Recognizing Handprinted Digit Strings: a Hybrid Connectionist/Procedural Approach*

Thomas Fontaine and Lokendra Shastri
Computer and Information Science Department
University of Pennsylvania
Philadelphia, PA 19104-6389

Abstract

We describe an alternative approach to handprinted word recognition using a hybrid of procedural and connectionist techniques. We utilize two connectionist components: one to concurrently make recognition and segmentation hypotheses, and another to perform refined recognition of segmented characters. Both networks are governed by a procedural controller which incorporates systematic domain knowledge and procedural algorithms to guide recognition.

We employ an approach wherein an image is processed over time by a spatiotemporal connectionist network. The scheme offers several attractive features including shift-invariance and retention of local spatial relationships along the dimension being temporalized, a reduction in the number of free parameters, and the ability to process arbitrarily long images.

Recognition results on a set of real-world isolated ZIP code digits are comparable to the best reported to date, with a 96.0% recognition rate and a rate of 99.0% when 9.5% of the images are rejected.

Introduction

Although a device capable of handprint recognition has many applications (Schürmann, J., 1982), the problem of recognizing handprint by machine remains largely unsolved. The failure to develop a working solution to the problem can best be attributed to the excess of variance inherent in unconstrained handprint, such as differences in writing styles between authors, handedness, and quality of print.

Traditional connectionist models for character and word recognition operate on static images (Le Cun, Y. *et al.*, 1990)(Keeler, J. D. *et al.*, 1991). We present an alternative approach in which an image is considered to be a time-varying signal and is presented to a system in a piece-wise fashion over time. The *temporalized* image is processed by a *spatiotemporal* connectionist network.

The approach offers several advantages. First, the model offers shift-invariance along the temporalized dimension. Second, the model inherently retains the local spatial relationships in the image along the temporalized dimension without having to learn them explicitly. Next, the model is architecturally less complex than a similar model using two spatial dimensions. Finally, the model is capable of processing arbitrarily long inputs along the temporal dimension.

In most approaches to word recognition, a segmentation step is utilized prior to character recognition in which a word image is decomposed into its component characters. Since handprint samples often contain characters which overlap or are disjoint, adequate segmentation is a difficult problem. In general, a word recognition system is faced with a dilemma: a word image cannot be properly segmented without first recognizing individual characters, but individual characters cannot be recognized prior to segmentation.

We have suggested an alternative approach to deal with the segmentation/recognition dilemma (Fontaine, T., 1991), and have focused on the problem of recognizing handprinted digit strings. We present a scheme in which segmentation and recognition is performed by a hybrid system comprised of two connectionist networks and a procedural controller. The first network, the Coarse Recognition Device (CRD), assimilates a word image in a left-to-right fashion over time, estimating segmentation boundaries between characters, while performing coarse character recognition. The second network, the Refined Recognition Device (RRD), is specialized for isolated character recognition, and attempts to classify portions of the image hypothesized to be characters by the CRD. Both networks are governed by a traditional procedural controller, capable of fusing signals emanating from the two networks while incorporating domain knowledge.

The spatiotemporal approach

Most character recognition schemes operate on *static* character images whereby an image is pre-

*This work was supported by grant MCS-83-05211 and ARO grants DAA29-84-9-0027 and DAAL03-89-C-0031. We thank Gary Herring and John Hull for the USPS database and John Geist for the NIST database.

sented to the system as a time-invariant signal. An alternative viewpoint is to consider an image to be a time-varying signal which is presented to the system in a piecewise fashion over time. For example, consider a matrix containing M rows and N columns of data. One could envisage a *column-wise* input of data in which a system receives all M pieces of data contained in column i of the image at time i. Thus, N time steps would be needed to assimilate the entire matrix of data into the system. It was suggested by Shastri (1989) that in some visual recognition domains there may be inherent advantages to consider images as being *spatiotemporal* signals, akin to other time-varying signals such as speech, and to process them using spatiotemporal connectionist models.

Advantages of the approach

Shift-Invariance. During the column-wise input of an image over time, output emanates from the output layer at each time step. If the network is optimized such that a column of zero inputs (white space) does not significantly change the state of the network, then the time-integrated network output is independent of the spatial position of the character along the temporalized axis. Thus shift-invariance along the temporalized axis falls out as a natural byproduct of the scheme.

