INFS4205/7205 Advanced Techniques for High Dimensional Data

An Introduction to Multimedia Databases

Semester 1, 2021

University of Queensland

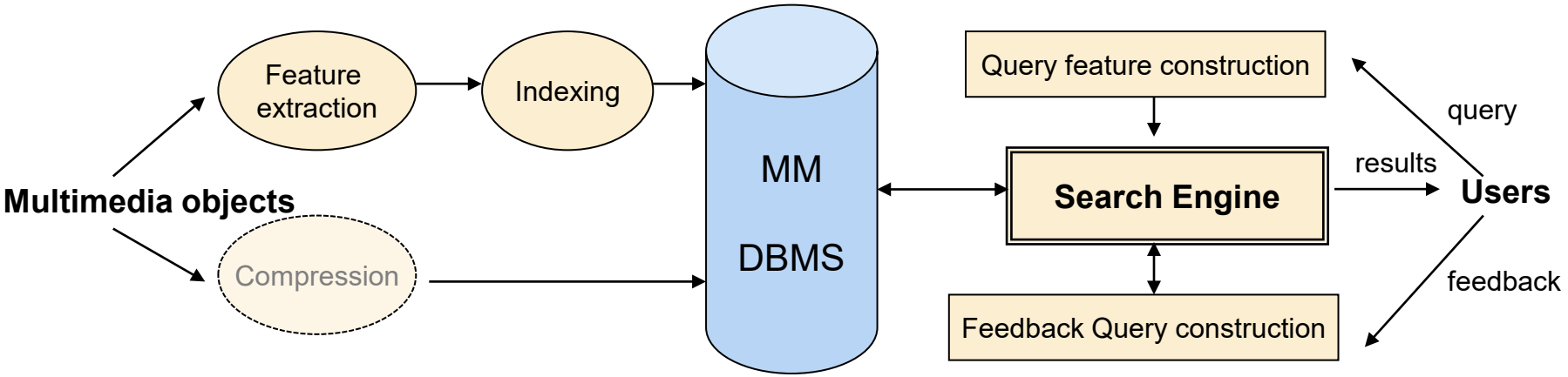
Advanced Techniques for High Dimensional Data

- Course Introduction
- Introduction to Spatial Databases
- Spatial Data Organization
- Spatial Query Processing
- Managing Spatiotemporal Data
- Managing High-dimensional Data
- Introduction to Multimedia Database
- Route Planning in Road Network
- When AI Meets High-Dimensional Data
- Trends and Course Review

Learning Objectives

- To understand what is Multimedia Databases
- To learn abstraction (feature representations) of MM
- To search in MM database (content-based retrieval)

