A Convolutional Neural Network Hand Tracker

Steven J. Nowlan
Synaptics, Inc.
2698 Orchard Parkway
San Jose, CA 95134
nowlan@synaptics.com

John C. Platt
Synaptics, Inc.
2698 Orchard Parkway
San Jose, CA 95134
platt@synaptics.com

Abstract

We describe a system that can track a hand in a sequence of video frames and recognize hand gestures in a user-independent manner. The system locates the hand in each video frame and determines if the hand is open or closed. The tracking system is able to track the hand to within ± 10 pixels of its correct location in 99.7% of the frames from a test set containing video sequences from 18 different individuals captured in 18 different room environments. The gesture recognition network correctly determines if the hand being tracked is open or closed in 99.1% of the frames in this test set. The system has been designed to operate in real time with existing hardware.

1 Introduction

We describe an image processing system that uses convolutional neural networks to locate the position of a (moving) hand in a video frame, and to track the position of this hand across a sequence of video frames. In addition, for each frame, the system determines if the hand is currently open or closed. The input to the system is a sequence of black and white, 320 by 240 pixel digitized video frames. We designed the system to operate in a user-independent manner, using video frames from indoor scenes with natural clutter and variable lighting conditions. For ease of hardware implementation, we have restricted the system to use only convolutional networks and simple image filtering operations, such as smoothing and frame differencing.

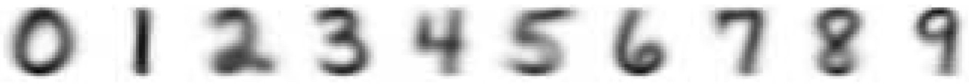

Figure 1: Average over all examples of each of the 10 classes of handwritten digits, after first aligning all of the examples in each class before averaging.

Our motivation for investigating the hand tracking problem was to explore the limits of recognition capability for convolutional networks. The structure of convolutional networks makes them naturally good at dealing with translation invariance, and with coarse representations at the upper layers, they are also capable of dealing with some degree of size variation. Convolutional networks have been successfully applied to machine print OCR (Platt *et al*, 1992), machine print address block location (Wolf and Platt, 1994), and hand printed OCR (Le Cun *et al*, 1990; Martin and Rashid, 1992). In each of these problems, convolutional networks perform very well on simultaneously segmenting and recognizing two-dimensional objects.

In these problems, segmentation is often the most difficult step, and once accomplished the classification is simplified. This can be illustrated by examining the average of all of the examples for each class after alignment and scaling. For the case of hand-printed OCR (see Fig. 1), we can see that the average of all of the examples is quite representative of each class, suggesting that the classes are quite compact, once the issue of translation invariance has been dealt with. This compactness makes nearest neighbor and non-linear template matching classifiers reasonable candidates for good performance.

If you perform the same trick of aligning and averaging the open and closed hands from our training database of video sequences, you will see a quite different result (Fig. 2). The extreme variability in hand orientations in both the open and closed cases means that the class averages, even after alignment, are only weakly characteristic of the classes of open and closed hands. This lack of clean structure in the class average images suggested that hand tracking is a challenging recognition problem. This paper examines whether convolutional networks are extendable to hand tracking, and hence possibly to other problems where classification remains difficult even after segmentation and alignment.

2 System Architecture

The overall architecture of the system is shown in Fig. 3. There are separate hand tracking and gesture recognition subsystems. For the hand tracking subsystem, each video frame is first sub-sampled and then the previous video frame (stored) is subtracted from the current video frame to produce a difference frame. These difference frames provide a crude velocity signal to the system, since the largest signals in the difference frames tend to occur near objects that are moving (Fig. 5). Independent predictions of hand locations are made by separate convolutional networks, which look at either the intensity frame or the difference frame. A voting scheme then combines the predictions from the intensity and difference networks along with predictions based on the hand trajectory computed from 3 previous frames.

