Image and Video Processing

Handout 1

Course Notes for Integrated Systems Design

Dr. A. C. Kokaram

Department of Electronic and Electrical Engineering,

University of Dublin Trinity College.

Contents

1 Introduction and Course Scope

This course consists of 27 lectures which introduce the basics of image and video processing including a working introduction to elements of image coding. The following will be covered

IMAGE ANALYSIS AND CODING

- Human Visual Perception
- 2D Fourier Xform, 2D DCT
- Basic Image Manipulation/Characterisation: Histograms, Filtering, Overlapped Processing
- Simple Compression Example [Haar Transform]
- DCT and JPEG
- Filterbanks and Wavelets

VIDEO ANALYSIS AND CODING

- Video Formats: CCIR Rec 601, Composite Video, Interlacing
- Sampling in space/time
- Motion Estimation (Image Centric): Direct Search, Optic Flow, Multiresolution methods
- Video Filtering (Motion compensated and non-motion compensated)
- Changing the sampling rate: Deinterlacing/Upsampling
- Basics of MPEG Schemes

There are two main references for this course [1, 2]. The coursework involves exercises using both Matlab and written material. The overall goal is to allow the student to diagnose problems in the design and implementation of video and image processing systems such as MPEGx codec chipsets and turnkey visual consumer/broadcast devices.

1.1 The rise of the Digital Image

Work in Digital Image Processing (DIP) began as early as the 1960's with NASA programme of space exploration. Pictures were broadcast (painfully slowly) pixel by pixel and sometimes the colour/intensity map of the image was created by hand by simply colouring in a square on a very large sheet of paper with the relevant colour. Digital computers were used to correct for the blurring caused by lenses and enhancement techniques were developed for improving the picture detail. It is interesting to note that the rise of DIP follows closely the development of picture reproduction and display devices[1] e.g. laser (dry ink) printing, dye sublimation printing, wax transfer and now ink jet printing.

The last 5 years has seen an explosion in the availability of digital visual devices and hence digital visual media. Digital Television set top boxes are now available free from SKY in Ireland, SKY has been broadcasting DTV for about 3 years now. DVD (Digital Video Disk) sales have outstripped CD sales at a comparable point in their releases. PDAs (Personal Digital Assistants) like Compaq Imode, HP Journada, Visor, PalmPilot now seek to include camera add ons as well as wireless. WebCams have been available for a long time, and Digital Cameras are getting better. Medical analysis is now exclusively in the digital domain as it is much easier to manipulate and enhance the images for diagnosis. Words like JPEG, MPEG, AVI, Graphics seem in as common use by under 20's. Special fx at the movies are now the norm.

These devices and media all require the design of systems which are growing increasingly complex, both from the point of view of data management and data processing itself. Image and Video compression in particular are key technologies that have enabled consumer devices since the media data rate would be otherwise too large for practical design. The designer needs to understand the compromises which must be made in handling visual media, and the key concepts which make that handling possible in the first place. That is what this course is about.

1.1.1 The Hot (Research?) Areas

The proliferation of digital capture devices and consumer video equipment has fuelled new industrial research trends as users are exposed to these novel devices and media.

- MPEG4/H.264 is more error-resilient than MPEG2 and gives better image quality at the same bitrate. This implies it is suitable for Wireless Video in particular, and Internet Streaming from the point of view of standardisation. MPEG4 though, needs high level

[1] Why do DIP if not many people can see the results?

analysis tools for building the encoder and more error correction at the decoder in real situations. **Video Object Segmentation and Error Concealment** important.

- **Digital Cinema** allows the motion picture houses more control, faster distribution and better quality reproduction. Compression is a serious issue here as each frame is very high resolution e.g. $2048 \times 1152 \times 3 = 6.75\text{Mb} = 168\text{MB/sec}$!

- It is very very hard to search digital media in the same way that search engines can search text for instance. A movie, or a photograph means different things to different people. It is possible to attach keywords to media stream to describe it, but these keywords only describe what is in the mind of the cataloguer.

  Thus we may wish to access photos from a nature reserve (for instance) which show wildlife hunt chases, but no one had bothered to index their photos like this at the time. Is it possible to design image processing routines which can extract this information by 'understanding' the picture content?

  Video Editors are swamped with clips which they need to compile into movies, and more and more amateurs make movies as cheap DV (Digital Consumer Video Tape) devices become available[2]. How does one make this task simple?

  Automated **Content Analysis/Management/Retrieval** has become the BIGGEST multimedia research area in the last 5 years. MPEG7 addresses a mechanism for describing content which could be related to this idea.

- Broadcasters are starved for content despite available media channels (DTV, DVD, Internet Streaming). Archives hold that content. But it is in bad condition and needs re-touching or *restoration*. **Film and Video Restoration** is becoming more important.

- Noise reduction of video/film allows better compression. Several companies now offer 'cleaning' services for DVD.

- Digital media is too easy to copy and thus redistribute without the manufacturer's authority. It is possible to digitally watermark such media to allow it to be traced. **Digital Watermarking** is a key component in security measures for the protection of digital media[3] Seminal work in this area was performed at the EEE Dept., TCD around 1996.

- Online video processing throught the www is becoming of incresing interest. This is being pushed by availability of decent broadband and continually better mobile phones.

[2] Blair Witch Project, BBC location broadcasts/documentaries, Real-Life Docudramas

[3] Although by itself it is useless. Standards play an important role here, as does security key generation and validation.