A Generic Architecture of MMDBMS

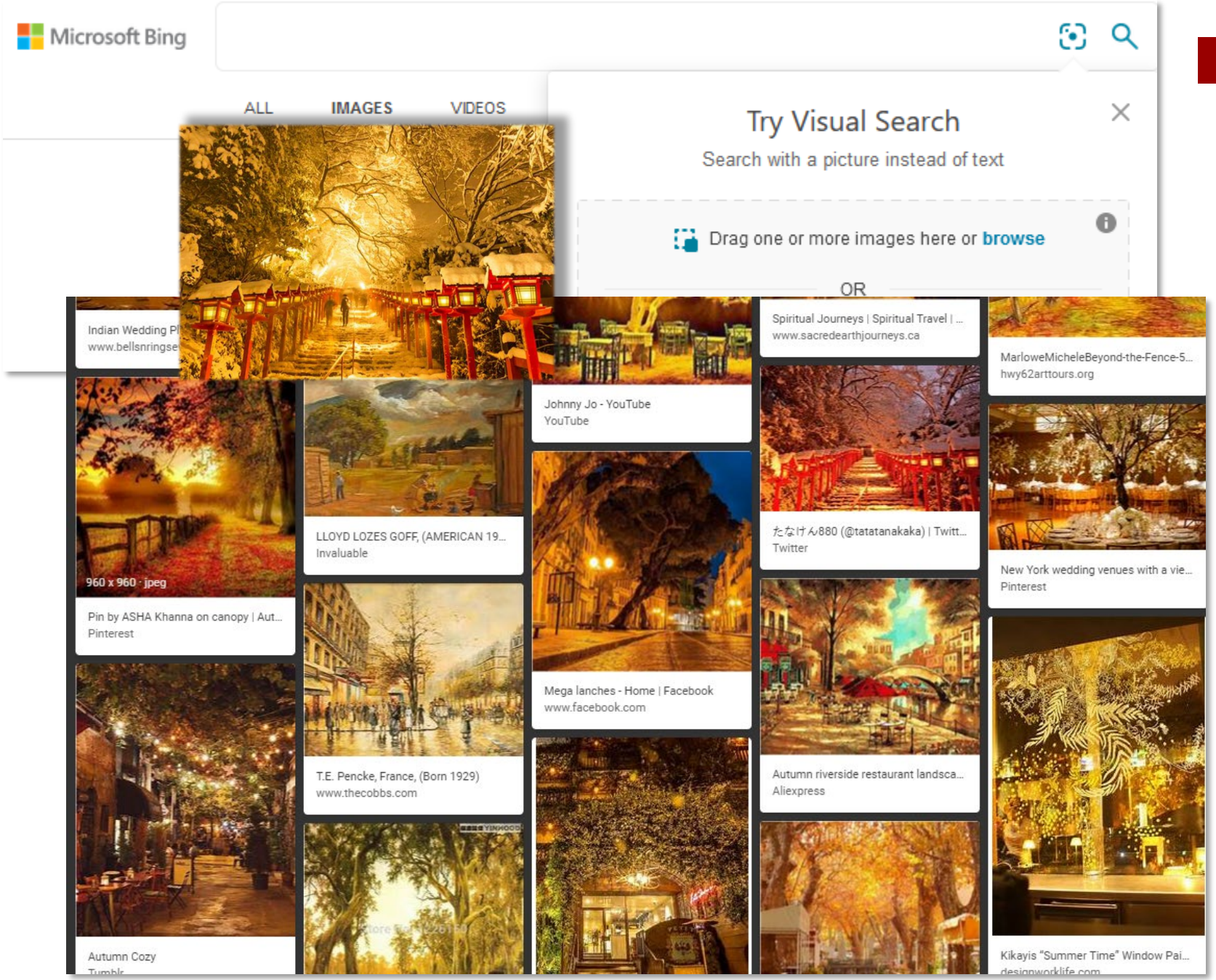


+ Multimedia data

- Multimedia data enables information to be represented through text, audio, graphics, image, animation, and video…
- Types of media
  - Text and documents
    - Include: words, sentences, paragraphs, …, structure, link, etc. (e.g., xml/html tags, hyperlink)
    - Now also include (micro)blogs, SMS
  - Image and graphics
    - Digital equivalents of drawings, paintings, photographs (including medical imaging), or prints
  - Audio and video
    - Considers temporal characteristics
    - Speech
      - Audio in natural language
      - Speech-to-text and text-to-speech tools

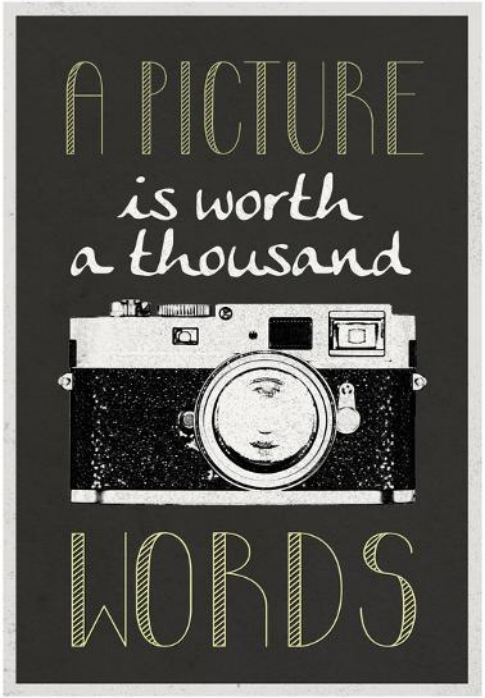
+ Data Characteristics

- Volume, velocity and variety
  - An image feature is typically a high-dimensional (tens or hundreds or more) feature vector
  - A short video may contain tens or hundreds of frames/images
- Similarity-based search
  - Unlike traditional “exact search” in relational database, users usually ask for similar objects based on their contents
  - A multimedia object may contain multiple features, and similarity can be subjective and content-dependent


Microsoft Bing
ALL
IMAGES
VIDEOS
Try Visual Search
Search with a picture instead of text
Drag one or more images here or browse
OR
Indian Wedding Pl
www.bellsnringse
Spiritual Journeys | Spiritual Travel | ...
www.sacredearthjourneys.ca
MarloweMicheleBeyond-the-Fence-5...
hwy62arttours.org
Johnny Jo - YouTube
YouTube
960 x 960 · jpeg
Pin by ASHA Khanna on canopy | Aut...
Pinterest
LLOYD LOZES GOFF, (AMERICAN 19...
Invaluable
たなけん880 (@tatatanakaka) | Twitt...
Twitter
New York wedding venues with a vie...
Pinterest
Mega lanches - Home | Facebook
www.facebook.com
T.E. Pencke, France, (Born 1929)
www.thecobbs.com
Autumn riverside restaurant landsca...
Aliexpress
Autumn Cozy
Tumblr
Kikayis "Summer Time" Window Pai...

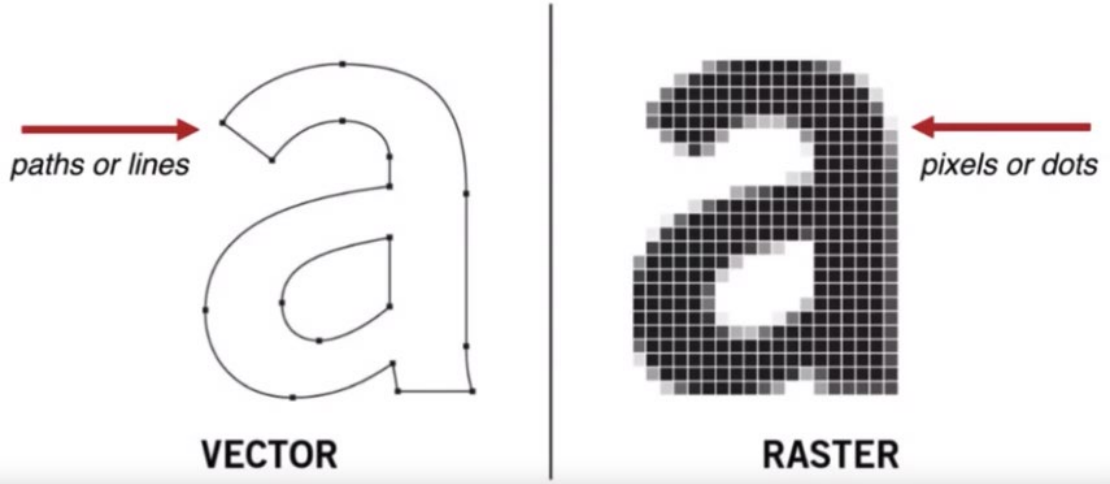
+ Motivation for MM Databases

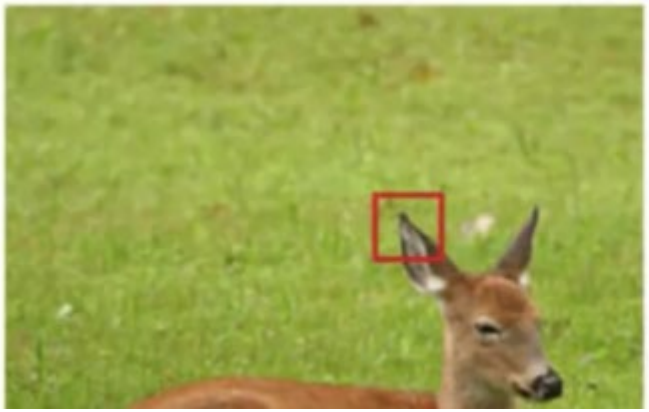
- Multimedia is a much more powerful communication tool than traditional text in our daily life
  - Image showcase, graphic design, TV commercial, short video, speech, movie, mobile phone multimedia message, etc
- We need to organize, manage and search these new multimedia data
  - RDBMS are no longer suitable for complex multimedia data
  - Need for robust systems which can manage and search multimedia data in a reliable and efficient way