Retention of local spatial relationships. Consider a unit in the first hidden layer of a traditional (static) network. The activations received by this unit from units in the input layer are unlabeled levels of activation, and hence, this unit cannot determine which inputs come from spatially neighboring pixels and which do not. As far as this hidden unit is concerned, the input it receives from an image I is indistinguishable from the input it receives from an image I' obtained by permuting I. Now consider a hidden unit in a spatiotemporal model. The inputs to such a unit from two adjacent pixels (along the temporalized dimension) become available to the unit in adjacent time steps. Consequently, the spatial structure of the input (along the temporalized dimension) is made explicit to the hidden unit.

Reduction in network complexity. In the spatiotemporal scheme, a spatial dimension is effectively exchanged for a temporal dimension, thereby decreasing the complexity of the network. During network training, the number of links in a network corresponds to the number of free parameters in an unconstrained nonlinear optimization. A substantive reduction in network complexity can dramatically decrease the dimensionality of the optimization.

Processing arbitrarily long inputs. Developing a word recognition system within a traditional feed-forward connectionist framework has proven difficult, although recent progress has been made by replicating and tessellating network substructures to accommodate images with multiple characters (Keeler, J. D. *et al.*, 1991). The ability to process arbitrarily long images is inherent in our approach, and offers an alternative means to process word images within a connectionist framework.

Spatiotemporal networks

Spatiotemporal data representation necessitates working within a framework capable of processing time-varying signals. The connectionist model we employed was inspired by the *Temporal Flow Model* (Watrous, R. and Shastri, L., 1986) which has achieved good results in speech recognition (Watrous, R., 1988).

In order to regain the spatial relationships after an image is temporalized, the signal needs to be respatialized by temporally integrating the signals from different time slices. Temporal integration in our networks is obtained in two ways: self-recurrent links on the units provide a record of the past activation of the unit, and multiple links between units, each with a distinct propagation delay, allow the receiving unit to compute a function over several time slices.

Training spatiotemporal networks

Spatiotemporal networks generate varying outputs over time, and hence a target output is necessary at each time step. This target sequence is known as the *target function*, and the development of a suitable set of target functions is a temporal credit assignment problem. Although the selection of target functions for a particular problem can be guided by domain knowledge and human intuition, selection is primarily experimental. The target functions used in the development of the RRD and CRD are described in the next section.

The word recognition system

Our system, depicted in Figure 1, is comprised of three components: the Refined Recognition Device (RRD), Coarse Recognition Device (CRD), and Procedural Controller (PC).

Refined recognition device (RRD)

The RRD is responsible for accurate recognition of handprinted digits. We have developed the RRD in a modular manner in order to incorporate domain knowledge, reduce the number of free parameters, and simplify network analysis.

The RRD is comprised of ten individually trained Single Digit Recognition Networks, each of which is responsible for the detection of a particular digit. Each Single Digit Recognition Network consists of four Single Scan Networks, each of which assimilates data from a different "scan" of the image. A Single

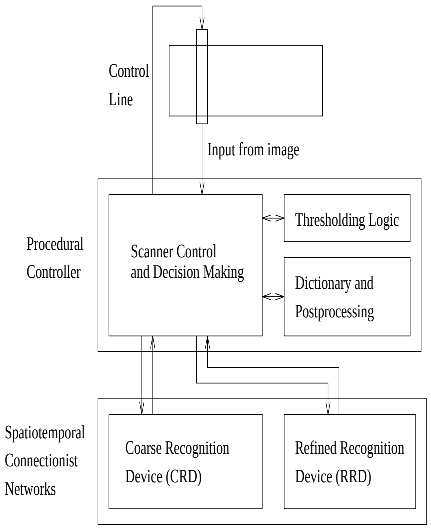

Figure 1: The Hybrid System. The image is processed by the CRD in a left-to-right fashion, monitored by the Procedural Controller. Once a confident signal arises (prescribed by Thresholding Logic), the most recently scanned portion of the image is sent to the RRD for verification. If verified, the classification is recorded and the networks reset. After scanning is complete, postprocessing algorithms make the final word classification decision.