- Conversion and Scaling is required to convert between HD and SD and Mobile phone pictures. Doing this well is important for Flat Screen TV technology as well as crazy people who want to watch Mobile phone footage on their 40inch HD plasma display.

1.1.2 The big Research/Industrial Conferences

For industrials and researchers alike a number of important conferences have arisen in the last 10 years which are excellent venues to observe the state of the art in new developments. Here are a few:

- International Conference on Acoustics Speech and Signal Processing (ICASSP): Annual (In 2008 : Vegas). The biggest Digital Signal Processing conference `www.icassp2001.org`: Deals with everything from Filter Design/Implementation to Video Networking and Compression
- International Conference in Image Processing (ICIP), Annual (In 2008:) Deals with all things image and video. The biggest image and video processing conference.
- NAB (National Association of Broadcasters), US Based Industrial Broadcaster Conference (Video, Audio). (Vegas)
- IBC (International Broadcast Convention), European Based like NAB (Amsterdam)

1.1.3 Useful Journals and Technical Publications

- IEEE Transactions on Image Processing, Circuits and Systems for Video Technology, Multimedia, Pattern Analysis and Machine Intelligence, Signal Processing, Signal Processing Letters
- Computer Vision and Image Processing (CVGIP) (Academic Press)
- EURASIP Signal Processing
- EURASIP Image Communication

2 A start and Terminology

Digital pictures are made up of individual **pic**ture **el**ement**s** called *Pixels*. In this course we will abbreviate this to *pel*. A digital image is typically made up of a number of lines each

containing a number of pixels. A PAL (European) television frame consists of 576 lines each having 720 pixels.

In *greyscale* images, at each pixel site, the intensity of a pel is given by some number, typically 8 bit for broadcast applications and 12 bit or more for medical applications. Thus 0 is black and 255 is bright white.

An image is a 2D function of space, and a Digital image can be represented as a matrix of numbers. The coordinate of a particular site will be denoted $[h, k]$ where h is the line number and k is the pixel in that line. This is the usual Matlab notation for 2D matrices. The intensity of a pixel in a grey scale image at a site $[h, k]$ for instance will be defined as $I(\mathbf{x})$ where $\mathbf{x} = [h, k]$ (a position vector). Thus if the intensity at $\mathbf{x}_1 = [5, 4]$ is 156, then $I(\mathbf{x}_1) = 156$.

2.1 Matlab

In matlab you will be using images as matrices. Remember that if `A` is an image (matrix) of size 576×720 then `A(:,20)` is a column vector with 576 elements consisting of the 576 pixels along the 20th column, and `A(100,:)` is a row vector with 720 elements consisting of the 720 pixels along the 100th row.

Colour images are stored as 3 planes of data `A(:,:,3)`. Get used to converting from the `uchar` default data type (8bits) for images read in with `imread()` to `double` data type for doing arithmetic on the pictures.

When you can't remember a command call in Matlab, use `help <command name>`.

3 The Human Visual System (HVS) and Perception

Understanding the importance of human perception of picture material is crucial since the result of any processing is to be presented to a human and that entity is decidedly non-linear in its perception of light and shade. In fact it is through the perceptual characteristics of the HVS that compression schemes can achieve much of their performance. Furthermore, all image communication industries are necessarily obsessed with maintaining image quality. By modelling the HVS quantitatively, it is possible to design objective models which assess the perceptual quality of processed or received images thus allowing a DTV broadcaster for instance to verify automatically that the transmitted images are achieving the quality required by the governing body for the media. Issues like Quality of Service (QoS) over video communication networks can be handled much more readily if such quantitative mechanisms

for assessing picture quality were available.

Work in the measurement of the HVS has been conducted by Visual Psychologists since the 1920's [3] and it was in the early 1970's that Netravali [4] and others began to exploit these ideas in the design of image compression schemes. There are a few broad ideas which deserve consideration Contrast/Intensity Sensitivity, Frequency Sensitivity and Perceptual Masking.

First of all it should be noted that the measurement of the HVS response to different stimuli is *notoriously* difficult[4]. This is because it is difficult to generate the conditions under which the human subject is *consistently* able to separate his higher level visual processing (e.g. "That's a computer monitor I'm looking at", "Who cares about that spot anyway?", "I know I am looking for circles so let's just spot circles.') from the low level image processing which goes on inside the visual cortex (e.g. spatial frequency response cells, colour receptors etc). Lighting conditions and the *luminance* (in Candela/m^2) function of the screen used (emitted light energy as a function of greyscale) are crucial and must be strictly controlled.

3.1 Contrast/Intensity Sensitivity

Weber's law relates the perceived brightness of an object to the brightness of its background. The law can be derived by measuring the 'Just Noticeable Difference' between two visual stimuli. A psychovisual experiment to measure this is as shown in figure 1. Consider a foreground object with intensity I_f and a background with intensity I_b. We can define the threshold ΔI as the intensity difference $\Delta I = I_f - I_b$ at which the foreground object is *just* visible (with a 50% probability) by a human. Weber's law states that $\Delta I/I = k$ where k is a constant (about 0.02). For a relatively large range of intensity this law holds and implies that for a bright background a large intensity difference is needed to resolve a foreground object than if the background were dark.

