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.

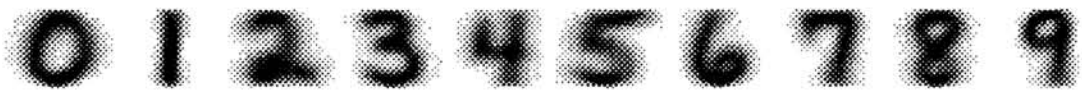

Figure 1: Average over all examples of each of the 10 classes of handwritten digits, after first aligning all of the examples in each class before averaging.

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.

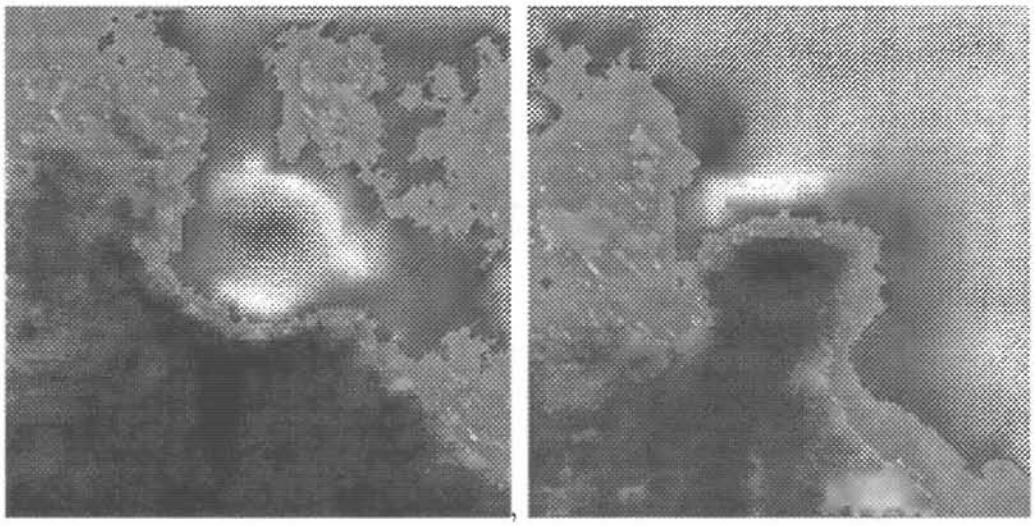

Figure 2: Average over all examples of open and closed hands from the database of training video sequences, after first aligning all of the examples in each class before averaging.

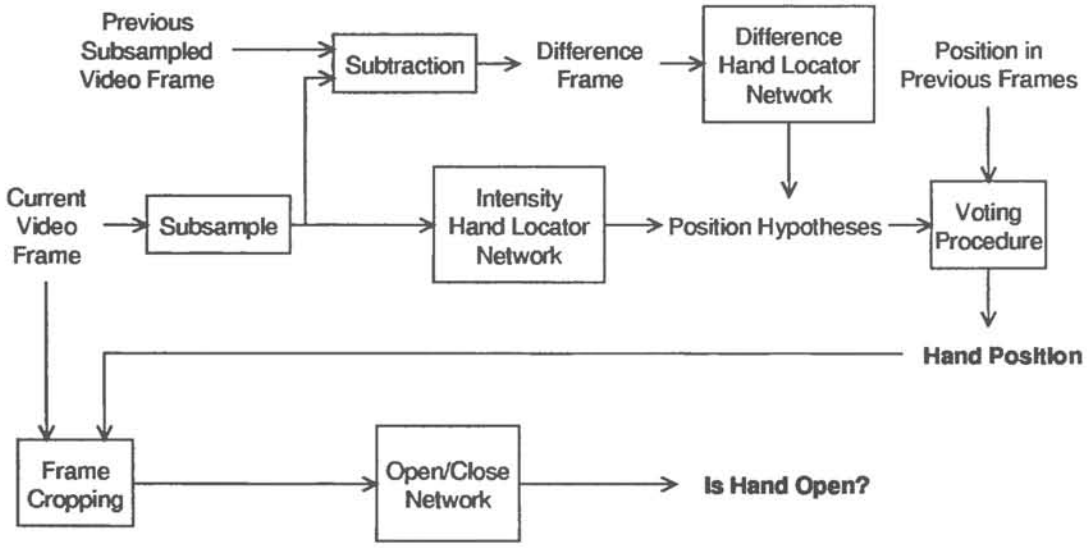

Figure 3: Architecture of object recognition system for hand tracking and open-versus-closed hand identification.