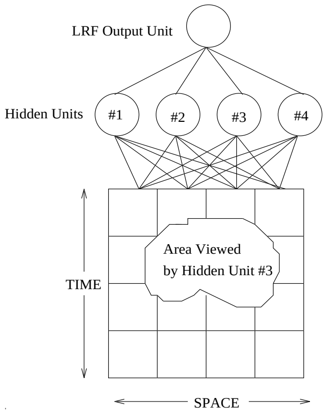

Figure 2: A generic Local Receptive Field (LRF). Hidden Unit n is able to view a spatial field of width 4 and a temporal window of width n, thereby respatializing the temporal signal. Unit #3, for example, views the first three rows and all columns of the image.

Scan Network is constructed from a number of adaptable layers, operating in conjunction with a number of pretrained Feature Detection Modules. A Feature Detection Module is formed by the replication and tessellation of a pretrained Local Receptive Field.

Feature detection modules. Many of the Arabic numerals can be approximately written using four simple stylus strokes: horizontal, vertical, slash, and backslash. The simplicity and recurrence of these strokes suggests the utility of developing pretrained feature detection modules, which can be integrated into a larger network. A separate *Local Receptive Field* module, or LRF, was pretrained to detect each of these four features over a localized area.

The generic LRF module is seen in Figure 2. It receives input over a spatial field of 4 inputs, a temporal field of 4 time steps, and consists of 4 input units, 4 hidden units, and a single output unit. Hidden unit n receives information from all input units, and utilizes n links from each input unit, with respective delays of $1, 2, \ldots, n$, creating a spatial window of width n into the temporal signal. As long as a feature to be detected by an LRF is present in its 4 by 4 receptive field, the LRF will emanate an output signal, albeit with a slight lag. Each specific LRF to detect horizontal, vertical, slash, and backslash strokes was trained using the same generic LRF module until a low error was achieved on a large set of strokes of various orientations.

One technique for extending local detectors to act upon larger fields of interest is to replicate and tessellate them (Le Cun, Y. *et al.*, 1990). We refer to a group of identical and tessellated LRFs as a *Feature Detection Module*, or FDM. Spatial information along one dimension is retained since an FDM is comprised of several spatially differing LRFs, while spatial information along the other dimension is encoded by the temporal sequence of LRF firings.

Single scan networks. Consider scanning an M by M image of an isolated digit using a left-to-right column-wise scan. Although useful discriminatory information may be present in the rightmost columns of the image, this information is not detected by the network until the final time steps. It may be useful to employ multiple scans in a variety of directions, where each scan feeds information into a separate group of input units.

The RRD employs four scans of the image: row-wise, column-wise, reverse row-wise, and reverse column-wise. Data from each scan is processed by separate, but topologically identical, network modules. Figure 3 illustrates the configuration of this module, referred to as a *Single Scan Network*. Two pretrained FDMs (a horizontal and slash stroke detector) are employed, along with several unstructured hidden layers.

Single digit recognition networks. Information from each scan is processed independently and concurrently by the four SSNs and the output of each SSN is passed to a single output unit. We refer to this complete network as a *Single Digit Recognition Network*.

Each Single Digit Recognition Network was trained to recognize a single digit class, and reject all others. Training samples were drawn from the USPS OAT Handwritten ZIP Code Database (1987)

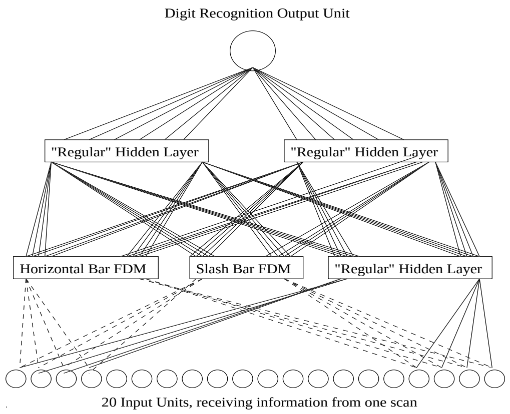

Figure 3: A Single Scan Network (SSN) Module. The input units pass information along links which are either frozen, if they are part of a pretrained FDM (dashed lines), or trainable, if they are "regular" links (solid lines). A local hierarchical structure is used to detect higher order features as information propagates towards the output unit.

and the NIST Special Database 3. The training set for each Single Digit Recognition Network consisted of approximately 2000 positive examples of the digit to be recognized, 125 negative examples of each other digit, and 900 negative examples of images containing partial or multiple digits. Images were preprocessed via smoothing, deskewing, scaling, and skeletonization (Fontaine, T. and Shastri L., 1992b).