Practical use for this idea can be found in quantisation for compression and noise reduction for restoration. Weber's law implies that noise is less visible in the bright regions than in the dark regions of a picture. Thus less noise reduction (coarser quantization) can be tolerated in bright regions than in dark regions.

3.2 Spatial Frequency Sensitivity

Figure 2 shows a staircase image intensity profile. The greyscale values are constant in bands across the image. However, the HVS perceives that each vertical stripe looks brighter on the left and darker on the right. This is an effect known as 'Mach banding'. It is as a direct

[4]And therefore extremely interesting

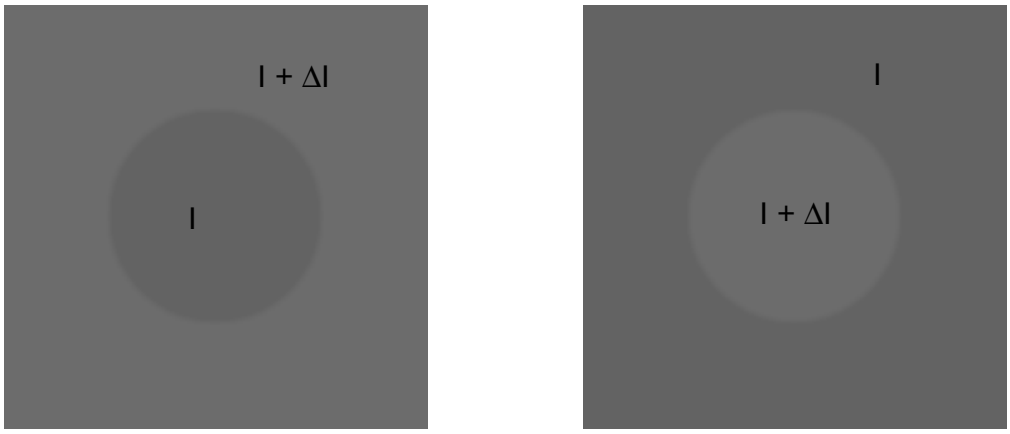

Figure 1: The setup for a psychovisual experiment to test Weber's law.

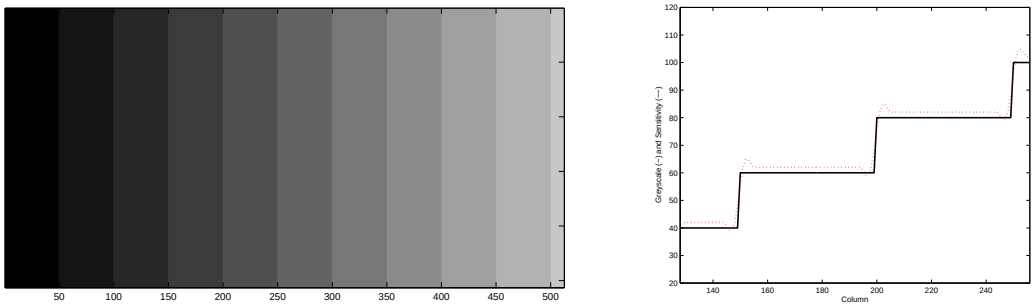

Figure 2: Left: Staircase pattern of greyscale showing the Mach banding effect. Right: Cross section of the greyscale for a subset of columns (black), and the result of filtering with a low pass filter having a SINC response (red). The low pass filter effect approximates the HVS response: each flat band is perceived due to the effect as if the left hand side were brighter than the right hand side. *Note that the filter response has been slightly shifted vertically so that it is more easily viewed.*

result of spatial filtering in the visual cortex. This can be partially understood through the effect of a spatial filter on the image (also shown in figure 2). In the horizontal direction a low pass filter (with a symmetric impulse response) will cause undershoot and overshoot at sharp edges, hence a distortion of the perceptual quality of the edge.

3.2.1 The meaning of spatial frequency

It will become necessary to understand what spatial frequency means from the HVS perspective. The units are in *cycles per degree* or *cycles per radian*. The underlying understanding is that pictures are being formed on the retina of the eye, and it is there that the visual processing begins. Thus frequencies should be measured in terms of their relationship to the

degrees of arc subtended by one luminance cycle on the retina.

CCIR[5] recommendations 500 require viewing at 5 times the screen height. At this distance the viewable height of a monitor h would subtend $\tan^{-1}(1/5) = 0.1974$ rad or 11.3 degrees. Consider a monitor set at 1024×768 pixel resolution. This implies that the resolution of a pixel on the retina is about $\theta = 11.3/768 = 0.0147$ degrees. Thus the sampling frequency of the screen is $1/\theta = 68$ cycles/degree (pixels/degree). A vertical sinusoidal grating having a frequency of .05 cycles per pixel on the screen would therefore have a frequency of $1/(0.05 \times 0.0147) = 0.3$ cycles per degree on the retina.

It is the cycles per degree which matter for the HVS since it is *at the retina* that processing starts. This explains why the perception of images depends on viewing distance and thus why International Committees define the viewing distance for regulating the measurement of picture quality.

3.2.2 Frequency Response of the HVS

The frequency response of the HVS is shown in figure 3. It shows the sensitivity of the HVS to stimuli with varying frequencies. Essentially there is a bandpass behaviour, the HVS is more sensitive to midfrequencies than low or high frequencies and is least sensitive to high frequencies.