+ Image data

- Vector image vs. Raster image



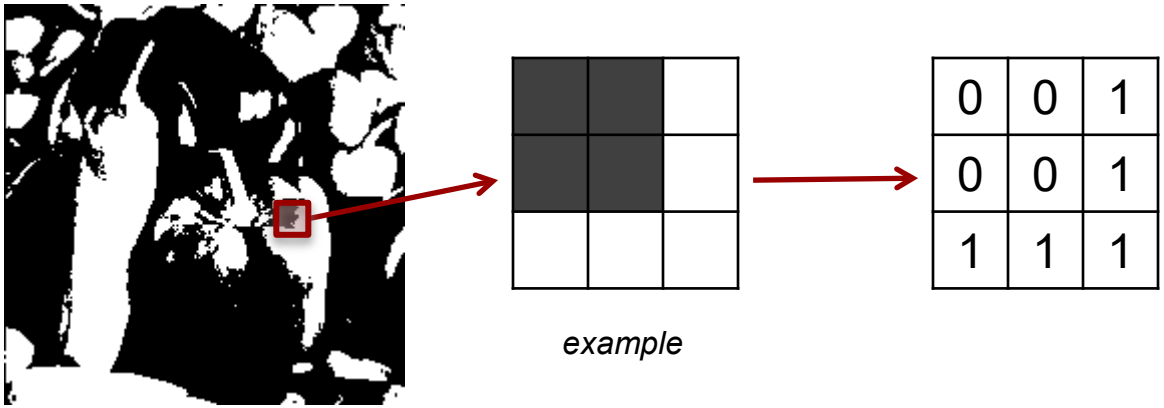


+ Raster Image Characterises

- Resolution
  - The number of pixels in a given physical distance
  - DPI– dots per inch
- Dimension
  - Size in pixels, 512x512, 1024x768, 1560x1024…
- Bit depth
  - How many bits are used to represent one pixel
- Colour mode
  - RGB, CMYK…

Monochrome Image

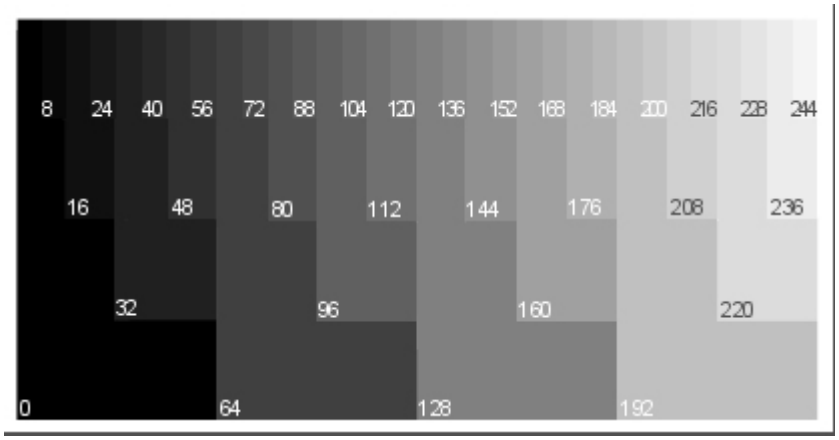
- Each pixel is stored as a single bit (0 or 1)
- 0: Black 1:White



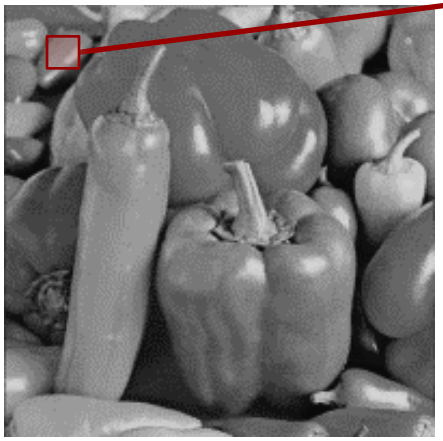
- A 640 × 480 monochrome image requires 37.5 KB of storage
  - Size : 640 × 480 = 307,200 (pixels)
  - Each pixel: 1 bit
  - 307,200 bits = 38,387 bytes = 37.5 Kb
    (1Kb = 1024 bytes, 1 byte = 8 bits)

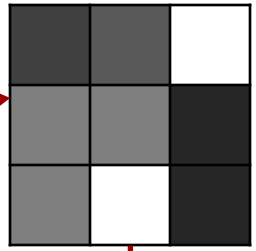
+ Gray-level Image

- Each pixel is usually stored as a byte = 8 bits (encode integer value between 0 to 255)
  - 0: Black 255: White
- A 640 × 480 monochrome image requires 300 KB (=37.5*8) of storage



