Guion 2: Compresión sin pérdida  
Código de Huffman

Información sobre la entrega de la práctica

Las prácticas se entregarán en un único fichero comprimido Practica02ApellidoNombre.zip. El fichero contendrá:

* Las funciones de Matlab a realizar en ficheros .m con los nombres de las funciones que se indiquen en el guion.
* Los trozos de código a realizar, que se entregarán todos en los pasos correspondientes de un único fichero .m llamado Practica02ApellidoNombre.m . Este fichero lo crearás modificando el fichero .m Practica02MolinaRafael.m en el servidor.
* Las discusiones y respuestas solicitadas en el guion se entregarán en un único fichero pdf. El nombre del fichero será Practica02ApellidoNombre.pdf. Lo construirás editando Practica02MolinaRafael.doc y salvándolo en formato pdf.

En este guion estudiaremos el código de Huffman como método de compresión de datos sin pérdida, es decir, si se comprimen y descomprimen los datos se vuelven a obtener los datos originales.

# Sobre la Codificación Huffman

Puedes codificar ficheros de datos mediante codificación Huffmanutilizando el software desarrollado por [Michael Dipperstein](http://michael.dipperstein.com/huffman/index.html). En la página web de [Michael Dipperstein](http://michael.dipperstein.com/huffman/index.html) se puede encontrar una clara explicación del algoritmo y las consideraciones de implementación del mismo. Aunque nosotros haremos las prácticas en Matlab te recomendamos que visites el sitio de [Michael Dipperstein](http://michael.dipperstein.com/huffman/index.html),contiene unas páginas muy buenas sobre compresión.

Codificación Huffman usando Matlab

**Paso 1**

Vamos a comenzar suponiendo que tenemos una fuente de la que conocemos los símbolos que genera y la probabilidad de cada una de ellos. Matlab nos permite generar el código de Huffman de una forma muy sencilla

A=[1:5];

p=[0.2 0.4 0.2 0.1 0.1];

D=huffmandict(A,p);

for i=1:length(A)

fprintf('LetraAlfabeto %s Código %s\n',num2str(D{i,1}),num2str(D{i,2}))

end

La salida de este trozo de código sería

Letra Alfabeto 1 Código 0 0 0

Letra Alfabeto 2 Código 1

Letra Alfabeto 3 Código 0 1

Letra Alfabeto 4 Código 0 0 1 1

Letra Alfabeto 5 Código 0 0 1 0

**Paso 2**

Podemos utilizar otros símbolos para los elementos del alfabeto. Halla y muestra, siguiendo el esquema del paso anterior, el código de Huffman correspondiente al alfabeto {P,Q,R,S,T} con probabilidades p=[0.2 0.4 0.2 0.1 0.1]. Cuya ejecución debe producir la salida que se muestra más abajo. Incluye el código en el paso 2 de Practica02ApellidoNombre.m

Letra Alfabeto P Código 0 0 0

Letra Alfabeto Q Código 1

Letra Alfabeto R Código 0 1

Letra Alfabeto S Código 0 0 1 1

Letra Alfabeto T Código 0 0 1 0

Incluye el código aquí también (además de en el paso 2 del fichero .m)

clear all;

A=uint8(['P' 'Q' 'R' 'S' 'T']);

p=[0.2 0.4 0.2 0.1 0.1];

D=huffmandict(A,p);

for i=1:length(A)

fprintf('Letra Alfabeto %s Código %s\n', char(D{i,1}), num2str(D{i,2}))

end

**Paso 3**

Podemos utilizar la función huffmandict para construir diccionarios de máxima o mínima varianza. El siguiente ejemplo construye el código de Huffmancon mínima varianza para el alfabeto A=[1:5] y probabilidades p=[0.2 0.4 0.2 0.1 0.1]. Entiende las opciones de la función huffmandict.