The effect is demonstrated in figure 3. A sinusoidal grating is shown which increases with frequency horizontally and decreases in intensity vertically. The visibility of j.n.d. boundary between the bright and dark vertical lines directly measure the reader's Modulation Transfer Function (modulated by the printing/photocopying process.). If there was no dependence on spatial frequency the boundary between visibility and non-visibility would be a straight line.

Note that the HVS sensitivity is orientation dependent with maximum sensitivity for vertical and horizontal orientations. Other orientations are at most 3dB off the peak hence the frequency response can be approximated as isotropic.

Important features are

- Maximum frequency sensitivity occurs at around 5 cycles/degree. This corresponds to striped patterns with a period of about 1.8 mm at a distance of 1 m ($\sim$ arm's length)
- There is very little response above 100 cycles per degree which corresponds to a stripe width of 0.17 mm at 1m. This implies about 1800 pels per line on a computer display. SVGA of 1024×768 is a little more than half of this. Laptop displays have a pel size

[5] A regulatory body

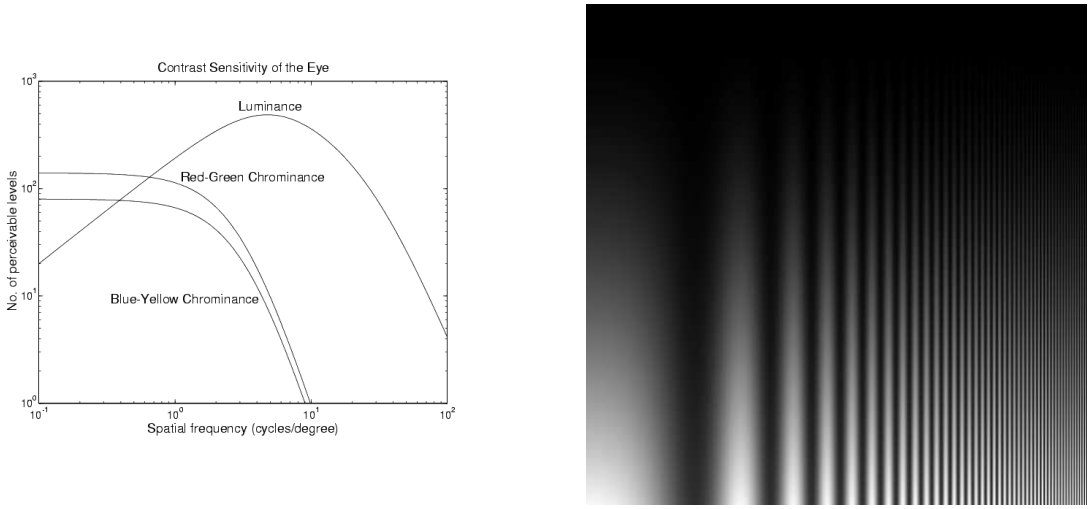

Figure 3: Left: Contrast sensitivity of the HVS. Horizontal axis represents the frequency of an alternating pattern of parallel stripes with sinusoidally varying intensity. The vertical scale shows the contrast sensitivity of the HVS which is the ratio of the maximum visible range of intensities to the minimum discernible peak-to-peak intensity variation at the specified frequency. Right: Sinusoidal grating for demonstrating HVS Frequency response.

of 0.3 mm but are pleasing to view because of the sharpness of the pixels and the lack of flicker.

- Sensitivity to luminance drops off at low spatial frequencies, showing that we are not good at estimating absolute luminance levels *as long as they do not change with time*. The luminance sensitivity to temporal fluctuations (flicker) dos not fall off at low spatial frequencies.

3.3 Colour

The HVS perceives colour using receptors (cones) in the retina which correspond to three broad colour channels in the region of red, green and blue. [ROYGBIV] ([5] page 61).

Other colours are perceived as combinations of RGB and thus monitors use RGB to form almost any perceivable colour by controlling the relative intensities of R, G, and B light sources. Therefore electronic representation of colour images require the representation of three intensities (R G B) at each pixel site.

The numerical values used for these intensities are usually chosen such that equal increments in value result in approximately equal apparent increases in brightness. In practise this means that the numerical value is approximately proportional to the log of the true light

intensity (energy of the wave). This is another statement of Weber's law. These numerical values will be referred to as intensities in the rest of the course as it is convenient to refer to a subjectively linear scale.

3.3.1 Alternative Colour Spaces: YUV

The eye is much more sensitive to luminance (brightness) than to colour changes. Usually even if we remove all the colour from a picture, it is still intelligible using greyscale only. Thus black and white TV was acceptable for a long time until colour technology became sufficiently cheap (1952). The luminance of a pel Y may be calculated as

$$Y = 0.3R + 0.6G + 0.1B \tag{1}$$

These are only approximate values, and different references quote slightly different values e.g. $0.299R + 0.587G + 0.114B$.

The YUV transformation mapping was used in the 1950's so that those with Black and White TV sets could still view colour TV signals.

RGB representations are usually defined so that if $R = G = B$ the pel is some shade of grey. Thus if $Y = R = G = B$ in these cases, the coefficients used in equation 1 should sum to unity.