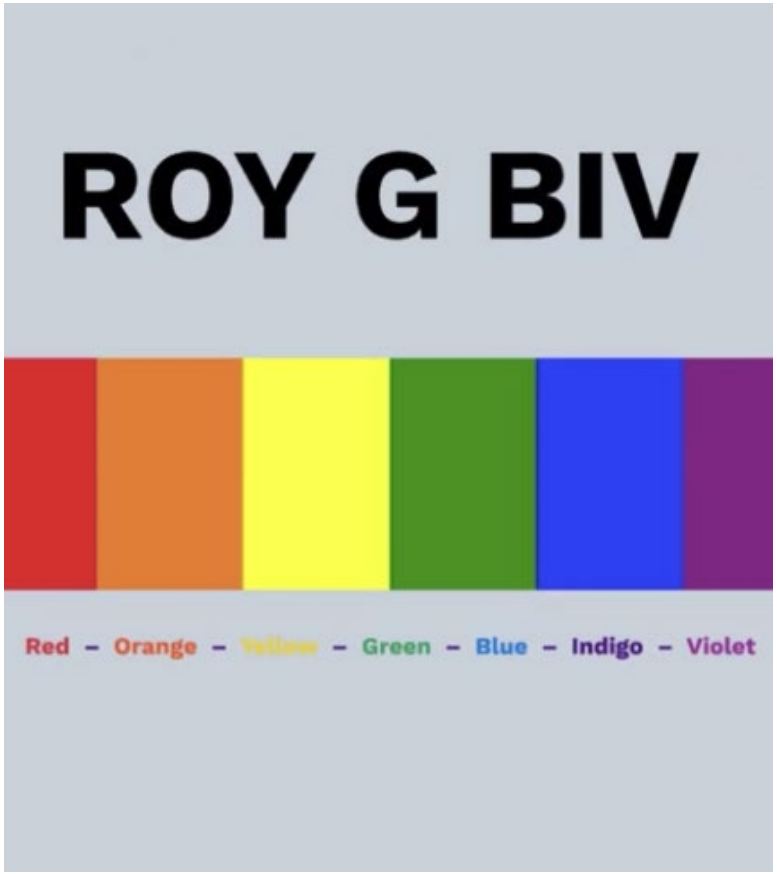
8-bit Gray scale





30	48	255
104	104	0
104	255	0

Dispersion of Light



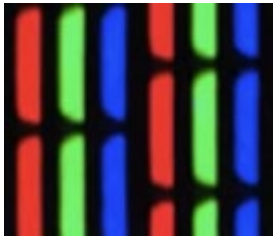
Isaac Newton, 1666

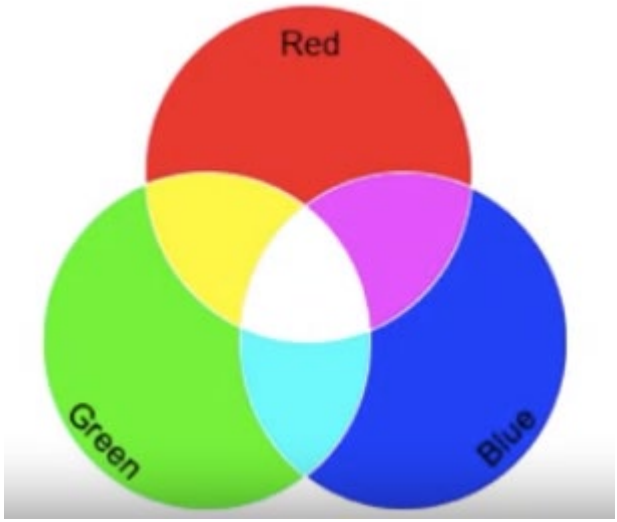
+ Colour models

- Colours can be represented as tuples of numbers, typically three or four values
- Several major models:
  - RGB : <**r**ed, **g**reen, and **b**lue> – monitor, scanner …
  - HSV**:** <**h**ue, **s**aturation, **v**alue> – artists
  - CMYK: <**c**yan, **m**agenta, **y**ellow and black> – printing inks
  - Other popular ones such as CIE (Commission Internationale de L'Eclairage) and YUV (luma Y and two chrominance UV)

RGB Model

- Three primaries: **R**ed, **G**reen, and **B**lue
- An additive colour model
  - Two colours add to each other
  - Mainly used for TV and monitor





- A colour image is usually an array of (*R,G,B*) integer triplets
  - White (*255,255,255*)
  - Black(*0,0,0*)
  - Yellow (*255,255,0*).
- HEX colour model
  - White #FFFFFF
  - Black #000000

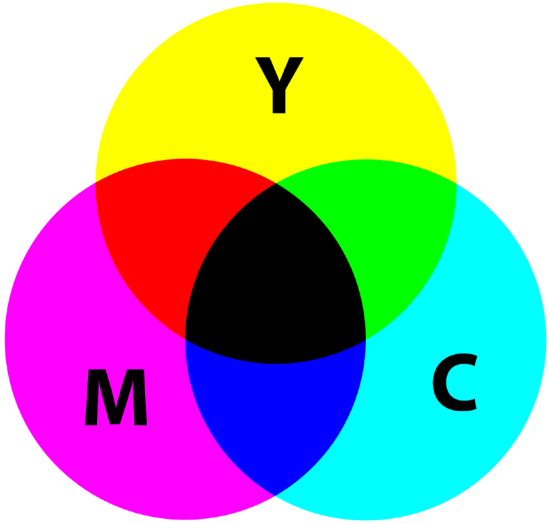
	A	B	C	D	E
1	R	G	B	HEX	Colour
2	255	0	0	FF0000	
3	255	128	0	FF8000	
4	191	255	0	BFFF00	
5	0	255	11	00FF0B	
6	0	255	191	00FFBF	
7	0	64	255	0040FF	
8	128	0	255	8000FF	

