Media Processing

DSP and its application to Speech and Image Processing

Sunil Kumar Kopparapu

SunilKumar.Kopparapu@TCS.COM

Speech and Natural Language Processing Group

TCS Innovation Labs - Mumbai

Advanced Technology Applications Group,

Yantra Park, Thane (West), Maharastra

July 2007

What is Media?

What is Media?

- What is Media?
  - A plural of medium [Dictionary]
  - Formats for **presenting information** [Wiki]
  - a means or instrumentality for *storing* or *communicating information* [WordNet]

What is Media?

- What is Media?
  - A plural of medium [Dictionary]
  - Formats for **presenting information** [Wiki]
  - a means or instrumentality for *storing* or *communicating information* [WordNet]
- Information is the key element
- Media can be loosely defined as any information in electronic form

What is Media?

- What is Media?
  - A plural of medium [Dictionary]
  - Formats for **presenting information** [Wiki]
  - a means or instrumentality for *storing* or *communicating information* [WordNet]
- Information is the key element
- Media can be loosely defined as any information in electronic form

So, what holds information?

Information in Media

- What are the different media that hold information?
  - text (including on-line script)
  - graphics (the digital representation of an *imaginary scene*)
  - audio / sounds
  - images (the digital representation of a *real scene*)
  - videos (moving images or graphics)

Information in Media

- What are the different media that hold information?
  - text (including on-line script)
  - graphics (the digital representation of an *imaginary scene*)
  - audio / sounds
  - images (the digital representation of a *real scene*)
  - videos (moving images or graphics)
- Multimedia ....
  - the ability to combine text, graphics, audio, and (moving) images in meaningful ways.
  - Probably this is one of the powerful aspect of computing technology

Multimedia - Components, Aspects

- Components
  - text
  - audio
  - images (sequence of images - video)

Multimedia - Components, Aspects

- Components
  - text
  - audio
  - images (sequence of images - video)
- Some Aspects
  - Storage / Compression (mp3, jpg, mpeg)
  - Transmission (codec, mpeg4)
  - Multimedia signal processing (including annotation)
  - Search / Retrieval

Aspects in Media Processing

- Multimedia Analysis, Processing, and Retrieval

Aspects in Media Processing

- Multimedia Analysis, Processing, and Retrieval
- Signal Processing for Media Applications

Aspects in Media Processing

- Multimedia Analysis, Processing, and Retrieval
- Signal Processing for Media Applications
- Image and Video Processing, Speech, Audio and Music Processing

Aspects in Media Processing

- Multimedia Analysis, Processing, and Retrieval
- Signal Processing for Media Applications
- Image and Video Processing, Speech, Audio and Music Processing
- Real-time Multimedia

Aspects in Media Processing

- Multimedia Analysis, Processing, and Retrieval
- Signal Processing for Media Applications
- Image and Video Processing, Speech, Audio and Music Processing
- Real-time Multimedia
- Interactive Media and Games, 3D-TV, Stereo Systems

Aspects in Media Processing

- Multimedia Analysis, Processing, and Retrieval
- Signal Processing for Media Applications
- Image and Video Processing, Speech, Audio and Music Processing
- Real-time Multimedia
- Interactive Media and Games, 3D-TV, Stereo Systems
- Audio/video Streaming

Aspects in Media Processing

- Multimedia Analysis, Processing, and Retrieval
- Signal Processing for Media Applications
- Image and Video Processing, Speech, Audio and Music Processing
- Real-time Multimedia
- Interactive Media and Games, 3D-TV, Stereo Systems
- Audio/video Streaming
- Media Content Distribution, Wireless Multimedia

Some Trend in Multimedia

The next wave of multimedia could be in terms of creating new approaches for the

- acquisition,

Some Trend in Multimedia

The next wave of multimedia could be in terms of creating new approaches for the

- acquisition,
- *processing*, and

Some Trend in Multimedia

The next wave of multimedia could be in terms of creating new approaches for the

- acquisition,
- *processing*, and
- display of new types of image

Some Trend in Multimedia

The next wave of multimedia could be in terms of creating new approaches for the

- acquisition,
- *processing*, and
- display of new types of image

The ultimate goal is to

Some Trend in Multimedia

The next wave of multimedia could be in terms of creating new approaches for the

- acquisition,
- *processing*, and
- display of new types of image

The ultimate goal is to

- change the way visual information is captured from real scenes and

Some Trend in Multimedia

The next wave of multimedia could be in terms of creating new approaches for the

- acquisition,
- *processing*, and
- display of new types of image

The ultimate goal is to

- change the way visual information is captured from real scenes and
- presented to the human observer

Some Trend in Multimedia

The next wave of multimedia could be in terms of creating new approaches for the

- acquisition,
- *processing*, and
- display of new types of image

The ultimate goal is to

- change the way visual information is captured from real scenes and
- presented to the human observer

Achieved through media processing

Example: Trends in Multimedia

- depth-perception (stereo image processing),

Example: Trends in Multimedia

- depth-perception (stereo image processing),
- visualization without aids (eg. virtual reality glasses),

Example: Trends in Multimedia

- depth-perception (stereo image processing),
- visualization without aids (eg. virtual reality glasses),
- offering the viewer the capacity to visually *fly around* an object or a scene (representation)

Example: Trends in Multimedia

- depth-perception (stereo image processing),
- visualization without aids (eg. virtual reality glasses),
- offering the viewer the capacity to visually *fly around* an object or a scene (representation)
- view a complete sphere of view at multiple points in a natural scene, while preserving photo-realistic quality

Example: Trends in Multimedia

- depth-perception (stereo image processing),
- visualization without aids (eg. virtual reality glasses),
- offering the viewer the capacity to visually *fly around* an object or a scene (representation)
- view a complete sphere of view at multiple points in a natural scene, while preserving photo-realistic quality