As described in the previous section, a temporal target function is required to train spatiotemporal networks. A constant target of 0.05 was chosen for negative examples, while the target function for positive examples was sigmoidal, rising from a target of 0.05 at the onset of the digit to 0.95 at the end of the digit. Networks were trained using the BFGS algorithm (Fletcher, R., 1980), until a predetermined error on the training set was achieved.

After each Single Digit Recognition Network was trained to recognize its respective digit, all networks were combined to produce the RRD, capable of recognizing all ten digits. For complete details, see (Fontaine, T. and Shastri L., 1992a).

Coarse recognition device (CRD)

Construction of the CRD was similar to the RRD, but only one Single Scan Network was used, receiving information from a left-to-right columnar scan. The same target function used to train the RRD was used for the CRD, although training examples were padded on the left and right with white space in order to enforce shift-invariance, and the target function amended accordingly.

The training set consisted of approximately 2025 positive examples of the digit to be recognized and 225 negative examples of each other digit, and optimization was terminated at a predetermined error.

Procedural controller (PC)

In many problems, there is abundant domain knowledge which can best be utilized by traditional procedural techniques. We therefore employ a Procedural Controller, responsible for system control.

In our basic scheme, the PC monitors the output of the CRD (which assimilates a word image in a left-to-right fashion) and waits for the CRD to build up confidence in classification hypotheses. When a threshold is met, the PC sends the most recently scanned portion of the image to the RRD for verification. If the RRD accepts the estimate, the digit is recognized, the CRD is reset, and the process continues. If the RRD rejects the estimate, however, the CRD must either continue processing, or backtrack. For example, if a continued scan increases confidence in the current hypothesis, it can again be sent to the RRD for verification. If a continued scan decreases confidence, then thresholds can be altered to be less pessimistic and a portion of the image rescanned, if desired.

Control and classification decisions are made by procedural algorithms driven in part by signals arising from the connectionist networks, and in part by domain-specific knowledge and algorithms. In ZIP code recognition, for example, a given image will usually contain either 5 or 9 digits, constraining the recognition problem. In many domains, only a subset of all possible strings are legal, in which case a dictionary of legal strings can guide recognition and aid in classification (Shingal, R. and Toussaint, G. T., 1979).

Fusing the components

Fusing the components, fast and robust connectionist networks perform recognition under the guidance of a procedural component capable of incorporating domain knowledge. The hybrid system, shown in Figure 1, assumes a dictionary of legal strings and suitable postprocessing routines. The "Thresholding Logic" box represents the algorithms to accept or deny the decisions of the CRD, and to control the scanner positioning. A primitive thresholding method is presented in the next section. In future work, we envisage the derivation of thresholding logic by statistical analysis of the output of the CRD on a large training set.

Results

One must be cautious when comparing recognition results achieved using different databases. Ideally, a test database should be widely available, voluminous, and contain samples printed by a diversity of authors, unaware that their printing may be used to test a recognition device. The USPS OAT Handwrit-

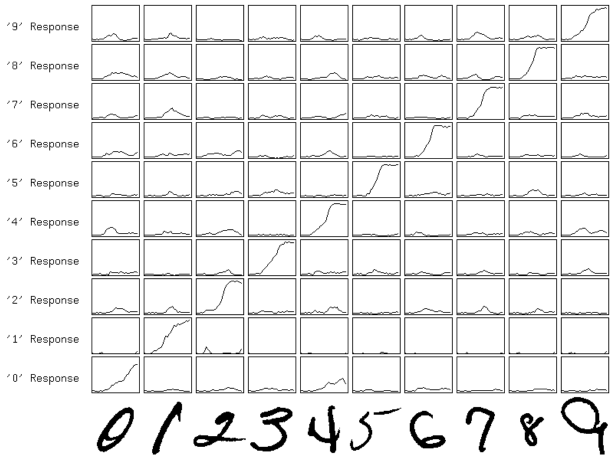

Figure 4: Digit recognition system output unit response, over time, to a typical set of ZIP code digit images.