The chrominance of a pel is defined by U and V as below (for PAL[6]).

$$\begin{aligned} U &= 0.5(B - Y) \\ V &= 0.625(R - Y) \end{aligned} \tag{2}$$

Grey pels will always have $U = V = 0$

The transformation between RGB and YUV colour spaces is linear and can be expressed as follows

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \mathbf{C} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

$$\text{Where} \quad \mathbf{C} = \begin{bmatrix} 0.3 & 0.6 & 0.1 \\ -0.15 & -0.3 & 0.45 \\ 0.4375 & -0.3750 & -0.0625 \end{bmatrix} \tag{3}$$

[6]Phase Alternate Line: the European Colour TV format

The inverse relationship is derived by taking the inverse of the transformation matrix $\mathbf{C}$ on the right of equation 3 to give

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \mathbf{C}^{-1} \begin{bmatrix} Y \\ U \\ V \end{bmatrix}$$

$$\text{Where} \quad \mathbf{C}^{-1} = \begin{bmatrix} 1 & 0 & 1.6 \\ 1 & -0.3333 & -0.8 \\ 1 & 2 & 0 \end{bmatrix} \tag{4}$$

The allowed range of Y is from 16 to $255 - 16$, U, V range between ± 128 and so are shifted by 128 to allow storage as an 8 bit number.

3.3.2 YIQ, YDbDr Colour Space

The US standard for video NTSC (National Television Systems Committee)[7] uses a slightly different transformation into a space known as YIQ in which

$$\mathbf{C} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.523 & -0.311 \end{bmatrix} \tag{5}$$

$[Y\ Db\ Dr]$ is used for Secam (Sequential Couleur a Memoire) which is employed in France, Russia. Here $Db = 3.059U$ and $Dr = -2.169V$.

3.3.3 HSV Colour Space

The YUV representation is used for Digital TV and Digital Media communications in general, however it does not create a fully *orthogonal* colour space w.r.t. perception. A more useful colour space to operate in from the point of view of perception in particular is the HSV : Hue Saturation and Intensity (or Value) space. This space describes a colour in terms of its Brightness (Intensity or Value), its Hue (its 'redness', 'greenness' etc) and its Saturation (deep red, light red etc). The difference between HSI and HSV colorspaces is the intensity calculation.

Whereas the RGB colour space can be thought of as a cube with the three principal axes being R,G,B; the HSV space is a cone with the vertical axis being Brightness (I), Saturation (S) varies with the radius, and Hue (H) varies around the circumference.

[7] Otherwise known as *Never The Same Colour*

For HSV, the transformation from RGB is more complicated than for YUV. Hue has a range for 0 to 360 degrees, S has a range from 0 to 1 and V has a range from 0 to 1. Matlab has an HSV2RGB function which you can experiment with, and the lecture will include a demonstration using color wheels for which the 'm' script is available.

See [1, 5, 6] for more information.

3.4 Colour Sensitivity

Figure 3 also shows the sensitivity of the eye to Chrominance (U,V) components.

- The maximum chrominance sensitivity is much lower than the maximum luminance sensitivity.
- The chrominance sensitivities fall off above 1 cycle/ degree thus requiring a much lower spatial bandwidth than luminance.

This illustrates why it is better to use the YUV domain for video communication since the eye is less sensitive to U and V. Hence the U and V components may be sampled at a lower rate than Y and may be quantised more coarsely. U and V also tend to be much more 'smooth' functions than Y and this adds to their compressibility.

3.5 Activity Masking

In general the sensitivity of the HVS to any image feature changes depending on the *character* of the background. Therefore edges for instance are less visible in textured areas than against a uniform background. Thus the contrast sensitivity to a given pattern is reduced by the presence of other patterns in the same region. This is called *activity masking*.

This is a complicated subject and the masking effect depends on the similarity between the given pattern and the background activity. In general, however, the higher the variance of pels in a small region (typically 8 to 16 pels across) the lower the contrast sensitivity.

Thus compression schemes which adapt quantisation to local image activity achieve better compression rates than those using uniform quantisation. Furthermore image enhancement techniques which allow for masking effects generally show less visible damage to the picture than would otherwise be the case.

A good example of this is in noise reduction. Typically, noise reduction causes blurring of edges and good noise reducers achieve a compromise between this blurring and reducing

the noise level. However, by allowing for the masking effect of noise by edges and texture, more noise can be left in textured and edge regions than in other areas of the image yielding a noise reduced, but *perceptually* sharper image.

However, it is *very* difficult to allow for masking effects quantitatively in image analysis/enhancement tasks as the perceptual distortion measure can complicate the analysis and implementation.

4 A Note on Compression

Image data is typically of a high bandwidth. A typical raw PAL TV signal will require 20 MB/sec for transmission. A single Digital Cinema format frame is 7MB large! The task of image compression is to reduce the amount of data used to store an image *without* objectionable degradation in the percieved quality of the image. The term *objectionable* depends of course on the use of the image data. Sometimes fast/real time transmission is more important than image quality e.g. video over wireless. Other times quality is paramount as in Digital Cinema and DTV. To make matters more complicated the same image shown in different formats looks quite different e.g. when one converts from one DTV format to another, degradations in the conversion do not show up on television sets, but if the same image is used for creating a Digital Cinema production [film format ads created from TV ads] then the degrdations are very noticeable.

Image compression is of vital importance for communications and media storage. There is a growing demand for IC designs that implement the core elements of image compression systems, in particular for low power and low memory applications like PDAs and Mobile phones.

Image compression is possible because of the nature of images and the development of efficient coding techniques. The three enabling ideas are listed as follows.

1. There is alot of **statistical redundancy** in images. For instance, in local image regions say 8×8 blocks, the data tends to be 'flat' or typically homogenous much of the time. This redundancy can be removed without affecting the image substantially, thus reducing the amount of data to be stored.