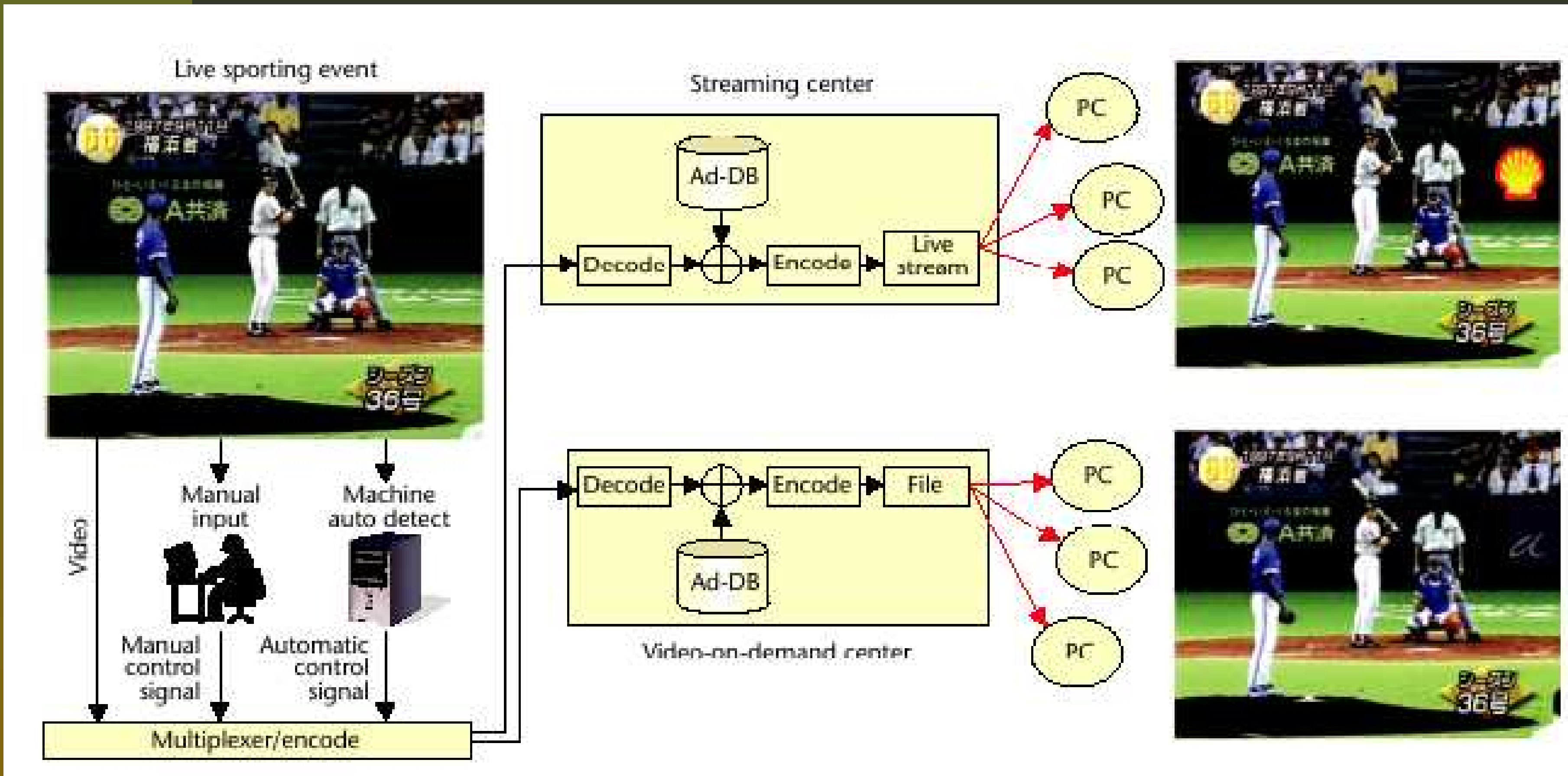
Achieving this goal will rely on processing visual information generated by using input images captured from *real* scenes

Media Processing at Work

Advertising Insertion in Sports Webcasts (IEEE Multimedia)

Media Processing at Work

Advertising Insertion in Sports Webcasts (IEEE Multimedia)



Advertising Insertion in Webcasts

- Webcasts can reach a demographic segment.

Advertising Insertion in Webcasts

- Webcasts can reach a demographic segment.
- Webcasts combine traditional media's familiarity with the Internet's one-to-one interactivity.

Advertising Insertion in Webcasts

- Webcasts can reach a demographic segment.
- Webcasts combine traditional media's familiarity with the Internet's one-to-one interactivity.
- Webcast viewers are likely to be more tolerant of extraneous video effects such as advertising insertions than TV viewers because webcasts' video quality is generally lower than TV's.

Advertising Insertion in Webcasts

- Webcasts can reach a demographic segment.
- Webcasts combine traditional media’s familiarity with the Internet’s one-to-one interactivity.
- Webcast viewers are likely to be more tolerant of extraneous video effects such as advertising insertions than TV viewers because webcasts’ video quality is generally lower than TV’s.
- Webcast audiences are generally more technologically savvy, affluent, and likely to spend money on advertised items than TV audiences.

Multimedia at Work

- Multimedia Video Blog (Hosted on IEEE Multimedia website)

Multimedia at Work

- Multimedia Video Blog (Hosted on IEEE Multimedia website)

- New mode of knowledge distribution

Multimedia at Work

- Multimedia Video Blog (Hosted on IEEE Multimedia website)

- New mode of knowledge distribution

How do we access this kind of knowledge?

What Media? (1)

- Speech
  - **non-linguistic**, (*who said it*)
    gender, emotional states, speaker name
  - **linguistic** (*what he said*)
    Language name and what was said (written language)
  - **paralinguistic** (*how well said* – manner, clarity or accent, aspects related to quality)
    deliberately added by the speaker, and not inferable from the written text.

Goal: Automatically extract information in speech signal

What Media? (2)

- Image
  - An image is a *digital representation* of a *real-world scene*.
  - Composed of discrete elements called picture element (pixels)
  - Pixels are parametrized by
    - position
    - intensity
    - time
  - These parameters define (a) still images, (b) video, (c) volume data and (d) moving volumes

Image Processing

Digital Photographs

- Two spatial parameters
  - x, or horizontal position
  - y, or vertical position
- Three intensity parameters
  - Red
  - Green
  - Blue

Digital Photographs

- Two spatial parameters
  - x, or horizontal position
  - y, or vertical position
- Three intensity parameters
  - Red
  - Green
  - Blue

Ultrasound

- Two spatial parameters - x and y
- One intensity parameter - ultrasound reflection
- One time parameter (ultrasound printouts do not show this, but the exam does)

Ultrasound

- Two spatial parameters - x and y
- One intensity parameter - ultrasound reflection
- One time parameter (ultrasound printouts do not show this, but the exam does)

1

Digital X-Ray

- Two spatial parameters x and y
- A single intensity parameter - attenuation of x-ray

Digital X-Ray

- Two spatial parameters x and y
- A single intensity parameter - attenuation of x-ray

1

Digital Video

- Two spatial parameters
  - x, or horizontal position
  - y, or vertical position
- Three intensity parameters
  - Red
  - Green
  - Blue
- One time parameter - frame

Other types of images

- Thermal image - 2D image based on temperature

Other types of images

- Thermal image - 2D image based on temperature

1

Other types of images

- Thermal image - 2D image based on temperature
- Satellite Aperture Radar (SAR) image
- Computed Tomography
  3 dimensional x-ray images of the human body
- Range image
  captures the depth of the object from camera
- Functional Magnetic Resonance Images
  3 dimensional images of the human body over time
- Speech Spectrogram
  3 dimensional image of acoustic signal

What is Image Processing?