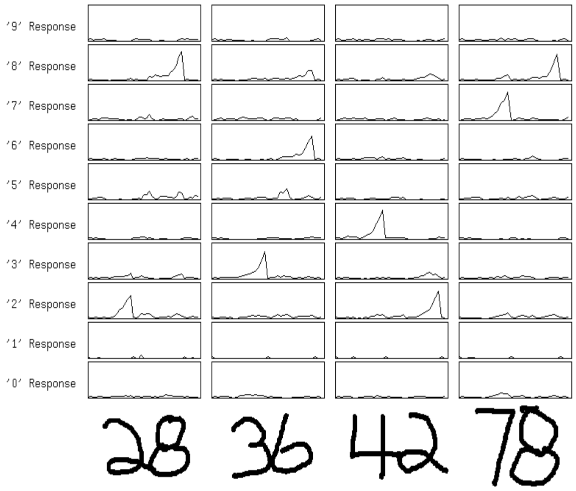

Figure 5: Output unit response of the Coarse Recognition Device in response to a set of images depicting touching or overlapping pairs of digits. Sharp peaks in response correspond to recognition of a digit and subsequent resetting of the CRD.

ten ZIP Code Database (1987) is such a database of unconstrained handprinted samples. A set of 2700 digits from the database not used in training was reserved for testing.

Refined recognition device (RRD)

To test the RRD, a winner-take-all scheme was chosen in which classification was performed by choosing the class corresponding to the output unit which generated the highest integrated activation. Figure 4 depicts the output of our system in response to a set of ten digit images. The plot shows the output unit response, over time, upon assimilation of the images. On the test set of 2700 images, an accuracy of 96.0% was obtained with no rejections. When 9.5% of the images were rejected, the error rate was decreased to 1% on the remaining images.

Coarse recognition device (CRD)

On the 2700 images in the USPS test set, the pilot CRD achieved a 94.5% recognition rate, using the same classification decision as the RRD. After testing on the USPS dataset, we informally tested the CRD on a small set of handprinted digit pairs to determine the potential of the CRD to provide adequate recognition and segmentation hypotheses. We chose a crude thresholding logic: if the activation of the output unit designed to recognize digit i exceeded half its maximum output, the portion of the image was accepted as belonging to digit class i and all activity in the network was reset to zero. Figure 5 depicts the output of the pilot CRD on a set of touching and overlapping digit pairs. Stiff peaks correspond to the acceptance of a digit and the subsequent resetting of the CRD. Initial results suggested that even without the PC, RRD, or sophisticated thresholding logic, the scheme possessed the potential to process digit strings.

Pilot system

To test the capability of our scheme to segment and recognize overlapping and touching digits, we implemented a primitive Procedural Controller. The PC monitored the CRD outputs, and permitted a hypothesis if a CRD output unit had achieved half its maximum value. If the RRD rejected the hypothesis, scanning continued. If the RRD accepted the hypothesis, the hypothesized segmentation boundary expanded until confidence in the RRD classification decreased. At the point of acceptance, network activity was reset, and scanning continued.

Since pairs of digits which touch, or whose fields overlap, are not readily available, test data was synthesized from the isolated ZIP code digit images. We tested the system on six separate data sets, each comprised of 500 images. The sets differed depending on whether the digits were drawn from the training or testing set, and how much their fields overlapped (as a percentage of the width of the first digit in the pair). Table 1 shows recognition results. The percentage of each set containing digit pairs which touch is significant, since many traditional segmenters cannot cope with such samples. The reject column shows the percent of the images rejected by the system. In our pilot implementation, an image was rejected if either the CRD did not generate a hypothesis, or if the CRD and RRD could not agree on a hypothesis. The accuracy column shows the accuracy of the system on the images which were not rejected. Although domain knowledge, such as