2. The **HVS response** to image stimuli implies that one can introduce artefacts into images *without* them being seen. The colour demonstration accompanying the lecture illustrated this idea with colour subsampling. Thus techniques that remove statistical redundancy can apply that concept heavily in regions where the resulting defects will not be noticed. This further reduces the image data to be stored.

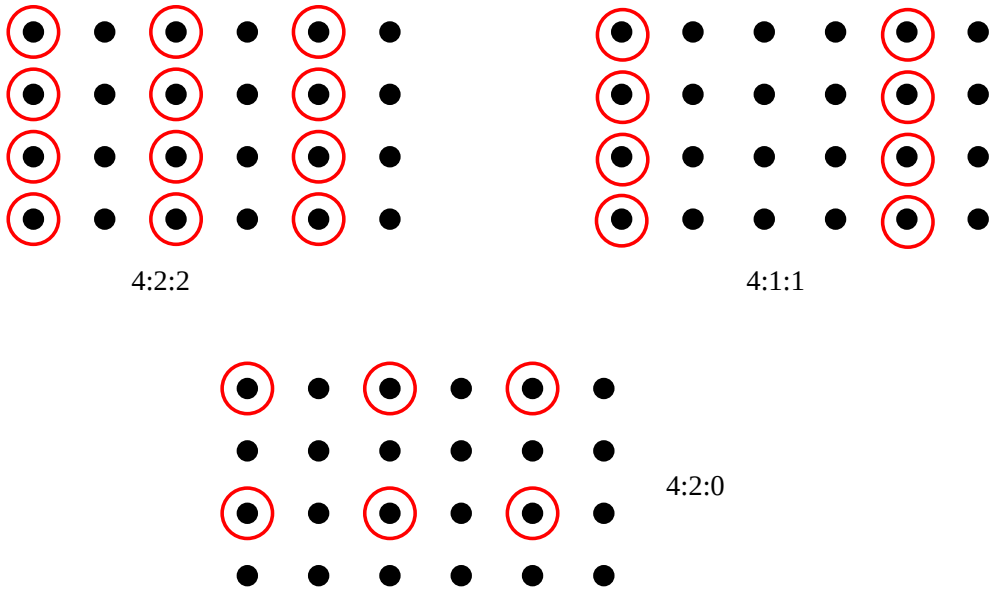

Figure 4: Sampling structures used for colour digital video formats. Luminance samples are black dots, colour samples are indicated by red circles.

3. **Efficient coding techniques** can be used to represent any data as a more compact stream of digits. This technology can be used both for compression and *error-resilience*.

The course will elaborate on image compression technology after some fundamental image processing concepts are introduced.

5 Impact on Digital Picture Formats

In a digital picture file, picture information is sampled spatially and perhaps temporally and also the resulting pixel intensities are quantised. CCD imagers provide a natural mechanism for spatial and temporal sampling and it is typical to use 8 bit quantisation in all digital video data formats. Higher sampling rates and smaller quantisation steps are used for medical images in which 12 bit quantisation is sometimes used. The Film and HD Broadcast markets are beginning now to use 16bit quantisation.

Digital video data is represented as three separate component data streams: RGB or YUV. Because of the subjectively lower sensitivity of humans to colour, colour information is typically sampled at a lower rate than the intensity information. When the colour information is downsampled by a factor of 2 *horizontally* from the full resolution intensity image, the picture sampling structure is called 4:2:2. When the colour information is sampled by a factor of 2 *horizontally and vertically* the sampling is called 4:2:0. The 4:4:4 sampling structure

represents video in which the colour components of the signal are sampled at the same rate as the luminance signal. 4:1:1 sampling yields 1 colour sample for every 4 horizontal luminance samples. Figure 4 shows the spatial arrangement of these sampling structures.

There are several digital video formats defined by the CCIR Recommendation 601. These are indicated in the table below.

Format	Total Resolution	Active Resolution	MB/sec
CCIR 601 30 frames/sec, 4:3 Aspect Ratio, 4:2:2, NTSC			
QCIF	214×131	176×120	1.27
CIF	429×262	352×240	5.07
Full	858×525	720×485	20.95
CCIR 601 25 frames/sec, 4:3 Aspect Ratio, 4:2:2, PAL			
QCIF	216×156	176×144	1.27
CIF	432×312	352×288	5.07
Full	864×625	720×576	20.74

The CIF and QCIF formats are approximately $2:1$ and $4:1$ downsampled (in both horizontal and vertical) directions from the full resolution picture sizes. Note that despite the differences in picture resolution between NTSC and PAL, the data bandwidth is approximately the same. This is due to the difference in frame rates. In modern mobile phones, the CMOS imagers are increasingly sophisticated and will soon capture 640×480 pictures. The LG ViewTY already does 120 frames per second (fps)! Currently most video over mobile phones is in QCIF format. CIF is the format used for video taken with still picture cameras, while DV Camcorders operate at Full resolution. See Figure 5 for a visual appreciation of the relative sizes of the pictures.

The data bandwidth required for these video streams are all on the order of MB/sec. Thus the successful transmission of digital video relies heavily on compression mechanisms and ultimately on the standardisation of compression schemes. This is what the MPEG committees worked on from about 1985-2000. A discussion of digital video formats is therefore not complete without a discussion of digital video compression standards, however this 'complicated' topic is left for the end of this series of lectures.