A=[1:5];

p=[0.2 0.4 0.2 0.1 0.1];

D=huffmandict(A,p,2,'min');

fori=1:length(A)

fprintf('LetraAlfabeto %s Código %s\n',num2str(D{i,1}),num2str(D{i,2}))

end

La salida de este código es

Letra Alfabeto 1 Código 1 1

Letra Alfabeto 2 Código 0 0

Letra Alfabeto 3 Código 1 0

Letra Alfabeto 4 Código 0 1 1

Letra Alfabeto 5 Código 0 1 0

**Paso 4**

En los ejemplos anteriores hemos supuesto que conocemos las probabilidades de cada símbolo. En condiciones reales estas probabilidades tendrán que ser estimadas/calculadas utilizando un fichero concreto o un conjunto de ficheros. Antes de ver como se estiman/calculan las probabilidades de cada símbolo recordemos una forma de generar símbolos. En el ejemplo siguiente simulamos la generación de 10 símbolos del alfabeto anterior

A=[1:5];

p=[0.2 0.4 0.2 0.1 0.1];

rng(0);

stream= randsrc(1,10,[A;p])

El resultado de este trozo de código es

4 5 1 5 3 1 2 2 5 5

Observa que podría no haber aparecido algún símbolo.

**Paso 5**

Vamos ahora a leer un fichero del que estimaremos la probabilidad de cada letra de un alfabeto. Primero leemos el fichero Don Quijote de la Mancha - Miguel de Cervantes.txt y lo almacenamos en un vector fila de bytes

clc; clearall;

fichero= 'Don Quijote de la Mancha - Miguel de Cervantes.txt'

fid=fopen(fichero,'r')

seq=fread(fid,'\*uint8');

fclose(fid);

seq=reshape(seq,1,length(seq)); %leidos datos en seq vector fila

fprintf('Tamaño del fichero original en bytes %d\n',numel(seq))

Tamaño del fichero original en bytes 2117266

**Paso 6**

Calculamos su histograma y lo mostramos

letras=[0:255];

histo=histc(seq,letras); %calculamos histograma e indices

bar(letras,histo);

axis('tight');

xlabel('letras [0:255]');

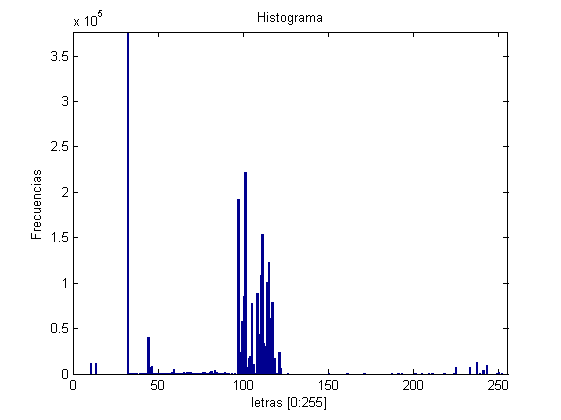
ylabel('Frecuencias');

title('Histograma');

xlabel('letras [0:255]');

ylabel('Frecuencias');

title('Histograma');



**Paso 7**

De las 256 letras del alfabeto solo usamos las que aparecen en el fichero'Don Quijote de la Mancha - Miguel de Cervantes.txt'. Hállalas y calcula sus probabilidades. Puedes utilizar la función find, por ejemplofind(histo>0) que devuelve las posiciones de la matriz histo que almacenan valores mayores que cero. No debes confundir las posiciones de las letras en letras que van de 1 a 256 con los valores de las letras que van de 0 a 255. Escribe el código correspondiente en el paso 7 de Practica02ApellidoNombre.m

Incluye el código aquí también (además de en el paso 2 del fichero .m)

pos\_letras\_usadas=find(histo>0);

letras\_usadas=letras(pos\_letras\_usadas);

hist\_letras\_usadas=histo(pos\_letras\_usadas);

prob\_letras\_usadas=hist\_letras\_usadas/sum(hist\_letras\_usadas(:));

**Paso 8**

A continuación calculamos el diccionario para las letras usadas y sus probabilidades y codificamos la secuencia que teníamos suponiendo que has almacenado en letras\_usadas las letras utilizadas (números en [0:255]) y sus probabilidades en prob\_letras\_usadas.

En elcódigo anteriorletras\_usadas puede ser elementos en el rango [0:255] olos caracteres correspondientes. Mira la documentación de huffmandict, yo he utilizado números en [0:255].