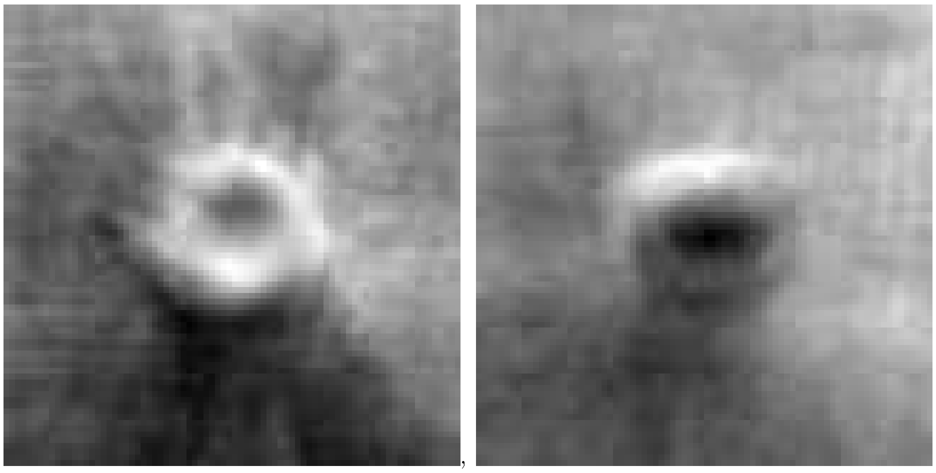

Figure 2: Average over all examples of open and closed hands from the database of training video sequences, after first aligning all of the examples in each class before averaging.

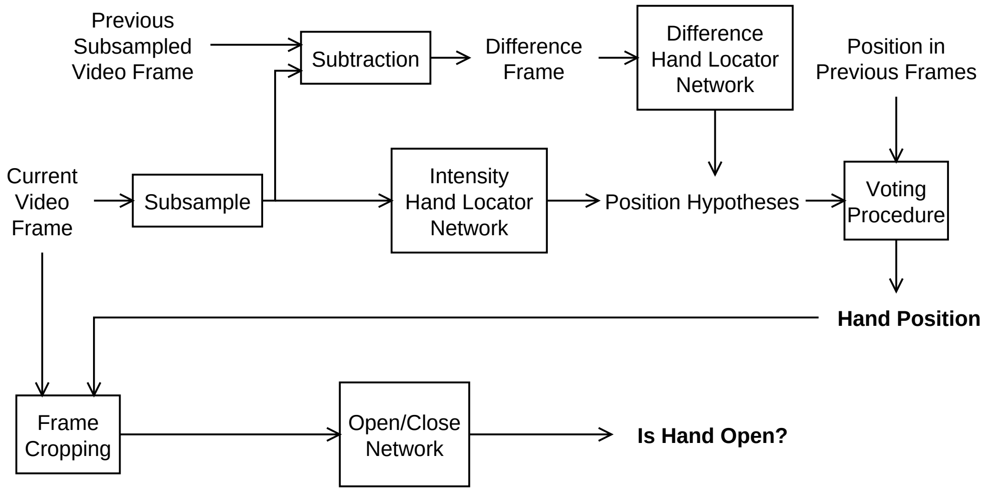

Figure 3: Architecture of object recognition system for hand tracking and open-versus-closed hand identification.

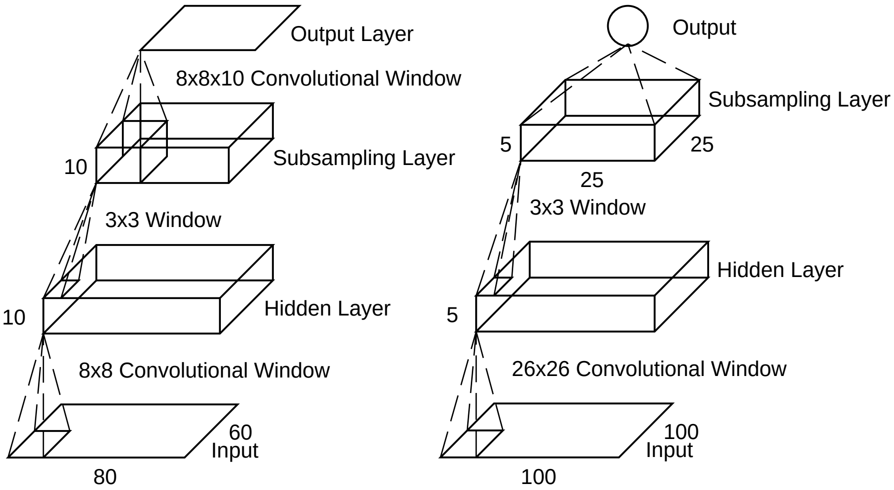

Figure 4: Network architectures. The network on the left will find a hand in either an intensity or difference low-resolution image. The output of the left network is a single convolutional image for encoding the presence of the hand at a location. The network on the right recognizes whether the hand is open or closed in a full resolution image. The output of the right network is a single value that identifies whether the hand is open or closed.

The gesture recognition subsystem takes a 100 by 100 pixel full resolution subimage of the current video frame as input. The position of this subimage is centered around the location output by the hand-tracking subsystem, thus the performance of the gesture recognition subsytem is dependent on the success of the tracking subsystem. A single convolutional network looks at this intensity subimage and indicates if the hand contained in the image is open or closed.