- Image processing is a form of information manipulation for which the input is an image
- Most image processing techniques involve treating the image as a 2D signal and applying standard signal processing techniques to it.
- Categories
  - **Image Processing**
    Output is an image
  - **Image Analysis**
    Output is a set of measurements
  - **Image Understanding**
    Output is a high-level description of the image

Example of Image Processing

Edge Detection 1, contrast enhancement, segmentation, restoration

Example of Image Analysis

A task as simple as **reading bar coded tags** or as sophisticated as **identifying a person from their face**

Input	Output
	Wikipedia
	Sunil

Example of Image Understanding

Understanding usually attempts to mimic the human visual system in extracting meaning from an image

Example of Image Understanding

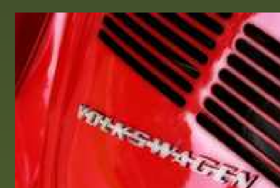
Understanding usually attempts to mimic the human visual system in extracting meaning from an image Given a reason for looking at a particular scene, understanding systems produce descriptions of (a) the image and (b) the world scene that the image represent. 1

Example of Image Understanding

Understanding usually attempts to mimic the human visual system in extracting meaning from an image

Given a reason for looking at a particular scene, understanding systems produce descriptions of (a) the image and (b) the world scene that the image represent. [1]

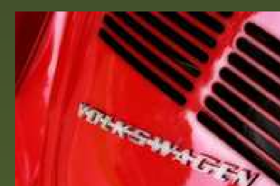




Example of Image Understanding

Understanding usually attempts to mimic the human visual system in extracting meaning from an image Given a reason for looking at a particular scene, understanding systems produce descriptions of (a) the image and (b) the world scene that the image represent. 1





System should be able to describe all these images as say **red car**. 1

Image Definitions

- An image in the *real world* can be considered to be a function of two real variables $A(s)$ where A is the amplitude (intensity) of the image and $s = (x, y)$ is the real coordinate position

Image Definitions

- An image in the *real world* can be considered to be a function of two real variables $A(s)$ where A is the amplitude (intensity) of the image and $s = (x, y)$ is the real coordinate position
- An image may be considered to contain sub-images sometimes referred to as regions-of-interest.

Image Definitions

- An image in the *real world* can be considered to be a function of two real variables $A(s)$ where A is the amplitude (intensity) of the image and $s = (x, y)$ is the real coordinate position
- An image may be considered to contain sub-images sometimes referred to as regions-of-interest.
- In a practical image processing system it should be possible to apply specific image processing operations to selected regions.

Image Definitions

- An image in the *real world* can be considered to be a function of two real variables $A(s)$ where A is the amplitude (intensity) of the image and $s = (x, y)$ is the real coordinate position
- An image may be considered to contain sub-images sometimes referred to as regions-of-interest.
- In a practical image processing system it should be possible to apply specific image processing operations to selected regions.
  Example, one part of an image might be processed to suppress motion blur while another part might be processed to improve color rendition.

Image Definitions (2)

The amplitudes (A) of a given image will almost always be either real numbers or integer numbers due to quantization.

- In certain image-forming processes, however, the signal may involve photon counting which implies that the amplitude would be inherently quantized.
- In other image forming procedures, such as magnetic resonance imaging, the direct physical measurement yields a complex number in the form of a real magnitude and a real phase.

Digital Image Definitions

- a digital image $a[m, n]$ is described in a 2D discrete space
- it is derived from an analog image $A(x, y)$ in a 2D continuous space
- through a sampling process (called digitization).

Digital Image Definitions

- a digital image $a[m,n]$ is described in a 2D discrete space
- it is derived from an analog image $A(x,y)$ in a 2D continuous space
- through a sampling process (called digitization).

Digital Image Definitions

- a digital image $a[m, n]$ is described in a 2D discrete space
- it is derived from an analog image $A(x, y)$ in a 2D continuous space
- through a sampling process (called digitization).

The continuous image $A(x, y)$ is divided into N rows and M columns.

Digital Image Definitions

- a digital image $a[m, n]$ is described in a 2D discrete space
- it is derived from an analog image $A(x, y)$ in a 2D continuous space
- through a sampling process (called digitization).

The continuous image $A(x, y)$ is divided into N rows and M columns.
The value assigned to the integer coordinates $[m, n]$ with $\{m = 0, 1, ..., M - 1\}$ and $\{n = 0, 1, ..., N - 1\}$ is $a[m, n]$.

Digital Image Definitions

- a digital image $a[m, n]$ is described in a 2D discrete space
- it is derived from an analog image $A(x, y)$ in a 2D continuous space
- through a sampling process (called digitization).

$M = 8; N = 7$

The continuous image $A(x, y)$ is divided into N rows and M columns.
The value assigned to the integer coordinates $[m, n]$ with $\{m = 0, 1, ..., M - 1\}$ and $\{n = 0, 1, ..., N - 1\}$ is $a[m, n]$.

Image Creation

The creation of images involves two main tasks

- spatial sampling, (determines resolution of image)
- quantization, (determines allowed intensity levels)

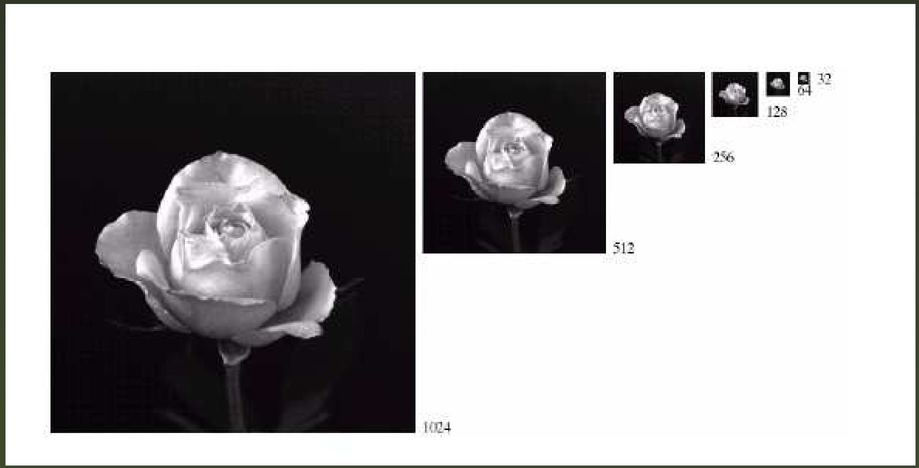
Spatial sampling – level of detail that is seen