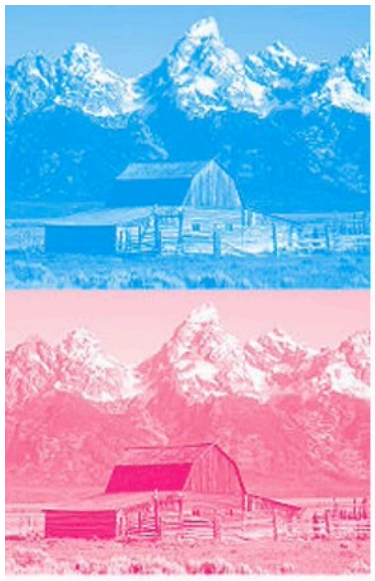
+ HSV Model

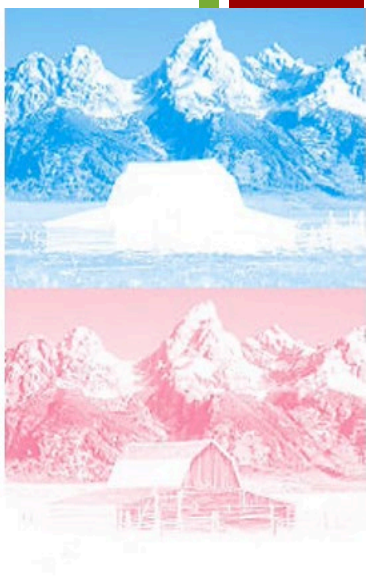
- Hue
  - Colours we know from the basic colour wheel (rainbow)
  - Usually represented in the range of 0 to 360 degrees
- Saturation
  - The brilliance/vividness/dullness of a colour
    - A pure colour has a saturation 100%
    - A white colour has a saturation 0%
- Value
  - Darkness of the colour
    - Lightest: 0%
    - Darkest: 100%

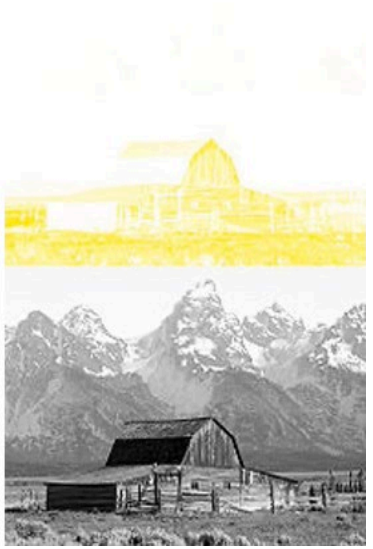
+ CMYK Model

- Cyan, Magenta, Yellow, Black
  - Pigment theory, Ink/Printing
  - Subtractive colour
    - How white light is absorbed and reflected through surfaces









+ Many Image Formats

Raster Image

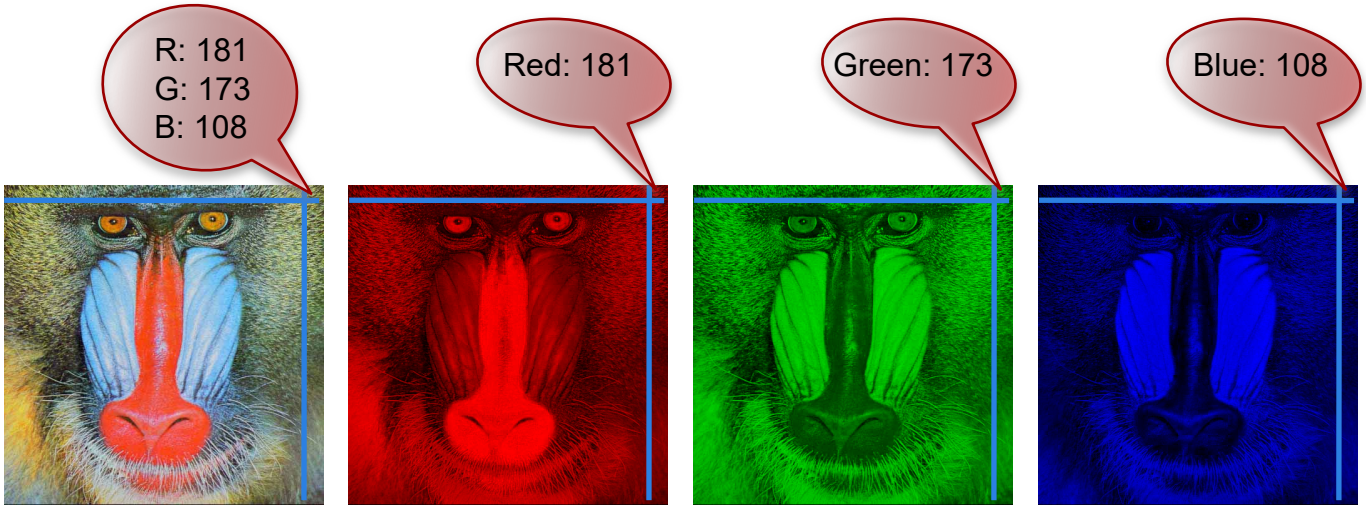
- BMP (Bitmap) – limited colour range
- GIF (Graphics Interchange Format) – 8-bit colours, transparent background, buttons…
- TIFF (Tagged Image File Format) – various colour models, printing, do not allow transparent background
- JPEG (Joint Photographic Experts Group) – various colour models, commonly for photographs, do not allow transparent background
- PNG (Portable Network Graphics) – networks, different degree of transparency

Vector Image

- EPS (Encapsulated PostScript) – mainly used for logos, transparent background

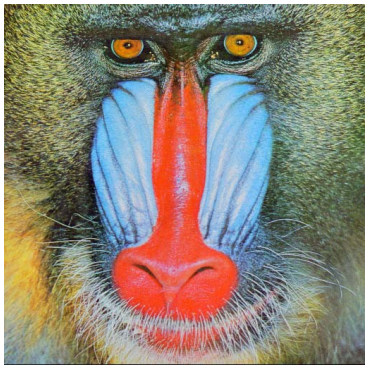
+ JPEG

- Each pixel is represented by 3 bytes
  - 24-bits, supporting 256 × 256 × 256 possible combined colors (or 16,777,216 different colors)
- A 640 × 480 24-bit colour image requires 921.6 KB of storage (without compression)