Finalmente, muestra el número de letras utilizadas y la longitud de la codificación

[dict,avglen] = huffmandict(letras\_usadas,prob\_letras\_usadas) ;

seq\_codificada = huffmanenco(seq,dict);

fprintf('Número de letras usadas %d\n',numel(letras\_usadas))

fprintf('Longitud de la secuencia codificada %d\n', length(seq\_codificada))

debedevolver como resultados

Número de letras usadas 111

Longitud de la secuencia codificada 9333683

**Paso 9**

¿Cómo guardaríamos el fichero codificado?. Podríamos utilizar una cabecera que tuviera las letras usadas (1 byte por letra) y sus probabilidades en formato double (8 bytes por letra) más una letra adicional, por ejemplo, 0 con probabilidad 0 (1 + 8 bytes más) para indicar que ya no tenemos más letras. A continuación guardaríamos la secuencia codificada que necesitaría  bytes

En nuestro ejemplo

tamagno\_comprimido=(length(letras\_usadas)+1)\*1+ ...

(length(letras\_usadas)+1)\*8+ ...

ceil(length(seq\_codificada)/8);

fprintf('Tamaño fichero comprimido en bytes %d\n',tamagno\_comprimido)

que produce

Tamaño fichero comprimido en bytes 1167719

Observa que no hemos discutido cómo terminar la decodificación. El uso de ceil nos dice que podemos tener unos bits al final del fichero codificado que no corresponden a ningún símbolo y que si los decodificamos nos habremos equivocado. Estos problemas los discutiremos al final del guión.

**Paso 10**

¿Cómo reconstruiríamos el fichero comprimido?. Leeríamos la cabecera y de ella sacaríamos las letras\_usadas y prob\_letras\_usadas, después leeríamos la secuencia codificada que tendríamos que desempaquetar y convertir en una matriz de doubles que sólo contiene 0s y 1s y que llamaríamos seq\_codificada, esta seq\_codificada va a coincidir con la que obtuvimos en el paso 8. A continuación con letras\_usadas y prob\_letras\_usadas construimos el diccionario que sería el diccionario dict que obtuvimos en el paso 8 y finalmente decodificamos la secuencia usando

deco=huffmandeco(seq\_codificada,dict);

**Paso 11**

Muy importante, comprobamos que la secuencia original y decodificada son iguales

fprintf('¿Coinciden original y comprimido 1(S) 0 (N)?, %d\n',...

isequal(seq,uint8(deco)))

¿Coinciden original y comprimido 1(S) 0 (N)?, 1

**Paso 12**

Incluye en el paso 12 del fichero Practica02ApellidoNombre.pdf el tamaño del fichero original y el comprimido. Incluye también el factor de compresión.

Escribe tus respuestas aquí.

Tamaño del fichero original (en bytes) = 2117266

Tamaño del fichero comprimido (en bytes) = 1167719

Factor de compresión = 1.81316

**Paso 13**

Crea ahora la siguiente estructura

clc; clear all;

N=9;

Fichero=cell(N);

Fichero{1}='constitucion española.txt';

Fichero{2}='Fundacion e Imperio - Isaac Asimov.txt';

Fichero{3}='Cinco semanas en globo - Julio Verne.txt';

Fichero{4}='ptt1.pbm';

Fichero{5}='ptt4.pbm';

Fichero{6}='ptt8.pbm';

Fichero{7}='camera.pgm';

Fichero{8}='bird.pgm';

Fichero{9}='bridge.pgm';

fichero='ptt8.pbm';

**Paso 14**

Para cada uno de los ficheros anteriores:‘constitucion española.txt’, ‘Fundacion e Imperio - Isaac Asimov.txt’,‘Cinco semanas en globo - Julio Verne.txt’, ptt1.pbm, ptt4.pbm, ptt8.pbm, camera.pgm, bird.pgm ybridge.pgmescribe en el paso 14 de Practica02ApellidoNombre.mel código Matlab que calcule el tamaño del fichero original, el comprimido usando el código de Huffman y el factor de compresión obtenida en cada uno de los casos. El código Matlab debe comprobar también que los ficheros originales y descomprimidos son iguales.

Completa en el paso 14 de Practica02ApellidoNombre.pdf la tabla siguiente, realiza una comparación crítica de los resultados obtenidos y explica la variación en el factor de compresión.

Escribe tus respuestas aquí

|  |  |  |  |
| --- | --- | --- | --- |
| Fichero | Tamaño original | Tamaño Comprimido | FC |
| constitucion española.txt | 112246 | 64095 | 1.751244 |
| Fundacion e Imperio - Isaac Asimov.txt | 461298 | 259641 | 1.776676 |
| Cinco semanas en globo - Julio Verne.txt | 487020 | 275634 | 1.766908 |
| ptt1.pbm | 513249 | 87435 | 5.870063 |
| ptt4.pbm | 513249 | 145557 | 3.526103 |
| ptt8.pbm | 513249 | 115572 | 4.440946 |
| camera.pgm | 65551 | 59961 | 1.093227 |
| bird.pgm | 65551 | 57069 | 1.148627 |
| bridge.pgm | 65551 | 65356 | 1.002984 |

Los archivos con patrones y repeticiones tienden a comprimirse bien, mientras que las imágenes con regiones grandes de píxeles del mismo valor también un buen rendimiento.

