

1. 予備知識

ハガキに書かれた郵便番号のように、ほぼ同じ大きさで手書きされた数字を、コンピュータで読み取ることを考えよう。ひとつの手書き数字をカメラで撮影し、コンピュータに画像として入力したとする。この画像があらわす数字を正しく決定できれば、手書き数字が読み取れたことになる。入力画像の数字をどのようにして決定すればよいであろうか。

ひとつの方法は、図1のようにあらかじめ0から9までの手本となる画像を用意しておき、その画像と比較してどれに最も似ているかを判定することである。最も似ている画像が定められれば、その手本画像があらわす数字を読み取り結果とすればよい。

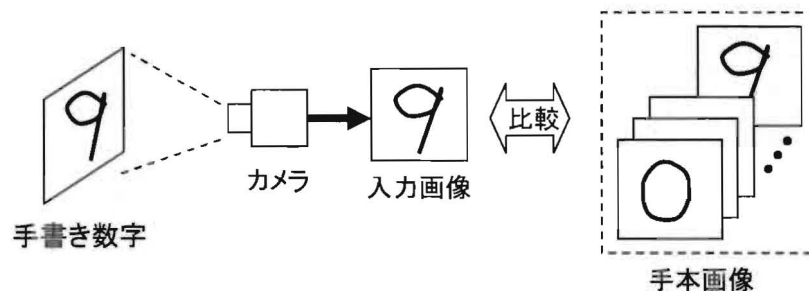


図1

では、撮影された未知の数字（以下ではテストデータと呼ぶ）と、手本とする既知の数字（以下では教師データと呼ぶ）の似ている程度（以下では類似度と呼ぶ）を、どのように測ればよいであろうか。最も簡単な方法としては、テストデータと教師データの画像を重ねてみて、文字部分が重なる割合を調べることが考えられる。重なる部分が多いほど似ている程度（類似度）が高いと見なすことができるであろう。ここでは、この方法をプログラムで実現しよう。

コンピュータで扱う画像とは、図2のように格子状に配置された画素の集まりである。以下では、ひとつの数字が縦14画素、横14画素の画像としてコンピュータに入力されると仮定する。背景（白）の画素を数値0で、文字（黒）の画素を数値1に置き換えて表現すると、ひとつの手書き数字の入力画像は、0または1の数値が縦に14個、横に14個、合計196個並んだものである。

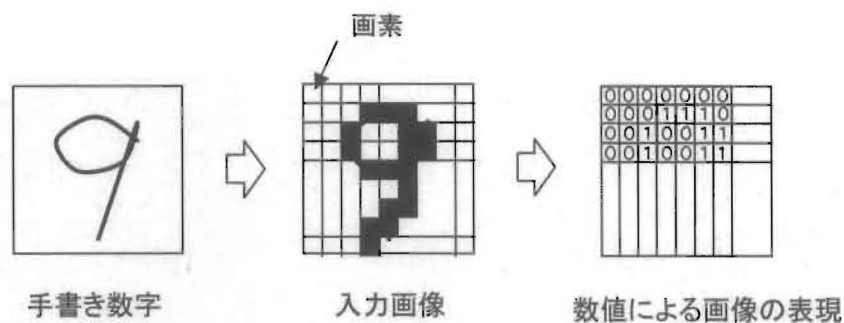


図2

これらの数値を一行に並べれば 196 次元のベクトルとみなすことができる。図3のように入力画像の 196 次元ベクトルと手本画像の 196 次元ベクトルを比較して、値が 1 である要素が両者で何個一致しているかを調べると、それは 2 つの画像に描かれた数字の重なり具合を示している。この値 1 の画素が一致している割合を類似度と定義し、与えられたテストデータとの類似度が最も高い教師データを決定して、手書き数字の読み取り結果としよう。

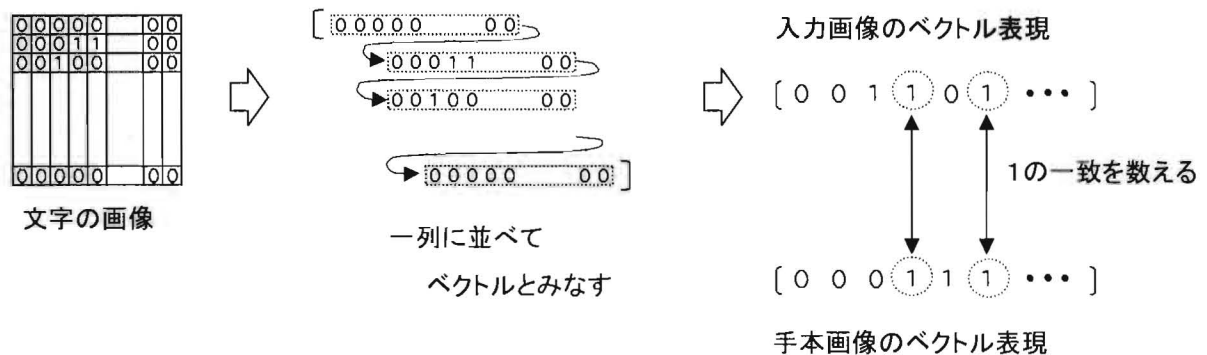


図3

2. 課題

ここでは問題を単純化するために、出現する数字は「8」と「9」の2種類だけであるとして、入力された数字の読み取りを考える。以下に与えられた課題を解くプログラムを作成し、課題ごとに指定された名前でソースファイルを保存するとともに、得られた結果を解答用紙に記入せよ。