- finer sampling allows for smaller detail ; more pixels - larger image size

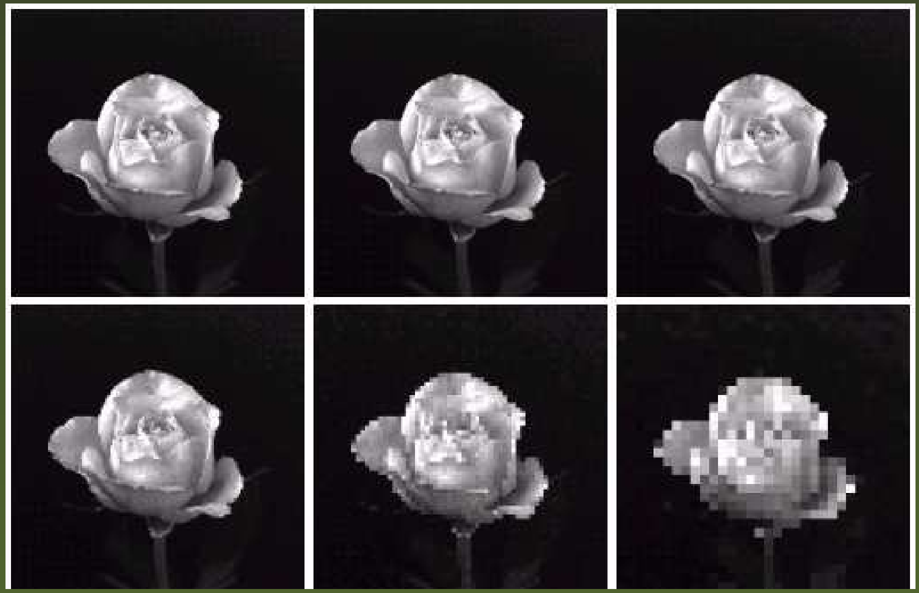
Quantization – how "smooth" the contrast changes are

- finer quantization will prevent "false contouring" (artificial edges)
- coarser quantization allows for compressing

Sampling Effect



Sampling Effect



Quantization Effect

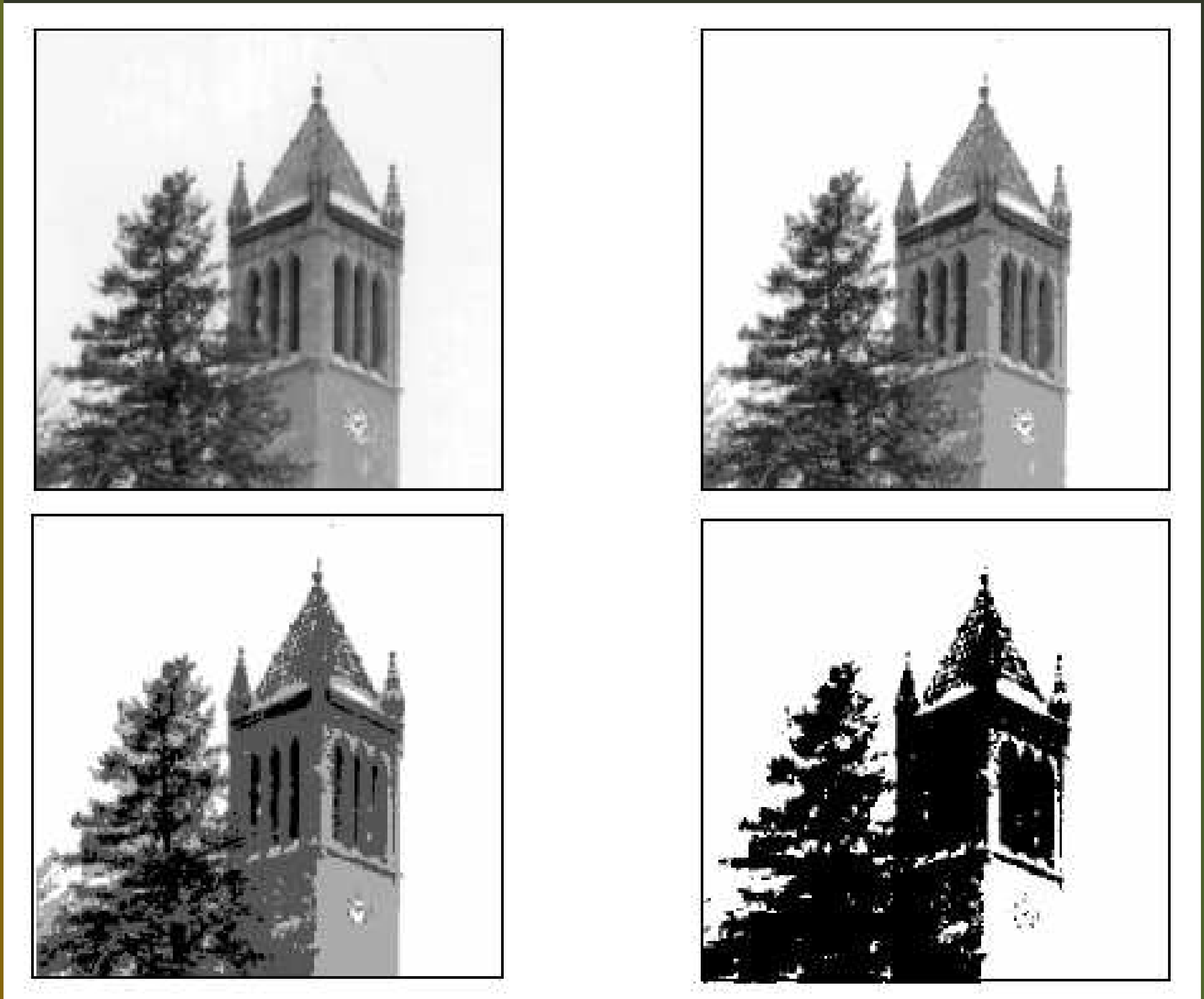


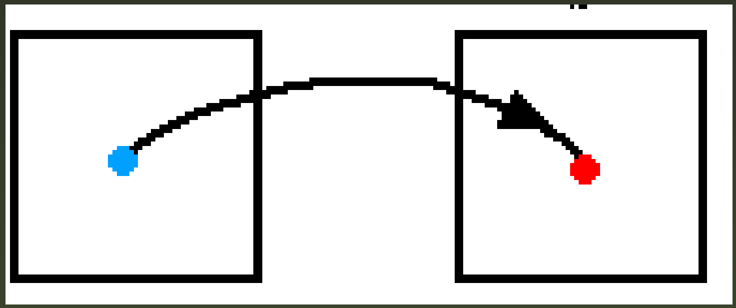
Image Processing Operations

The types of operations that can be applied to images to transform an input image $a[m,n]$ into an output image $b[m,n]$ can be classified into three categories