**Paso 15**

Vamos a usar el código de Huffman para comprimir una imagen agrupando de dos en dos sus valores. Lo haremos con unaimagen sintéticaque convertiremos a una secuencia, seq

clc; close all; clear all;

M=256;

img=uint8(zeros(M));

img(:,129:M)=255;

imshow(img);

img16=uint16(img);

img2=256\*img16(:,1:2:size(img,2))+img16(:,2:2:size(img,2));

[U,V]=size(img2);

seq=reshape(img2,1,U\*V);

**Paso 16**

Codifica la matriz img, que no agrupa los bytes de dos en dos, usando el código de Huffman, comprueba que coinciden la secuencia original y la decodificación de la comprimida, calcula el espacio que ocupa la imagen comprimida incluida la cabecera y el factor de compresión. Escribe el trozo de código en el paso 16 de Practica02ApellidoNombre.m.

Escribe en el paso 16 de Practica02ApellidoNombre.pdf el factor de compresión calculada.

Escribe tus respuestas aquí

Longitud la secuencia codificada 65536

Tamaño fichero comprimido en bytes 8219

Factor de compresión: 7.973719

¿Coinciden original y comprimido 1(S) 0 (N)?, 1

**Paso 17**

Codifica la matriz img2, que agrupa los bytes de dos en dos, usando el código de Huffman, comprueba que coinciden la secuencia original y la decodificación de la comprimida, calcula el espacio que ocupa la imagen comprimida incluida la cabecera y el factor de compresión. Escribe el trozo de código en el paso 17 de Practica02ApellidoNombre.m.

Escribe en el paso 17 de Practica02ApellidoNombre.pdf el factor de compresión calculada.

Escribe tus respuestas aquí

Longitud la secuencia codificada 32768

Tamaño fichero comprimido en bytes 4123

Factor de compresión: 7.947611

¿Coinciden original y comprimido 1(S) 0 (N)?, 0

**Paso 18**

¿Cuáles son los factores de compresión para las matrices img2 e img obtenidas en los dos pasos anteriores?.Discute los resultados obtenidos en los dos pasos anteriores en el paso 18 de Practica02ApellidoNombre.pdf.

Escribe tus respuestas aquí

FC img = 7.973719

FC img2 = 7.947611

La diferencia entre los dos factores de compresión significaría que en la primera imagen se alcanza mejor compresión.

**Paso 19**

Utiliza ahora la imagen goldhill.pgm

close all; clear all;

img=imread('goldhill.pgm');

imshow(img);

img16=uint16(img);

img2=256\*img16(:,1:2:size(img,2))+img16(:,2:2:size(img,2));

**Paso 20**

Incluye en el paso 20 del fichero Practica02ApellidoNombre.m el código en Matlab que calculalos factores de compresión para estas nuevas matrices img2 e img obtenidas a partir de la imagen goldhill.pgm del paso anterior. Incluye el tamaño de la cabecera. Si no pudieses codificar alguna de las dos matrices explica por qué.

Discute los resultados obtenidos en el paso 20 de Practica02ApellidoNombre.pdf.

Escribe tus respuestas aquí

1.03480

Sobre img:

Longitud de la secuencia codificada = 490984

Tamaño fichero comprimido en bytes = 63335

FC = 1.03475

Sobre img2:

Longitud del fichero original en bytes = 32768

Y no he podido obtener más datos porque daba este error: Maximum recursion limit of 500 reached. Use set(0,'RecursionLimit',N) to change the limit. Be aware that exceeding your available stack space can crash MATLAB and/or your computer. Y al aumentar el límite se me cerraba MATLAB.