The tracking and gesture recognition subsystems were trained separately, and will be described in detail in the following subsections.

2.1 Hand Tracking

The intensity pathway of the hand tracking subsystem takes the original video images subsampled to 1/4 their original resolution (Fig. 5). The network architecture used to locate the hand position is a fully convolutional network with one hidden layer, a subsampling layer, and an output layer (Fig. 4). The output layer contains a single convolutional unit, which when active above threshold indicates the (possible) presence of the hand at that location.

The difference image pathway of the hand tracking system subtracts the previous subsampled video image from the current subsampled video image, and uses these 1/4 resolution difference images as input (Fig. 5). The network used to locate the hand location in the difference pathway is identical in structure to that used for hand location in the intensity pathway (Fig. 4).

The intensity and difference hand location networks were trained independently to

Figure 5: Sample video frames used by the hand tracking subsystem. The upper frames are used by the intensity pathway. The lower frames are corresponding difference images used by the difference pathway. The white cross indicates the position of the hand predicted by the network.

locate the position of the hand in each training frame to within ± 10 pixels (at full resolution). The details of training and the performance of the trained networks is discussed in the following section (see Table 1). In order to provide adequate overall tracking performance, it was necessary to combine information from both hand location networks, using a voting scheme. This voting scheme also takes advantage of the fact that we are attempting to track a smoothly moving object, by using information from the estimated position in previous frames to predict the position of the hand in the current frame.

A simple rule based scheme is used to combine the predictions of hand location from the different sources. We first predict the current position of the hand based on a trajectory computed from 3 previous frames, and construct a plausible bounding box centered at this position. This bounding box represents about one third of the original image. We next find all above threshold network responses from both the intensity and difference networks. If there is a strong response from both of the networks in a similar spatial location, we choose that location. Otherwise, we use the location of the strongest response from the difference network. If there are no above threshold responses from the difference network, we use the response from the intensity network. If there are no above threshold responses from either network, we use the location predicted by the trajectory from previous frames. All thresholds used in this voting scheme are estimated from the training/cross-validation set.

2.2 Recognition of Open versus Closed Hands

The gesture recognition subsystem takes as input a 100 pixel by 100 pixel piece of the original video frame, centered at the location output by the hand tracking subsystem (and usually not centered on the hand itself). This image is only about

50 percent larger than the largest hands in our image database, so the gesture recognition subsystem is dependent on a high quality hand tracking system. The 100 by 100 pixel size chosen for input to the gesture recognition system allows positional errors of up to 25 pixels while still maintaining most of the hand in the input image. The largest positional error made by the hand location subsystem was 11 pixels, well within the tolerance of the gesture recognition subsystem.

The network architecture used to identify if the hand is open or closed is similar to the networks used to track the hand position with a convolutional hidden layer, a subsampling layer, and an output layer (Fig. 4). The primary difference from the hand location networks is that the output layer is non-convolutional and fully connected to the outputs of the subsample layer. The single output unit looks at this entire image at once, and if this unit is active above threshold, it indicates that the hand is open; if the unit is below threshold, the hand is closed.

3 Training and Performance

Our simulations have been conducted on a database of 900 video images from 18 different subjects. These images were captured from a video input in real time, at the rate of 10 frames per second with a resolution of 320 by 240 pixels per image. Each sequence represents a sample of 10 seconds of continuous motion from each subject. Subjects were requested to move one hand about freely, opening and closing the hand as it moved. Video of subjects was taken using a hand-held camcorder under natural lighting conditions in a variety of different rooms containing complex clutter, windows, etc. (Fig. 5). This varied and complicated background greatly increases the difficulty of segmenting the hand from background clutter in many circumstances. Thirty of the frames from each subject were used for training/cross-validation purposes with the remaining nineteen frames reserved for testing. In addition, we obtained a sequence of blind test frames from an individual not part of the original training set, and in a different room environment than any of the original training data.

Both location networks were training using the ISR (Keeler *et al* 1991) training procedure. The gesture network was trained using online back-propagation.

Table 1 summarizes the performance of the two tracking networks individually, as well as the performance of the overall system, on the set of 342 test images. The intensity network alone locates the hand to within ± 10 pixels in 91.8% of the test frames. A large portion of the errors made by this network are due to confusions caused by complex structured backgrounds that can have color and texture very similar to the hand in the intensity image. Another common source of errors are portions of the arms and face, particularly when the hand is closed.