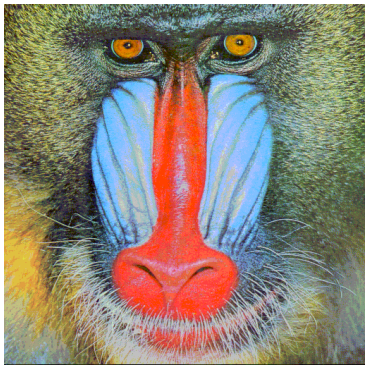


+ GIF

- 8-bit Colour Image (GIF)
  - One byte for each pixel (R: 3-bit; G: 3-bit; B: 2-bit)
  - Support 256 (i.e., 8 × 8 × 4) possible colors, acceptable colour quality
  - A 640 × 480 8-bit colour image requires 300 KB of storage (the same as 8-bit gray scale)



24-bit colour image



8-bit colour image

PBS - The Evolution of 8-Bit Art: https://www.pbs.org/video/off-book-evolution-8-bit-art/ Beginning with early Atari and Nintendo video games, the 8-bit aesthetic has been a part of our culture for over 30 years. As it moved through the generations, 8-bit earned its independence from its video game roots. No longer just nostalgia art, contemporary 8-bit artists and chiptunes musicians have elevated the form to new levels of creativity and cultural reflection.

Media Data Abstraction

- Abstraction: low-level representation of multimedia data
  - Also known as **features**
  - An abstraction captures the content of original media (but smaller in size)
- Indexing are usually built on top of the abstractions (also called features), instead of original data

Image Abstraction

- Colour feature
  - Analyzing the colour value of each pixel and capturing colour distribution
- Texture feature
  - Identifying the feature of a pattern (spatial arrangement of the gray levels of pixels)
- Shape feature
  - An object's shape/location
  - E.g. how many triangles, how many rectangles…
- There are many other types of features…

Feature Space for Images

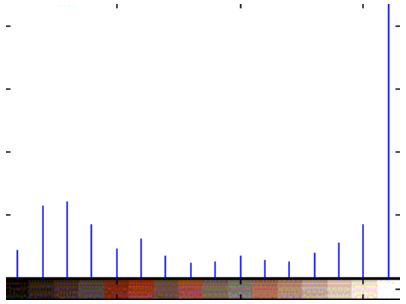
- A general definition
  - A feature is an attribute derived from transforming the original visual object by using an image analysis algorithm
  - It characterizes a specific property of interest of an image, such as its colour, texture, or shape…
  - An image is represented by a set of values called a feature vector
  - A feature space is useful from the retrieval point of view, as one can locate the images that are represented by a point in the feature space

+ Colour Feature

- Colour Histogram
  - Distribution of colours
  - Colour histogram describes
    - The colours and their pixels percentages in an image
      - All colors are quantized into colour bins
      - Pixel number for each colour bin is counted and normalized

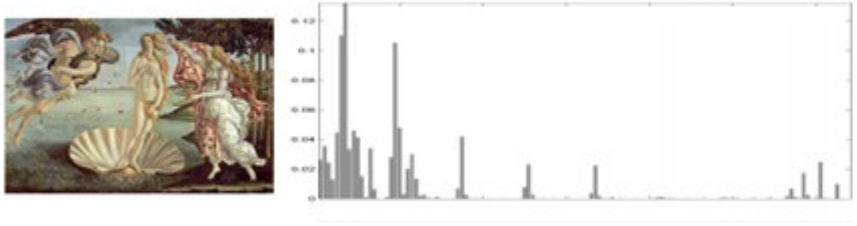
+ Colour Feature

- Number of bins
  - More bins are defined in the histogram, more discrimination power
    - It will increase the computation cost
      - R*G*B=256*256*256 = 16,777,216 bins
    - In some cases, it might not be necessary for image retrieval
  - Instead, use predefined colour bins + percentage

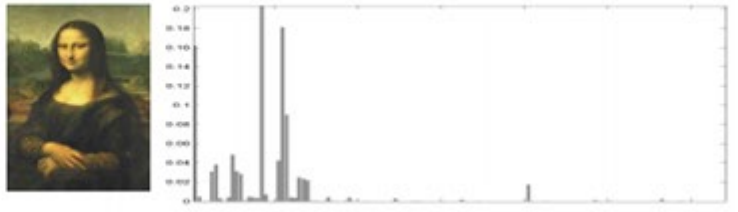


$$\left\langle (I_i, P_i) \middle| I_i \in ColorValue, 0 \le P_i \le 1, \sum_{i=1}^{N} P_i = 1 \right\rangle$$

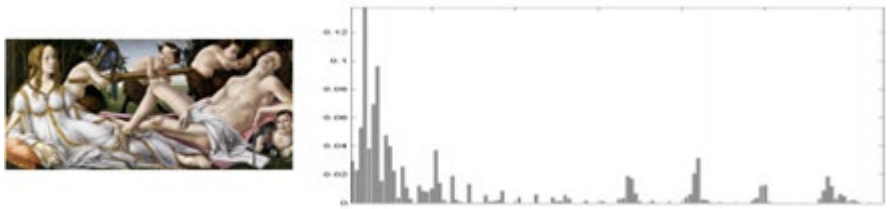
+ An Example



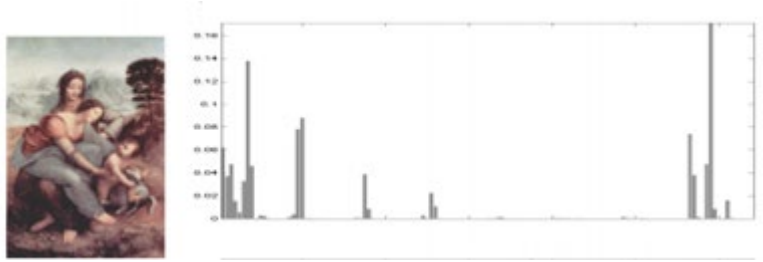
The Birth of Venus, Sandro Botticelli, probably mid 1480s



