Mats Rudemo, tel 772 3575 or 772 3500 or 772 3593

Examination in Statistical Image Analysis, August 27, 2004

Course code Chalmers: TMS016, Gothenburg University: Statistisk Bildbehandling

Written examination August 27, 2004, 14.15-18.15 in house M.

Literature and notes may be brought for this written examination. All types of pocket calculators are allowed but not computers. In the written examination there are two pages and two problems. You are supposed to answer both problems, and in the judgement they have the same weight. Answers may be given in English or Swedish.

Problem 1.

Figure 1 shows a detail of a 2D electrophoresis image and Figure 2 shows a perspective view of the same detail.

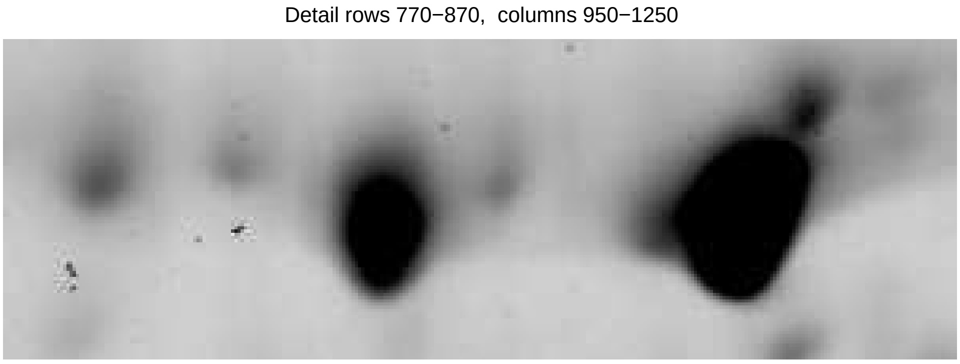


Figure 1: Detail of an intensity gray scale 2D electrophoresis gel image.

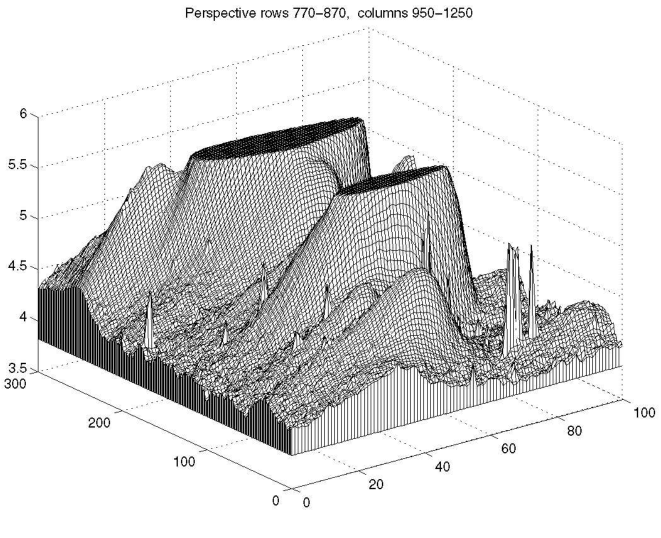


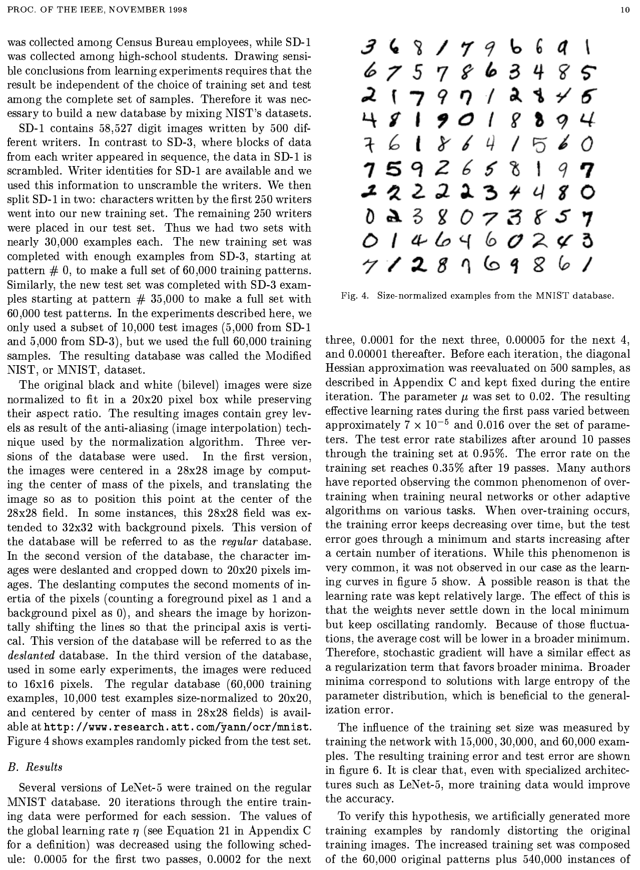
Figure 2: Perspective of the image in Figure 1 as seen from upper left corner

a) Ideally a 2D electrophoresis image should consist of a smoothly varying background and a superpositions of spots with shapes as densities of axis-parallell two-dimensional normal distributions. Try to describe how the image shown in Figure 1 deviates from this ideal.

b) The amount of protein in a spot should ideally be obtained by summing the intensity of the pixels in the spot with the local background subtracted. Suggest with suitable formulas how the amount of protein in the leftmost spot (with centre approximately in row 810 and column 980) in Figure 1 could be computed.

c) Suggest with suitable formulas how the amount of protein in the large spot slightly to the left of the middle in Figure 1 could be computed.