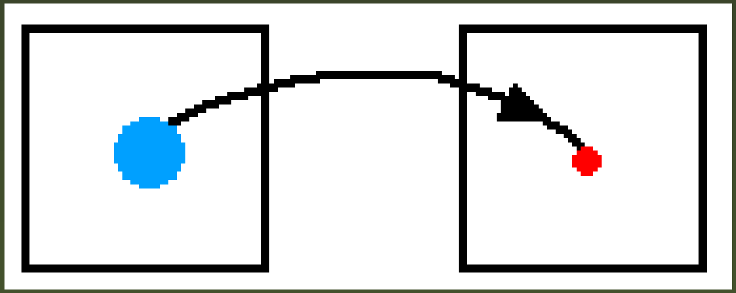
- *point* - the output value at a specific coordinate is dependent only on the input value at that same coordinate.
- *local* - the output value at a specific coordinate is dependent on the input values in the neighborhood of that same coordinate.
- *global* - the output value at a specific coordinate is dependent on all the values in the input image.

Image Processing Operations (2)

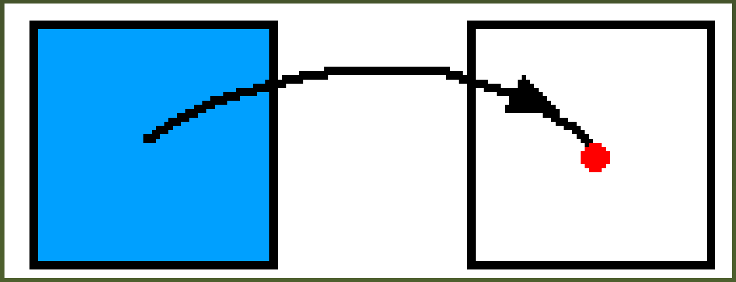
- point operation



- local operation



- global operation



Neighborhood of a pixel need to be defined.

Pixel Neighborhood

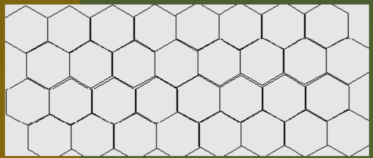
Neighborhoods that can be used to process an image

- *rectangular sampling* - image is sampled by laying a rectangular grid over an image

Pixel Neighborhood

Neighborhoods that can be used to process an image

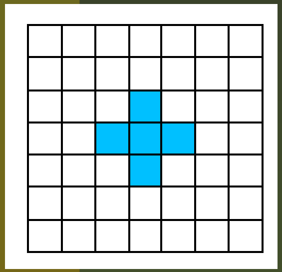
- *rectangular sampling* - image is sampled by laying a rectangular grid over an image
- *hexagonal sampling* - image is sampled by laying a hexagonal grid over an image

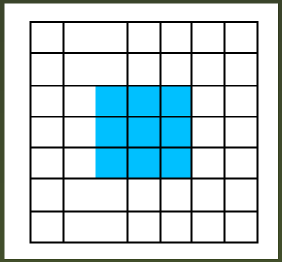


Pixel Neighborhood

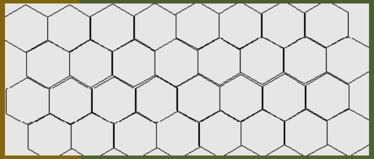
Neighborhoods that can be used to process an image

- *rectangular sampling* - image is sampled by laying a rectangular grid over an image


4 ngbd


8 ngbd

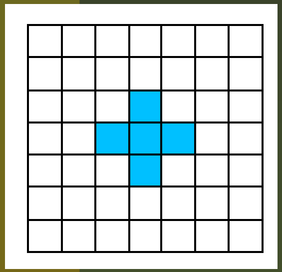
- *hexagonal sampling* - image is sampled by laying a hexagonal grid over an image



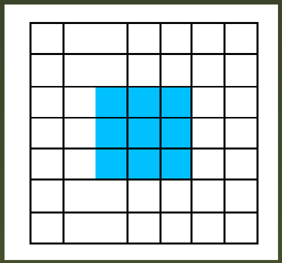
Pixel Neighborhood

Neighborhoods that can be used to process an image

- *rectangular sampling* - image is sampled by laying a rectangular grid over an image

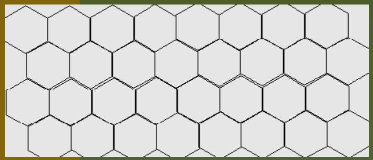


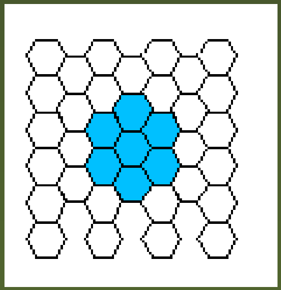
4 ngbd



8 ngbd

- *hexagonal sampling* - image is sampled by laying a hexagonal grid over an image





6 ngbd

Image Processing Tools

(Manipulation) tools are central to the processing of digital images. These include

- mathematical tools
  - convolution,
  - Fourier analysis
- statistical descriptions, and manipulative tools
  - chain codes
  - run codes

Convolution

$c = a \times b$

- 2D continuous domain

$$c(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(\alpha, \beta) b(\alpha - x, \beta - x) d\alpha d\beta$$

- 2D discrete domain

$$c[m, n] = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} a[i, j] b[m - i, n - j]$$

Convolution Properties

a, b, c, d are all 2D images

- Commutative

$$a \times b = b \times a$$

- Associative

$$(a \times b) \times c = a \times (b \times c) = a \times b \times c$$

- Distributive

$$a \times (b + d) = (a \times b) + (a \times d)$$

Fourier Transform

Fourier transform produces a representation of a (2D) signal as a weighted sum of sines and cosines