Much of the background complexity can be eliminated by looking at the difference images, which allows the difference network to outperform the intensity network. However, the difference network still has a fairly high error rate. These remaining errors are due to three factors: One cannot move the hand without also moving other parts of the arm and in many cases the head and torso. In addition, when the hand stops, as when reversing direction, it may disappear entirely from the difference frames. Finally, there are instances in our database in which large objects were

Table 1: Summary of test set performance for hand location.

Information Used	**Test Error Rate**
Intensity	8.2%
Difference	6.4%
Intensity + Difference	3.2%
Intensity + Difference + 3 Frames	0.3%

moving in the background of the scene.

The error rate is improved dramatically when the predictions from the intensity and difference networks are combined. This is a strong indication that the errors made by the intensity and difference networks are only weakly correlated, so a nearly maximal gain is obtained by combining their predictions. The most dramatic performance gain is obtained when the predictions of the network are combined with the trajectory information in the final voting procedure.

The gesture recognition network correctly identifies whether the hand is open or closed in 99.1% of the 342 test images. In contrast, a nearest neighbor classifier using the same training set and using a convolutional euclidean distance metric (*i.e.* the minimum euclidean distance between the training and test pattern allowing up to 10 pixels of misalignment in both x and y directions) could achieve only 43.2% correct classification. This empirically verifies that this is a difficult classification task even if misalignment is taken into account.

To provide a more rigorous test of the generalization ability of the hand tracking and gesture recognition system a second blind performance evaluation was performed, using a sequence of 50 frames from a subject who was not part of the original 18 subjects used for gathering the training and testing images. These frames were shot using a different video camera, and were shot in an open lab area, which was considerably different from the office environments most of the original training data was gathered in.

The hand tracking subsystem performed perfectly on this blind test set, correctly determining the position of the hand within ± 10 pixels in all 50 frames. The gesture recognition system correctly classified 94% of the test frames as open or closed, which corresponds to three errors out of 50. For these three frames, the hand was totally blended into the background, and a human observer could not determine from these individual frames whether the hand was open or closed.

4 Discussion and Conclusions

The system we have developed could be extended to recognize a broader range of hand gestures. One problem remaining to be solved is how to deal with relative motion signals in situations which involve moving background or camera.

The speed of the hand tracking and gesture recognition system is dominated by the time required to evaluate the two hand position networks and the gesture recognition network. The evaluation of these three networks requires approximately 25 million operations per frame, which would allow real time operation at video frame rates using existing convolutional hardware (Säckinger, 1992).

We have demonstrated that convolutional neural networks can be used to solve both the hand tracking and gesture recognition problems. The networks can find the hand in 99.7% of the test frames and recognize whether the hand was open or closed in 99.1% of the test frames. The system has demonstrated the ability to generalize to both novel users and novel indoor environments. In addition, the performance requirements of the system allow it to operate at video frame rates with existing hardware.

The high accuracy achieved on the hand tracking and gesture recognition tasks illustrates that convolutional networks can work on very general visual object recognition problems; problems where both segmentation and classification are difficult.

Acknowledgements

This work was carried out under an SBIR Phase I grant (contract N00014-93-C-0101) from the Office of Naval Research. We wish to thank Joel Davis of the ONR for useful suggestions and discussion.

References

J. D. Keeler, D. E. Rumelhart and W-K Leow. (1991) Integrated Segmentation and Recognition of Hand-Printed Numerals. In R. Lippmann, J. Moody, D. Touretzky (eds.), *Advances in Neural Information Processing Systems 3*, 557-563. San Mateo, CA: Morgan Kaufmann.

Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel. (1990) Handwritten Digit Recognition with a Back Propagation Network. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 2*, 396-404. San Mateo, CA: Morgan Kaufmann.

G. Martin, M. Rashid. (1992) Recognizing Overlapping Hand-Printed Characters by Centered-Object Integrated Segmentation and Recognition. In J. Moody, S. Hanson, R. Lippmann (eds.), *Advances in Neural Information Processing Systems, 4*, 504-511. San Mateo, CA: Morgan Kaufmann.

J. Platt, J. Decker, and J. LeMoncheck. (1992) Convolutional Neural Networks for the Combined Segmentation and Recognition of Machine Printed Characters, *USPS 5th Advanced Technology Conference,* **2**, 701-713.

E. Säckinger, B. Boser, J. Bromley, Y. LeCun, L. Jackel. (1992) Application of the ANNA neural network chip to high-speed character recognition, *IEEE Trans. Neural Networks,* **3**, (3), 498-505.

R. Wolf, J. Platt. (1994) Postal Address Block Location Using a Convolutional Locator Network. In J. Cowan, G. Tesauro, J. Alspector (eds.), *Advances in Neural Information Processing Systems 6*, 745-752. San Mateo, CA: Morgan Kaufmann.