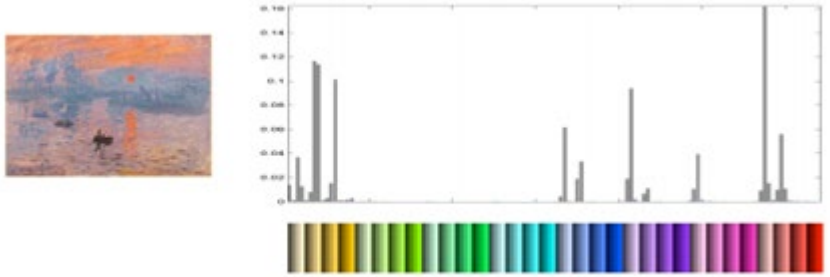
Mona Lisa, Leonardo da Vinci, 1503–1506, perhaps continuing until 1517



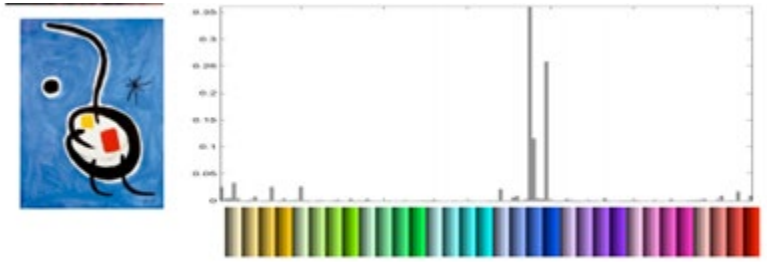
Venus and Mars, Sandro Botticelli, 1485



The Virgin and Child with Saint Anne, Leonardo da Vinci, probably 1503



Impression, Sunrise, Claude Monet, 1872



Figure, Star, Joan Miró, 1978

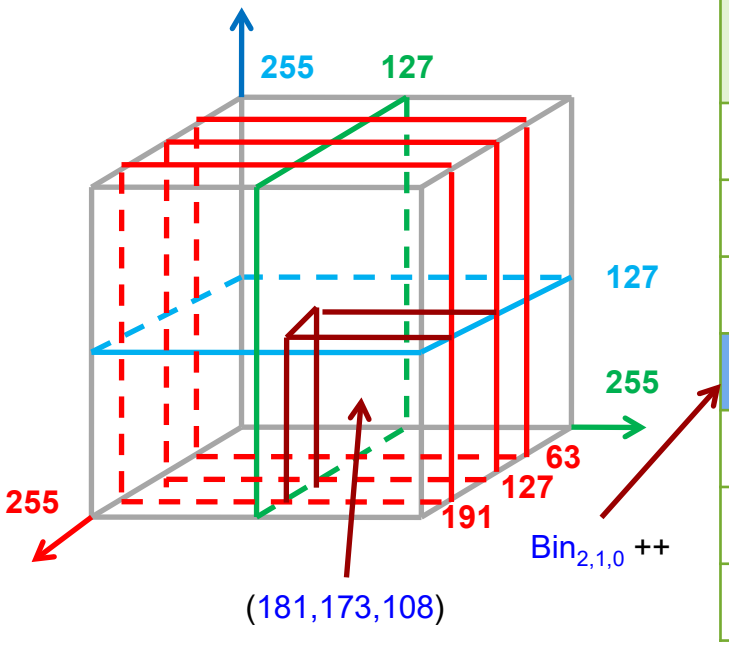
Colour Histogram Construction

- Quantization of the colour space
  - A bin is also referred as a dimension
- Example: map the 24bit RGB colour space to 16 bins
  - Define number of quantiles for each colour
    - R: 4-quantile
    - G: 2-quantile
    - B: 2-quantile
  - Generate a colour histogram of 16 dimensions (2^{24} colours quantized into 2^4 bins)
    - 4×2×2 = 16 dim

+ Illustration

R: 4-quantile
G: 2-quantile
B: 2-quantile

➡ 4×2×2 = 16 combinations → 16 dim



Bin	R	G	B	#pixs
C_0	0	0	0	2
C_1	0	0	1	10
		…		
C_{10}	2	1	0	15
		…		
C_{14}	3	1	0	3
C_{15}	3	1	1	0

$V_0 = (2,10,\ldots,15,\ldots,3,0)$

Normalization

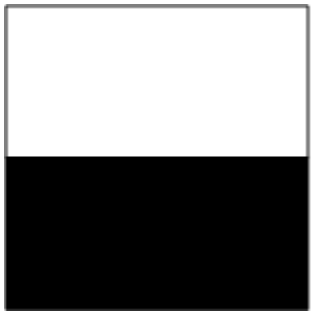
$V = (0.03,0.16,\ldots,0.23,\ldots,0.05,0)$

Colour Histogram Property

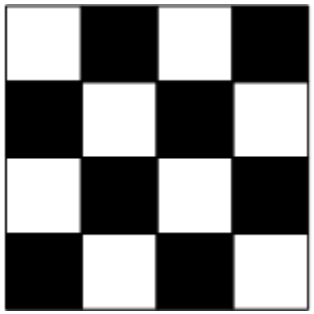
- A representation of the statistical distribution of colours in an image
- Colour histogram is relatively invariant to
  - Colour model
  - Rotation
  - Scaling
- Issues
  - Lost other information, such as location, boundary, etc.
  - Not good for some editing operations, such as cropping

+ Texture vs. Colour

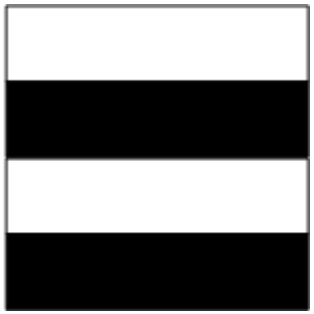
- Same colour distribution but different texture