Problem 2. Suggest a method for recognition of handwritten digits like the ones shown in the upper right corner in Figure 3, and discuss how you could conduct an experiment to evaluate the method. Give suitable formulas both for describing the recognition method and for the evaluation.



PROC. OF THE IEEE, NOVEMBER 1998 10

was collected among Census Bureau employees, while SD-1 was collected among high-school students. Drawing sensible conclusions from learning experiments requires that the result be independent of the choice of training set and test among the complete set of samples. Therefore it was necessary to build a new database by mixing NIST's datasets.

SD-1 contains 58,527 digit images written by 500 different writers. In contrast to SD-3, where blocks of data from each writer appeared in sequence, the data in SD-1 is scrambled. Writer identities for SD-1 are available and we used this information to unscramble the writers. We then split SD-1 in two: characters written by the first 250 writers went into our new training set. The remaining 250 writers were placed in our test set. Thus we had two sets with nearly 30,000 examples each. The new training set was completed with enough examples from SD-3, starting at pattern # 0, to make a full set of 60,000 training patterns. Similarly, the new test set was completed with SD-3 examples starting at pattern # 35,000 to make a full set with 60,000 test patterns. In the experiments described here, we only used a subset of 10,000 test images (5,000 from SD-1 and 5,000 from SD-3), but we used the full 60,000 training samples. The resulting database was called the Modified NIST, or MNIST, dataset.

The original black and white (bilevel) images were size normalized to fit in a 20x20 pixel box while preserving their aspect ratio. The resulting images contain grey levels as result of the anti-aliasing (image interpolation) technique used by the normalization algorithm. Three versions of the database were used. In the first version, the images were centered in a 28x28 image by computing the center of mass of the pixels, and translating the image so as to position this point at the center of the 28x28 field. In some instances, this 28x28 field was extended to 32x32 with background pixels. This version of the database will be referred to as the *regular* database. In the second version of the database, the character images were deslanted and cropped down to 20x20 pixels images. The deslanting computes the second moments of inertia of the pixels (counting a foreground pixel as 1 and a background pixel as 0), and shears the image by horizontally shifting the lines so that the principal axis is vertical. This version of the database will be referred to as the *deslanted* database. In the third version of the database, used in some early experiments, the images were reduced to 16x16 pixels. The regular database (60,000 training examples, 10,000 test examples size-normalized to 20x20, and centered by center of mass in 28x28 fields) is available at `http://www.research.att.com/~yann/ocr/mnist`. Figure 4 shows examples randomly picked from the test set.

B. Results

Several versions of LeNet-5 were trained on the regular MNIST database. 20 iterations through the entire training data were performed for each session. The values of the global learning rate η (see Equation 21 in Appendix C for a definition) was decreased using the following schedule: 0.0005 for the first two passes, 0.0002 for the next three, 0.0001 for the next three, 0.00005 for the next 4, and 0.00001 thereafter. Before each iteration, the diagonal Hessian approximation was reevaluated on 500 samples, as described in Appendix C and kept fixed during the entire iteration. The parameter μ was set to 0.02. The resulting effective learning rates during the first pass varied between approximately 7×10^{-5} and 0.016 over the set of parameters. The test error rate stabilizes after around 10 passes through the training set at 0.95%. The error rate on the training set reaches 0.35% after 19 passes. Many authors have reported observing the common phenomenon of over-training when training neural networks or other adaptive algorithms on various tasks. When over-training occurs, the training error keeps decreasing over time, but the test error goes through a minimum and starts increasing after a certain number of iterations. While this phenomenon is very common, it was not observed in our case as the learning curves in figure 5 show. A possible reason is that the learning rate was kept relatively large. The effect of this is that the weights never settle down in the local minimum but keep oscillating randomly. Because of those fluctuations, the average cost will be lower in a broader minimum. Therefore, stochastic gradient will have a similar effect as a regularization term that favors broader minima. Broader minima correspond to solutions with large entropy of the parameter distribution, which is beneficial to the generalization error.

The influence of the training set size was measured by training the network with 15,000, 30,000, and 60,000 examples. The resulting training error and test error are shown in figure 6. It is clear that, even with specialized architectures such as LeNet-5, more training data would improve the accuracy.

To verify this hypothesis, we artificially generated more training examples by randomly distorting the original training images. The increased training set was composed of the 60,000 original patterns plus 540,000 instances of

Fig. 4. Size-normalized examples from the MNIST database.

Figure 3: One page from a paper on recognition of handwritten digits. (Consider only the figure in the upper right corner of this page, but disregard the rest of the text.)