- 2D discrete domain $c[m,n] = \mathcal{F}(a[i,j])$

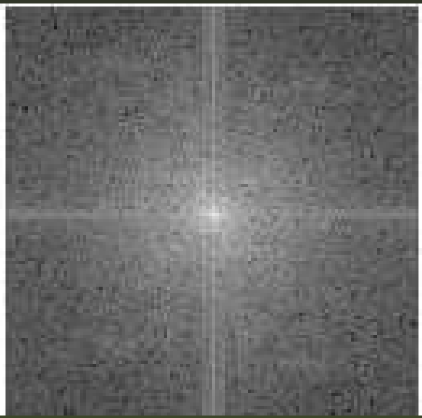
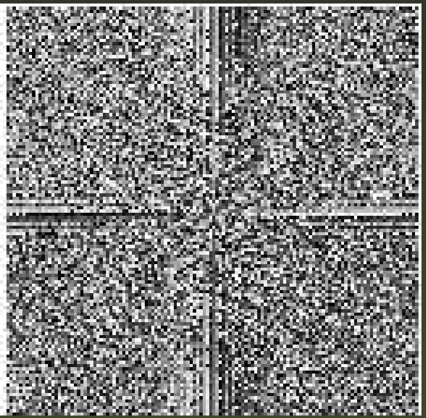
$$c[m,n] = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} a[i,j] \exp^{-(mi+nj)}$$

The Fourier transform is unique and invertible operation

$$a[i,j] = \mathcal{F}^{-1}(\mathcal{F}(a[i,j]))$$

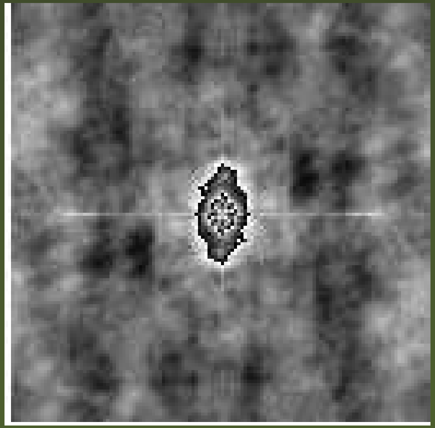
Notice that the Fourier transform is a complex function.

Fourier Transform (2)

Original magnitude phase

Reconstruction




Magnitude only; Phase only

Magnitude and phase required for reconstruction

Algorithms

Algorithms that are used in image processing can be divided into four categories
These could be implemented as a point, local or global operation)

- histogram based
- mathematics based
- convolution based
- morphology-based

What is a Histogram?

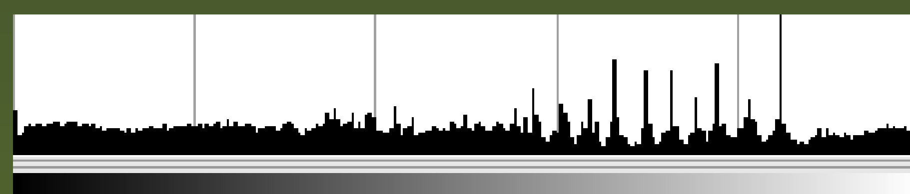
Histogram

x-axis gray level values 0 - 255 and y-axis is the number of pixels.

Equalization: Histogram based

Manipulation of the histogram to achieve a distinct goal.
Note that this is a point operation
Histogram equalization attempts to change the histogram of an image through the use of a function $b = f(a)$ into a histogram that is constant for all brightness values. Producing a brightness distribution where all values are equally probable. Unfortunately, for an arbitrary image, one can only approximate this result.





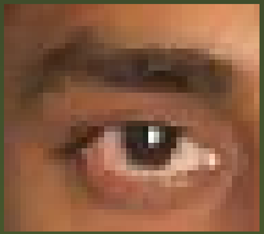
Other types of processing

- Image morphing
- Colouring (BW movies to colour)
- Wavelet processing (multiresolution)
- stereo processing (image)
- motion estimation
- camera calibration (importance)
- KL Transform
- Edge/Smooth filtering

Example: Face Feature identification

Identify the facial features, namely eyes, nose, lips, ears.

- Identify eyes
  - Photographs taken with a flash (why?)
  - Eyes are distinct by the presence of a white spot inside the pupil (this aspect can be exploited).



- Assume that the location of the eyes are know *a priori*
- Using Anthropometric data (statistical) to locate other features

Eye Detection (1)

Eye Detection (2)

Locate possible regions

Other Facial Features

- Hair
  Use colour; shape; anthropometry information
- Skin
  Use colour and anthropometry information
- Lips
  Use colour ; shape anthropometry information
- Ears
  Shape and anthropometry

Example: Hair Detection

Example: Lips Detection

Example: Skin Detection

Example: All Features

Processing in OCR Systems

- *Scanning and Pre-processing of Printed data* - Noise removal Binarization Skew detection and correction
- *Document layout analysis* - Text and non-text region (image/graphics) separation Single / multiple column document
- *Document segmentation* - Line, word and character segmentation
- *Character Recognizer* - Feature extraction Classification
- *Language dependent post processing* - Spell checker Letter n-gram, word n-gram models

Processing in OCR Systems

- Binarization schemes
  - Histogram based
  - Clustering based
  - Entropy based
  - Model based
- Skew detection
  - Projection profile based techniques
  - Entropy based techniques
  - Hough transform
  - Principal component analysis (PCA)
  - Radon Transform

Character Recognition

- K-NN classifier
  Back propagation
- Decision trees
  Bayesian Networks
- Probabilistic models
  Hidden Markov models

Speech Processing

Speech Signal Processing?

Speech signal processing refers to the acquisition, manipulation, storage, transfer and output of human utterances by a computer.

- **Speech recognition** focuses on capturing the human voice as a digital sound wave and converting it into a computer-readable format (speech to text conversion).
- **Speaker verification** focuses on verifying the identity of the speaker.
- **Speech synthesis** or **Text to Speech** is the reverse process of speech recognition. A TTS system converts normal language text into speech.

Speech Recognition: Overview

- **Input** Speech, 16kHz, 8 bit
- **Output**

1. a phoneme string — sil h au m a ch m ae k s i m a m a m A u n T sil k ae n ai w i D r ao th r U E T I e m sil
2. find word boundaries using dictionary — $\boxed{\text{hau}}$ $\boxed{\text{mach}}$ $\boxed{\text{maeksimam}}$ $\boxed{\text{amAunT}}$ $\boxed{\text{kaen}}$ $\boxed{\text{ai}}$ $\boxed{\text{wiDrao}}$ $\boxed{\text{thrUE}}$ $\boxed{\text{TIem}}$
3. converting the phoneme strings into text

How much maximum amount can I withdraw through ATM

Speaker Verification: Overview

- Is the process of verifying the claimed identity of a registered speaker using his voice characteristics.
- The speaker needs to enroll before using the system.
- During enrollment, the speaker speaks a given set of utterances, using which the systems builds statistical models representing the speaker's voice.
- A user claims he is X. Speaks a pass-phrase. The system gives a binary output YES (accept claimed identity) | NO.

