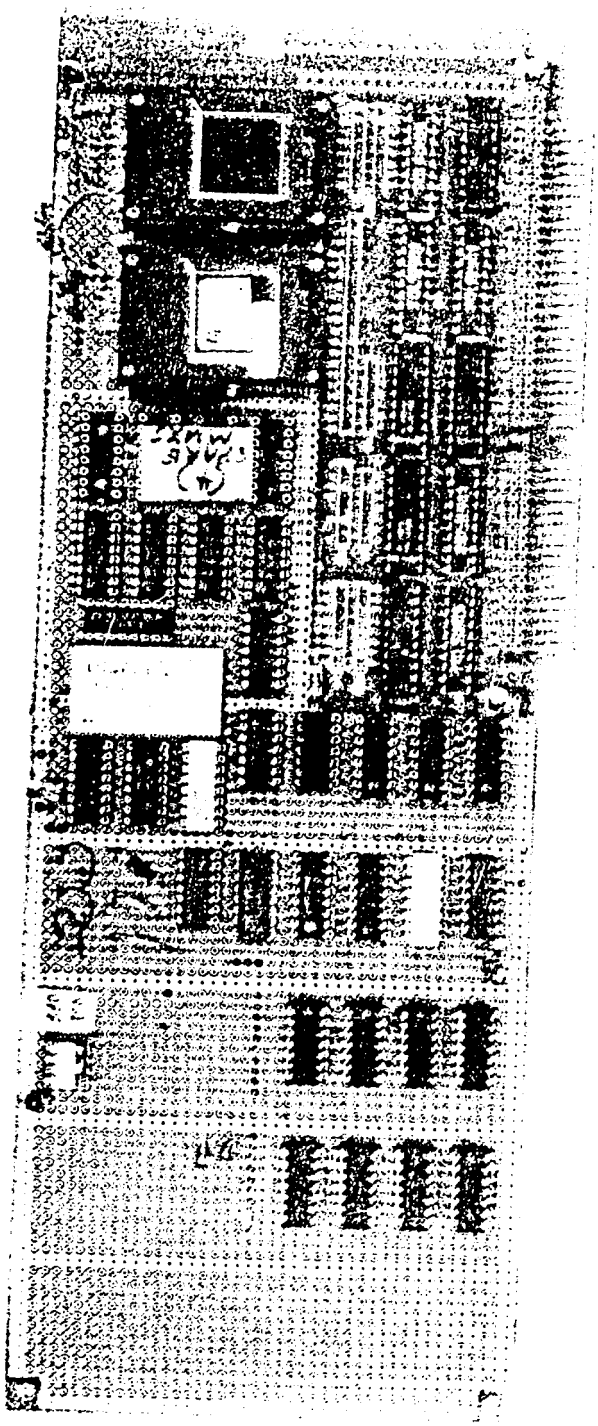

Figure 3. A X86 PC-compatible analog neuroprocessor-card.

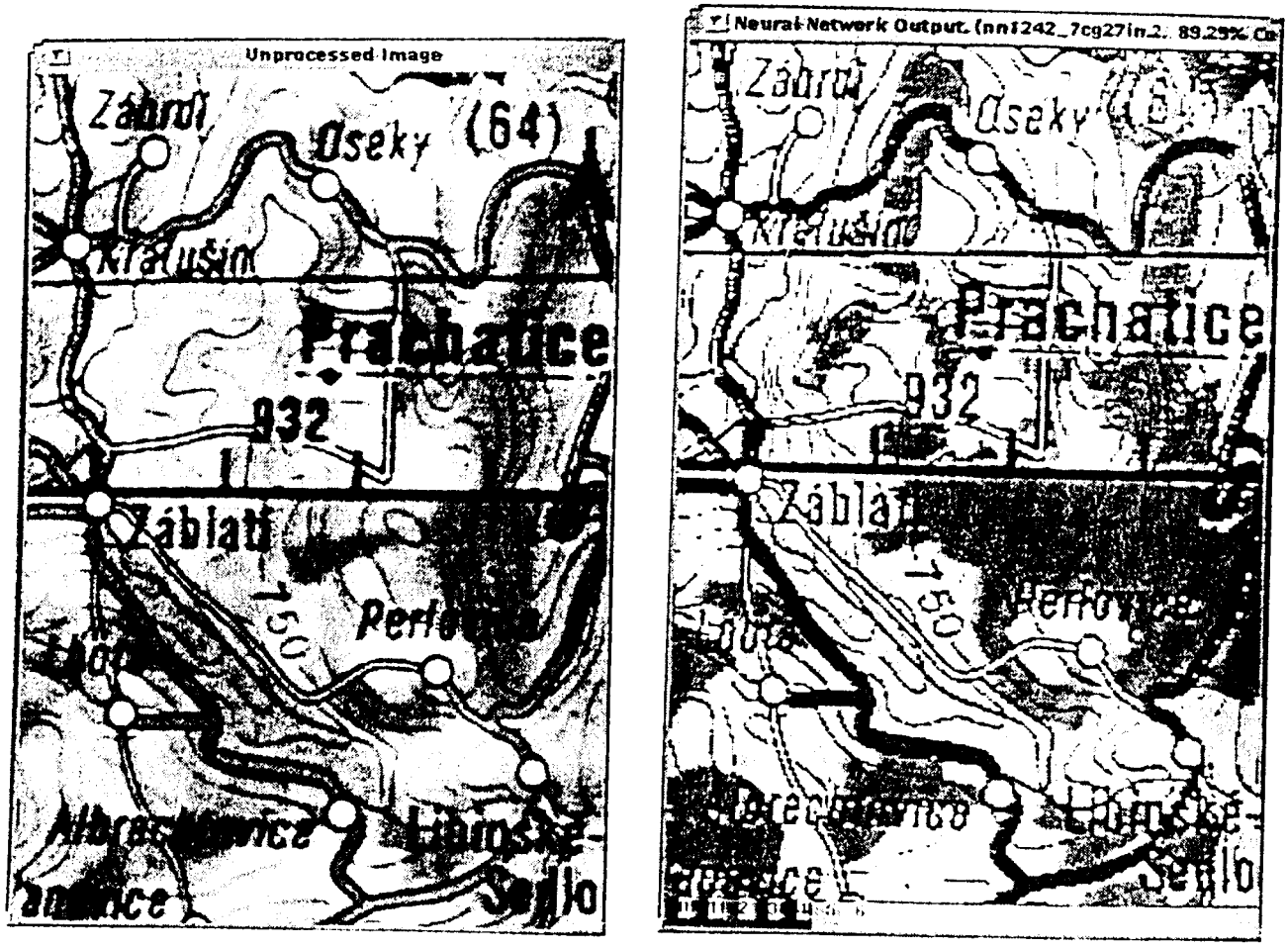

Figure 4a. A grey-scale rendition of a 305x200-pixel input map (3 colors/pixel and 8 bits/color) data.

Figure 4b. A grey-scale rendition of the hardware generated output with extracted seven features (one color/pixel)