6 High Definition (HD) Formats

There are two principal formats 1920×1080 and 1280×720 both at 16:9 aspect ratio. These come in different flavours : 1080i, 1080p, 720i, 720p. The suffixes stand for *interlaced* (i) and

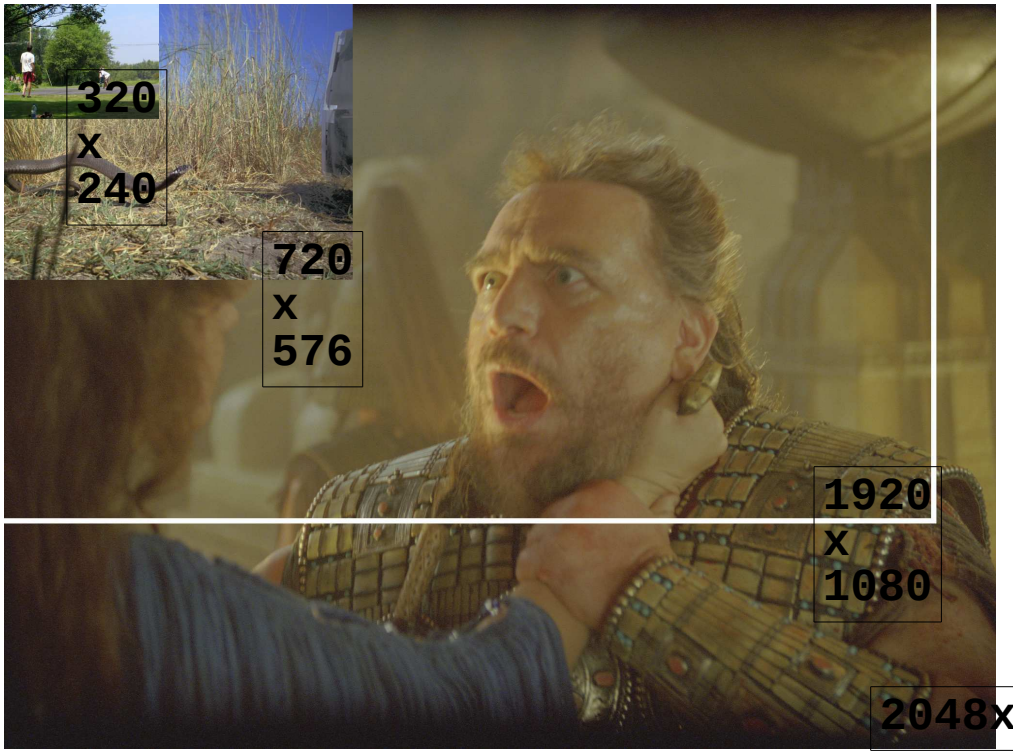

Figure 5: From small to large : Approximately CIF, SD (PAL), HD (1080p) and Digital Film Picture Sizes.

progressive (p) and refer to the scan pattern used in recording and displaying the picture. See Figure 7 and 6 for interlaced and progressive scan patterns. A more detailed discussion about interlaced and progressive video will emerge later in the course. 720p is brodcast at 23.976, 24, 29.97, 30, 59.94, 60 fps while 1080p is broadcast at 23.976, 24, 29.97, 30 fps (in theory). This is why an HD set top box is different from an SD set to pbox.

See http://www.microsoft.com/windows/windowsmedia/howto/articles/UnderstandingHDFormats.asp for more information.

7 Image Evaluation

The idea of any kind of signal processing is to achieve some kind of desired effect on the percieved information content in the signal. In image compression, the requirement is to reduce the amount of data required to represent the image *without* significant degradation. In image enhancement and restoration the idea is to process the data to improve illumination, or to remove defects like noise or Dirt on a film. A mechanism for assessing the *damage* done to the observed picture is important to evaluate the quality of the processed output.

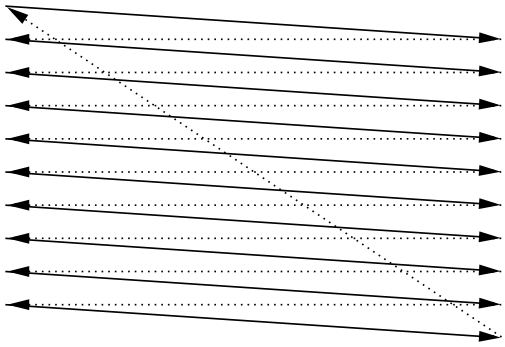

Figure 6: Progressive (Sequential) TV scanning

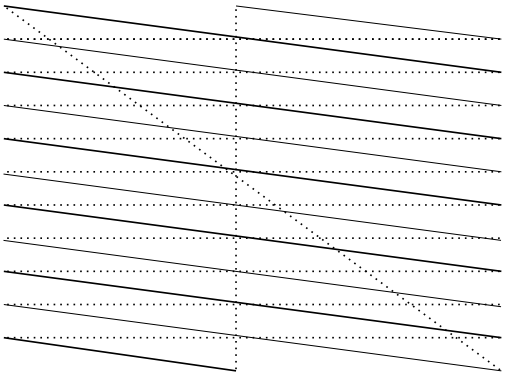

Figure 7: Interlaced TV Scanning

7.1 Subjective Assessment

The CCIR[8] recommendations 500-3 issued in 1986 a scale for subjective measurement adapted from early work by Bell Labs.