Need for threshold to be able to say Yes or No.

Types of Speaker Verification

1. **Fixed Phrase** – pre-determined phrase used for verification
2. **Fixed Vocabulary** – verification more flexible and practical; training and testing materials for a speaker are generated based on words of a fixed vocabulary
3. **Flexible Vocabulary** – a general set of sub-word phone models is created during speaker model training
4. **Text-Independent** – user is not constrained to say fixed or prompted phrases

Clearly, both complexity and security increases as we go from fixed phrase to text-independent.

Speech Synthesis (Text to Speech)

Speech Synthesis is the art of making a machine speak as well as an average literate human is capable of.

- **Input** – $\boxed{\text{hau}}$ $\boxed{\text{mach}}$ $\boxed{\text{maeksimam}}$ $\boxed{\text{amAunT}}$ $\boxed{\text{kaen}}$ $\boxed{\text{ai}}$ $\boxed{\text{wiDrao}}$ $\boxed{\text{thrUE}}$ $\boxed{\text{TIem}}$
- **Ideal Output** – Speech

The objective of speech synthesis is *deemed* complete when a human can not distinguish between a human spoken and a machine spoken speech.

Types of Speech Synthesis

- **Formant synthesis:** Formant synthesizers use a simple model of speech production and a set of rules to generate speech.
  The quality of formant synthesizers is robotic because it is difficult to reduce the speech acoustic context and quality to a simple set of rules.
- **Concatenative synthesis:** Concatenative synthesizers use speech segment units and achieve higher quality than formant synthesizers.
  This is labour intensive and does not lend easily itself to adaptation to the speaker characteristics, as there are thousands of different speech units, in different context.

Processing in Speech Recognition

Hidden Markov models (HMMs) are best suited for modeling speech

1. Statistical models (able to capture large variations which are possible in speech)
2. Able to preserve temporal information (important in speech)
3. Have been in use for several decades (with no visible replacements spare Artificial Neural Networks)
4. Their use has been successfully demonstrated (time and again)

Speech Pre-processing (1)

- Speech is non-stationary.
  Meaning, the statistics of the speech signal change with time.
- To develop a statistical model for speech we need to consider smaller *portions* of speech.
  Typically 10-20 ms of speech called speech frames where the signal can be considered to be stationary (a key assumption in all current speech recognition systems).

Speech Pre-processing (2)

- Dividing the speech signal into frames,
- Removing non-speech signal,
- Pre-emphasizing the signal to spectrally flatten the signal to make it less susceptible to finite precision effects in signal processing
  To offset 3 dB per octave fall due to the effect of radiation from the lips
- Tapering (Windowing) the frames (Hamming window)
  To minimize signal discontinuities at the beginning and end of the frame.

Processing in Speech

- Energy based methods
  speech - non-speech; vowel consonant separation
- Zero crossing based methods
  pitch detection, compression
- Frequency (Spectral) domain methods
  Fourier transform, Wavelet processing for compression and feature extraction
- Model or Learning
  HMM, Neural Networks, Dynamic time warping
- Unconventional Methods
  Error correcting codes for number recognition, speech spectrogram processing

Spoken Number Recog Accuracy

- Aim
  To increase the recognition accuracy of a connected digit recognizer without increasing the digit recognizing accuracy *per se*
- Basic idea
  Increase the number of digits in a number and use these appended digits to increase the overall accuracy of recognizing the number, as is done in the error correcting code literature

Problem Formulation

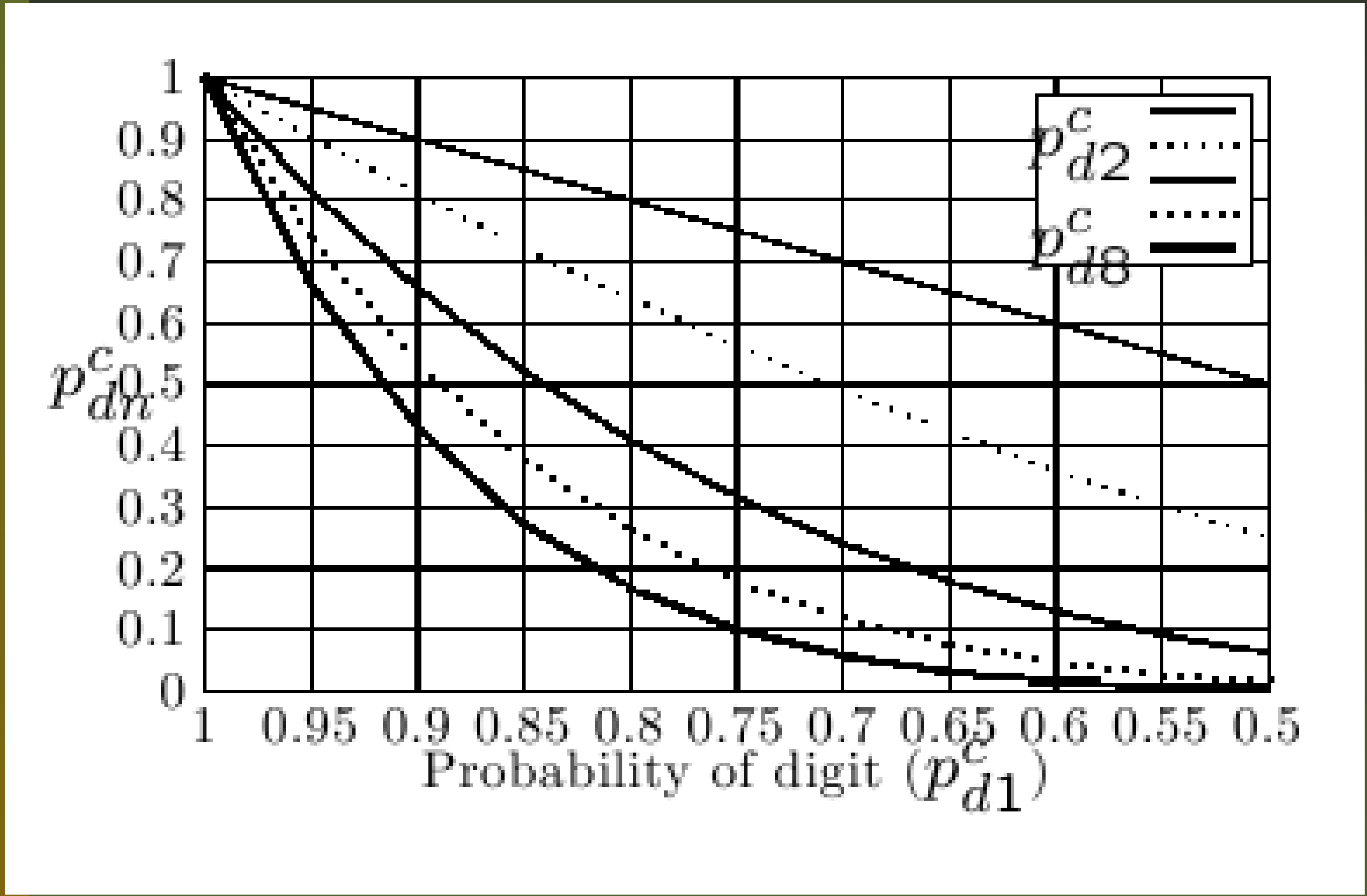
If p^c_{d1} is the probability of correct recognition of a single digit ($d1$), then the probability of recognizing a n digit connected number (dn) is