**Paso 21**

Lee la secuencia miss\_am.yuv, muéstrala y extrae la luminancia de los dos primeros fotogramas

clc, close all; clear all;

file\_name = 'miss\_am.yuv';

file\_format = 'QCIF\_PAL';

num\_of\_frames = 30;

[yuv\_movie, yuv\_array] = readYUV(file\_name, num\_of\_frames, ...

file\_format);

implay(yuv\_movie)

fotograma1=yuv\_array(:,:,1,1);

fotograma2=yuv\_array(:,:,1,2);

figure; imshow(fotograma1);

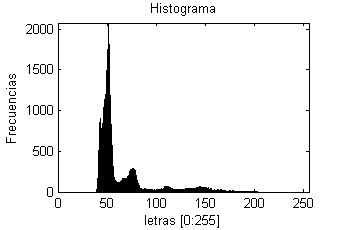
figure; imshow(fotograma2);

**Paso 22**

Utiliza el código de Huffman para codificar la matrizfotograma2. Escribe en el paso 22 de Practica02ApellidoNombre.m el código en Matlab para calcular el histograma de la imagen original, el tamaño de la matriz comprimida, incluida la cabecera, y su factor de compresión. Comprueba que la matriz original y la decodificación de la comprimida coinciden.

Incluye el histograma, el tamaño de la matriz original y el de la comprimida (con cabecera) en el paso 22 de Practica02ApellidoNombre.pdf. Escribe también el factor de compresión

Escribe tus respuestas aquí



Longitud la secuencia codificada 142360

Tamaño fichero comprimido en bytes 19262

Factor de compresion: 1.315751

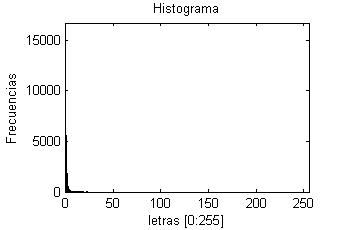
¿Coinciden original y comprimido 1(S) 0 (N)?, 1

**Paso 23**

Como puedes comprobar la nueva representación de la imagen ocupa menos espacio que la imagen original pero la reducción del tamaño no es grande. Escribe código en Matlabquedibuje el histograma de fotograma2-fotograma1, codifique la matriz diferencia usando Huffman, compruebe que fotograma2 coincide con su reconstrucción y calcule el tamaño de fotograma2-fotograma1 comprimido, incluida su cabecera. Incluye el código en el paso 23 de Practica02ApellidoNombre.m. Este ejercicio puedes hacerlo de dos formas según codifiques las diferencias. Realízalo comoconsideres más apropiado.

Incluye el histograma de fotograma2-fotograma1, y el tamaño de la matriz comprimida (con cabecera) en el paso 23 de Practica02ApellidoNombre.pdf. Escribe también el factor de compresión.

Escribe tus respuestas aquí



Longitud la secuencia codificada 39937

Tamaño fichero comprimido en bytes 5191

Factor de compresion: 4.882296

¿Coinciden original y comprimido 1(S) 0 (N)?, 1

**Paso 24**

Compara en el paso 24 del fichero Practica02ApellidoNombre.pdf los tamaños de las matrices fotograma2 y fotograma2-fotograma1 comprimidos obtenidos en los pasos 22 y 23. ¿A qué se debe la mejora?

Escribe tus respuestas aquí

Como el tamaño de fotograma2 comprimida es 19262 y de fotograma2-fotograma1 es 5191, se ve una mejora debido a que como fotograma1 y fotograma2 son tan similares la matriz de diferencias solo debe codificar unos pocos valores.

**Paso 25**

En el paso 9 nos dejamos pendiente detectar que habíamos decodificado toda la secuencia de 0s y 1s. Discute en el paso 25 de Practica02ApellidoNombre.pdf posibles soluciones a este problema.

Escribe tu respuesta aquí

Para garantizar que se ha decodificado toda la secuencia, se puede: Codificar la longitud de bits de la secuencia antes de los datos reales. Al decodificar, primero lee la longitud y luego la cantidad de bits especificada por la longitud. Cuando hayas leído esa cantidad de bits, sabrás que has terminado de decodificar. Esto asegura que no te saltes ningún bit.

**Paso 26**

En el paso 9 definimos una cabecera. Se te ocurre alguna(s) mejora(s) que ahorre(n) espacio?. Incluye tu respuesta en el paso 26 de Practica02ApellidoNombre.pdf.

Escribe tu respuesta aquí

**Codificación variable de longitud:** En lugar de asignar una longitud fija de bits a cada campo de la cabecera, podemos utilizar codificación variable de longitud.

**Compresión de la cabecera:** Podemos aplicar algoritmos de compresión de datos a la cabecera misma.