Table 1: Digit Pair Recognition Results

Set	Type	Overlap	Touch	Reject	Accuracy
1	Train	0	9.6	10.6	88.8
2	Train	5	46.6	7.0	74.8
3	Train	10	63.8	6.8	68.7
4	Test	0	10.2	10.8	89.9
5	Test	5	45.6	6.8	75.5
6	Test	10	59.8	6.4	67.3

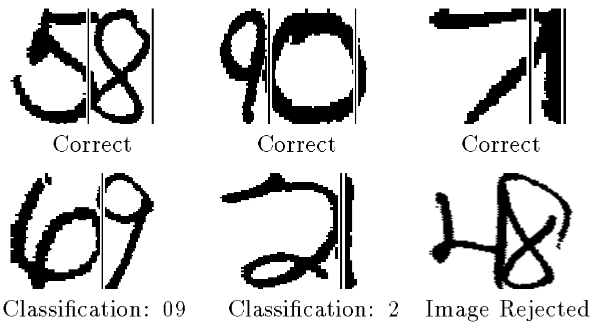

Figure 6: Examples of correctly and incorrectly classified or rejected pairs.

the expected number of digits present in an image, can significantly augment recognition, no such information was used. Examples of correctly and incorrectly classified pairs, drawn from Set 6, are shown in Figure 6 (vertical lines indicate points of segmentation).

Future work

The direction for future work is well defined based on these results. The CRD needs to be refined, and thresholding parameters derived based on statistical analysis of the output of the CRD on a training set. A confidence model for recognition, using the joint unit responses of the CRD and RRD, needs to be developed. Domain knowledge and algorithms should be incorporated, since domain knowledge can have a considerable impact on recognition. We are currently investigating ZIP code recognition, an application rich in domain knowledge.

Concluding remarks

We have presented an alternative approach to word recognition in which word images are viewed as time-varying signals and are processed by fast and robust spatiotemporal connectionist networks under the guidance of a traditional procedural controller.

Connectionist networks are elastic, offer fast recognition after training, and can be implemented on a single microchip. By employing spatiotemporal networks to process temporalized images, shift-invariance is gained, local spatial relationships are retained, a reduction in network complexity over static connectionist networks is achieved, and arbitrarily long images can be processed.

Recognition results of the Refined Recognition Device are comparable to the best reported to date on a difficult set of real-world handprinted digits. Initial CRD results demonstrate the power of the model to cope with strings of touching and overlapping digit strings, and initial results on touching pairs of digits are encouraging. We feel that our alternative approach to word recognition provides several distinct advantages, and offers interesting areas for future work.

References

Fletcher, R. *Practical Methods of Optimization.* John Wiley and Sons, 1980.

Fontaine, T. and Shastri L. Character recognition using a modular spatiotemporal connectionist model. Technical Report MS-CIS-92-24, University of Pennsylvania, March 1992a.

Fontaine, T. and Shastri L. Handprinted digit recognition using spatiotemporal connectionist models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 169–175, 1992b.

Fontaine, T. *A Hybrid Procedural-Connectionist Word Recognition System.* Thesis Proposal, University of Pennsylvania, 1991.

Keeler, J. D., Rumelhart, D. E., and Leow, W. K. Integrated segmentation and recognition of hand-printed numerals. In Lippman, Moody, and Touretzky, editors, *Advances in Neural Information Processing Systems*, volume 3, pages 557–563. Morgan Kaufmann, 1991.

Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 396–404. Morgan Kaufmann, 1990.

Schürmann, J. Reading machines. In *Proceedings of the International Conference on Pattern Recognition*, pages 1031–1044, 1982.

Shingal, R. and Toussaint, G. T. Experiments in text recognition with the modified Viterbi algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:184–193, 1979.

Watrous, R. and Shastri, L. Learning phonetic features using connectionist networks: An experiment in speech recognition. Technical Report MS-CIS-86-78, University of Pennsylvania, 1986.

Watrous, R. *Speech Recognition Using Connectionist Networks.* PhD thesis, University of Pennsylvania, 1988.