$$p^c_{dn} = (p^c_{d1})^n$$

Clearly, $p^c_{dn} < (p^c_{d1})^n$ for $n > 1$, $0 < p^c_{d1} < 1$.
Note: A 0.95 single digit accuracy results in a 6 digit number recognition accuracy of only 0.735 ($(0.95)^6$)!

Can we increase p^c_{dn} without increasing p^c_{d1}?

Number vs digit recog accuracy



How? — Append extra digits

The accuracy of recognition of a α digit number (probability of correct recognition of all the α digits) is

$$p^c_{d\alpha} = (p^c_{d1})^\alpha$$

Suppose we append $\beta > 0$ extra digits to the α-digit number; then the accuracy of recognition (probability of correctly recognizing all the $\gamma = (\alpha + \beta)$ digits) of the number is

$$p^c_{d\gamma} = (p^c_{d1})^\gamma$$

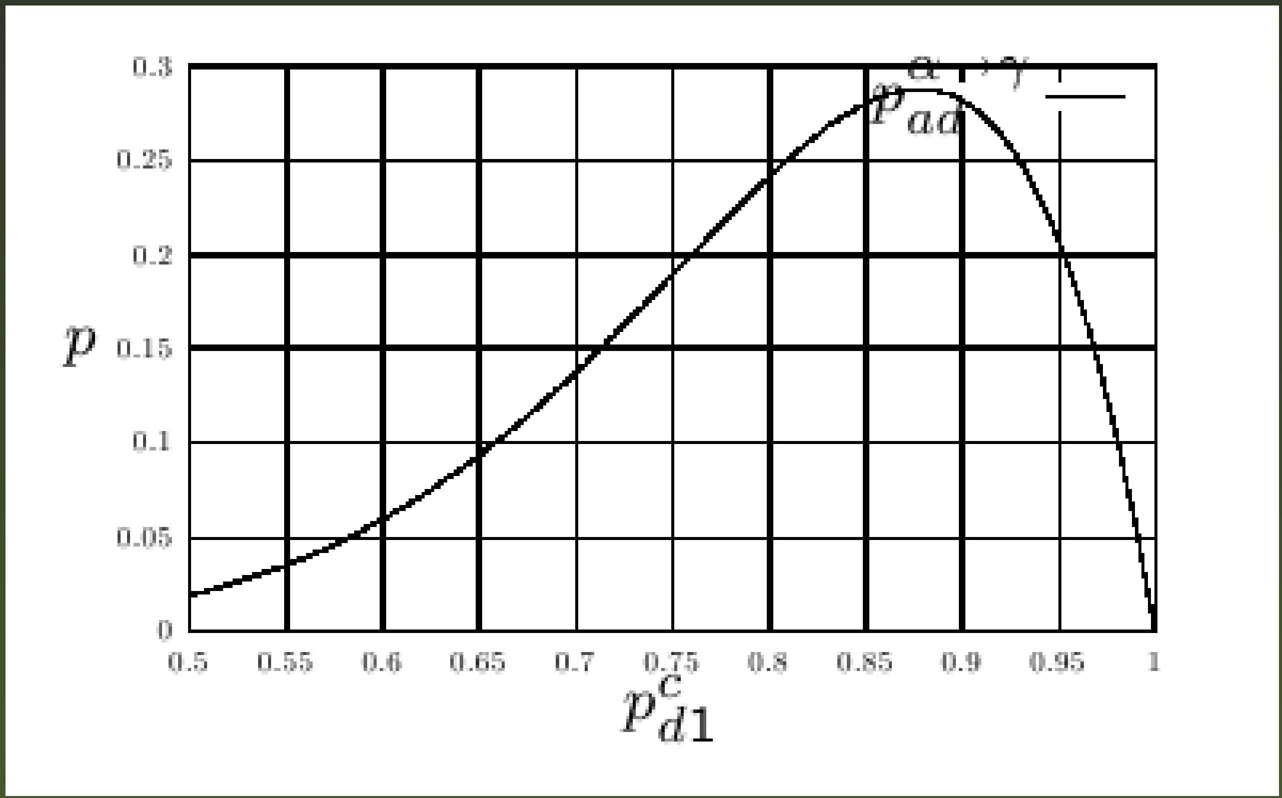
Clearly, adding extra digits reduces the accuracy of number recognition. Note that $p^c_{d\gamma} < p^c_{d\alpha}$ for $\gamma > \alpha$.

Append extra digits useful?

If the γ-digit number is constructed such that it is possible to identify and correct the k digits in error in the γ-digit number, then, the advantage of appending the extra β digits to the α digit number is

$$p_{ad}^{\alpha \rightarrow \gamma} = \left\{ p_{d\gamma}^{c} + \sum_{j=1}^{k} {}^{\gamma}C_{1} (p_{d1}^{c})^{\gamma - j} (1 - p_{d1}^{c})^{j} \right\} - p_{d\alpha}^{c}$$

Adv. single error digit correction



$$p_{ad}^{\alpha\rightarrow\gamma} = \left\{p^c_{d\gamma} + {}^{\gamma}C_1 (p^c_{d1})^{\gamma-1}(1-p^c_{d1})^1\right\} - p^c_{d\alpha}$$

Material Used from all sources

- http://facweb.cs.depaul.edu/research/vc/VC_Workshop/ presentations/pdf/ Jacob_tutorial1.pdf
- Image Processing – http://www.ph.tn.tudelft.nl/Courses/FIP/
- Wikipedia – http://en.wikipedia.org

Thank You

SunilKumar.Kopparapu@TCS.Com
TCS Innovation Labs - Mumbai
Advanced Technology Application Group
Yantra Park, Thane (West), Maharastra 400 601
http://www.tcs.com