課題を解くにあたってはC言語が利用できる。コンピュータの操作方法については別紙を参照し、疑問があれば手を上げて監督者に尋ねること。

練習1

この練習の結果については解答する必要はない。以下のサンプルプログラムをコンパイルして実行し、動作を確認せよ。ここで与えたサンプルプログラムは、以下の課題で自由に利用してよい。

サンプルプログラム1 (ソースファイル名: `sample1.c`)

数字「8」と数字「9」の教師データ (ファイル名: `8tehon.dat` と `9tehon.dat`) をファイルから読み込み、196次元ベクトルの要素を画面にそれぞれ表示する。また、値が1である画素の数 N_8 および N_9 を画面にそれぞれ表示する。

コンパイル方法:

「`gcc -o sample1 sample1.c`」 と入力して Return キーを押す。

実行方法:

「`./sample1`」 と入力して Return キーを押す。

課題1 (作成したソースファイルを `kadai1.c` として保存すること)

数字「8」と数字「9」の教師データ (ファイル名: `8tehon.dat` と `9tehon.dat`) を比較し、両者において一致した値1の画素の数を求めよ。

解答欄:

一致した値1の画素数

課題2 (作成したソースファイルを `kadai2.c` として保存すること)

テストデータ (ファイル名: `testdata.dat`) を読み込み, それを数字「8」および数字「9」の教師データ (課題1と同じファイル) とそれぞれ比較して, 一致した値1の画素の数 n_8 および n_9 をそれぞれ求めよ.

解答欄:

「8」の教師データと一致した値1の画素数

 n_8 :

「9」の教師データと一致した値1の画素数

 n_9 :課題3 (作成したソースファイルを `kadai3.c` として保存すること)

テストデータ (ファイル名: `testdata.dat`) を読み込み, それを数字「8」および数字「9」の教師データ (課題1と同じファイル) と比較して, 一致した値1の画素数の割合 (n_8/N_8 および n_9/N_9) をそれぞれ求めよ. また, その結果からテストデータがどちらの数字であるかを判定せよ.

解答欄:

「8」の教師データと一致した値1の画素数の割合

 n_8/N_8 :

「9」の教師データと一致した値1の画素数の割合

 n_9/N_9 :

判定結果

課題4（作成したソースファイルを kadai4.c として保存すること）

数字「8」であることがわかっているテストデータ 10 個（ファイル名：8tes0.dat～8tes9.dat）と、数字「9」であることがわかっているテストデータ 10 個（ファイル名：9tes0.dat～9tes9.dat）に対して、課題3と同じ方法で数字を判定し、合計 20 データのうち何個を正しく判定することができるかを求めよ。

解答欄：

正しく判定できたテストデータの数

練習2

この練習の結果については解答する必要はない。以下のサンプルプログラムをコンパイルして実行し、動作を確認せよ。

サンプルプログラム2（ソースファイル名：sample2.c）

数字「8」および数字「9」の教師データをそれぞれ 10 個ずつに増やす。数字「8」の教師データ（ファイル名：8tr0.dat～8tr9.dat）および数字「9」の教師データ（ファイル名：9tr0.dat～9tr9.dat）をそれぞれ読み込み、それぞれの教師データにおける値 1 の画素数 $N_{80} \sim N_{89}$ および $N_{90} \sim N_{99}$ を表示する。

課題5（作成したソースファイルを kadai5.c として保存すること）

テストデータとして（ファイル名：testdata.dat）を読み込み、サンプルプログラム2で用いた、数字「8」および数字「9」の教師データ（各 10 個）との類似度（一致した値 1 の画素の割合）をそれぞれ計算して、類似度が最も高い教師データのファイル名を表示せよ。

解答欄：

類似度が最も高い教師データのファイル名：

課題6（作成したソースファイルを `kadai6.c` として保存すること）

数字「8」であることがわかっているテストデータ 10 個（ファイル名：`8tes0.dat`～`8tes9.dat`）と、数字「9」であることがわかっているテストデータ 10 個（ファイル名：`9tes0.dat`～`9tes9.dat`）に対して、課題5と同じ方法でそれぞれ数字を判定し、合計 20 データのうち何個を正しく判定することができるかを求めよ。

解答欄：

正しく判定できたテストデータの数

課題7

課題6の方法で正しく読み取ることが難しい数字があるとすれば、それはどのような原因によると考えられるか、またそれを改善するにはどのような方法があるか、自分の考えを説明せよ。

解答欄：