1. Impairment is not noticeable
2. Impairment is just noticeable
3. Impairment is definitely noticeable but not objectionable
4. Impairment is objectionable
5. Impairment is extremely objectionable

Using this system, human subjects are asked to rank observed (processed) images in terms of this 5 point scale. Collating results from such experiments gives some quantitative measurement of the subjective perception of the images.

This process is tedious and difficult, requiring many participants and many reformulations of the order in which the images are presented.

Ideally some automated mechanism for assigning a *numerical* value to the percieved quality of the image is required. But as the above discussions on the HVS show, it is diffcult to quantify all the aspects of visual perception. A good recent effort at doing this was presented in 1993 [7].

[8]Now ITU: international Telecommunications Union

7.2 Objective Assessment

Letting the desired image be $I(h,k)$ and the result of processing be the image $\hat{I}(h,k)$, it is possible to employ simple *error* based schemes as an alternative to subjective evaluation. Thus defining an *error* as

$$e(h,k) = \hat{I}(h,k) - I(h,k) \tag{6}$$

allows the derivation of some simple measures for how close the processed image is to some 'ideal' $I(h,k)$.

In a compression example for instance, $\hat{I}(h,k)$ would be the image resulting after the original $I(h,k)$ is compressed and then decoded. Lossless compression would imply that $e(h,k)$ is 0 everywhere in the image, and lossy compression would imply that $e(h,k)$ was non-zero.

In image restoration, we do not kow $I(h,k)$ and the idea is to process some observed image to yield $\hat{I}(h,k)$. The question then is "How good is the restoration really?" Of course we do not have $I(h,k)$ in the first place, but that discussion is left for another time[9].

A number of simple measures can be generated to measure the 'observed error' as follows.

The *Mean Squared Error* (MSE) is

$$\text{MSE} = \frac{1}{NM}\sum_{\mathbf{x}} e(\mathbf{x})^2 \tag{7}$$

where the image size is N rows by M columns, the sum is over all the sites in the image, and $e(\mathbf{x}) = e([h,k])$. Thus the MSE is the mean of the squared error values across the entire image. In some cases, the **Mean Absolute Error** is used as follows.

$$\text{MAE} = \frac{1}{NM}\sum_{\mathbf{x}} |e(\mathbf{x})| \tag{8}$$

The **Signal to Noise** ratio is another popular objective measure and it has units of Decibels (dB).

$$\text{SNR} = 10\log 10\frac{\frac{1}{NM}\sum_{\mathbf{x}} I(\mathbf{x})^2}{\text{MSE}} \tag{9}$$

This is a ratio between the signal power, measured as the sum squared intensities in the original image I, and the 'noise' power measured as the MSE of the error, e.

A more popular version of the SNR used widely in image compression is the **Peak SNR** this is the log of the ratio between the peak signal (image) power and the noise power. It is also measured in dB.

$$\text{PSNR} = 10\log 10\frac{255^2}{\text{MSE}} \tag{10}$$

[9]The restoration problem is called ill-posed for this reason.

where it is assumed that the image is stored as 8 bit pixels in which case the peak value at a pixel site is 255.

Unfortunately, these error measures do not align well with the Human perception of images. It is a good rule of thumb however that in comparing images using these measures, large differences in objective measurements do tend to imply similar differences in human perception of the images. However when the differences between the same objective measure on several images are small then it is no longer the case that the differences would imply similar perceptual evaluations. A demonstration during the lecture will explain this more clearly.

8 Summary

This handout has covered the following ideas

1. Digital Image processing is more relevant with the increasing use of Digital Visual Media
2. Images are 2-D functions of space
3. Human visual perception is important for understanding what effects of a processing system will be visible.
4. The perception of image features depends on their brightness, frequency and the masking effects of the features nearby.
5. There are several different colour spaces for representing a colour digitally. YUV and HSV have applications in DTV and Image Analysis respectively
6. The HVS is less sensitive to colour than brightness (luminance)
7. Perceptual masking is one key to understanding why image compression is possible
8. Automated picture quality assessment is growing in importance. This is difficult because it is difficult to model the HVS and so assign to any arbitrary image an absolute measure of 'quality'.
9. CCIR Rec 500 proposes a 5 point subjective evaluation scheme which is one of the standards adopted by the DTV industry.
10. Objective image evaluation is convenient and MSE, MAE, SNR, PSNR are all used to assess the performance of image processing systems. It is understood that the comparisions using these measures may not follow human perception of the same images.

References

[1] Jae S. Lim. *Two-Dimensional Signal and Image Processing*. Prentice-Hall, 1990.

[2] A. Murat Tekalp. *Digital Video Processing*. Prentice Hall, 1995.

[3] S. Hecht. The visual discrimination of intensity and the weber-fechner law. *General Phisiology*, 1924.

[4] A. N. Netravali and B. Prasada. Adaptive quantisation of picture signals using spatial masking. In *Processings of the IEEE*, pages 536–548, April 1977.

[5] A.K. Jain. *Fundamentals of Digital Image Processing*. Prentice Hall, 1989.

[6] Keith Jack. *Video Demystified*. Hightext, 1993.

[7] A. A. Webster, C. T. Jones, and M. H. Pinson. An objective video quality assessment system based on human perception. In *Proceedings of Human Vision, Visual Processing and Digital Display IV*, volume 1913, pages 15–26. SPIE, 1993.