Block pattern



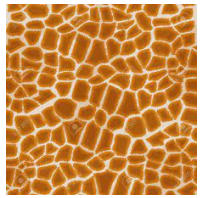
Checker board



Striped pattern

+ Texture Feature

- Texture is a repeated pattern of intensities
  - "An image region has a constant texture if a set of its local properties in that region is constant, slowly changing, or approximately periodic."
  - Applications: satellite images, medical images
- Useful for describing and reproducing contents of real-world images
  - Such as clouds, fabrics, surfaces, wood, stone patterns
- Texture is easy for human users to recognise, but hard to define and compute because:
  - Rotation and scale invariance



giraffe

tiger

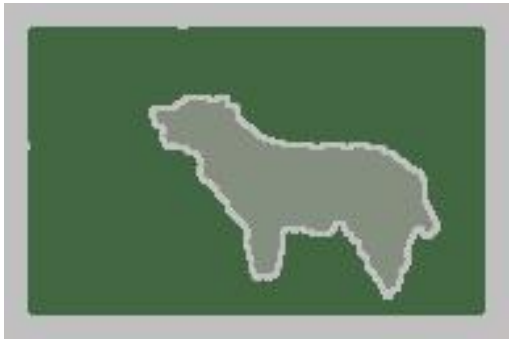
snake

fish

Peter Howarth, Stefan M. Rüger. Evaluation of Texture Features for Content-Based Image Retrieval. CIVR 2004: 326-334.

+ Shape Feature

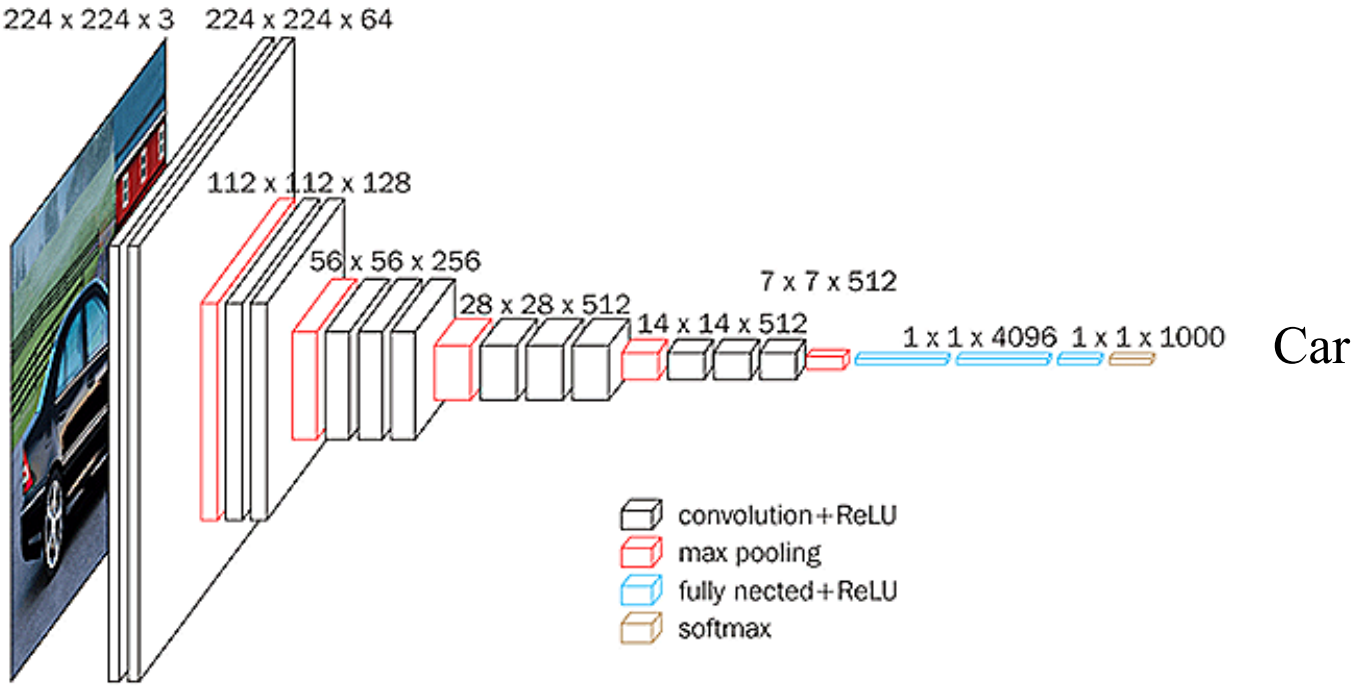
- Geometric properties for still image
- Use a set of shape descriptors
  - Shapes, their center, bounding box, circularity etc.
- Can be contour-based, or region-based in the space domain or a transformed domain



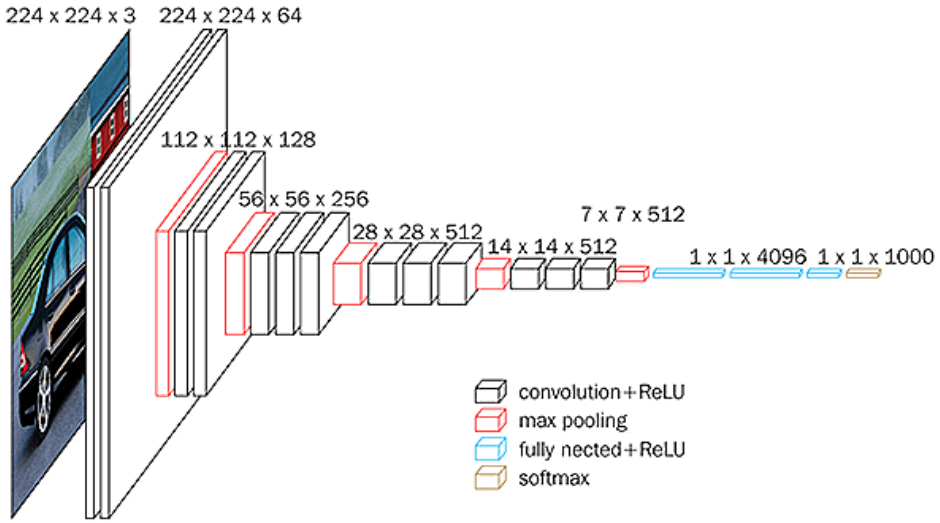
Mingqiang Yang, Kidiyo Kpalma, Joseph Ronsin. A Survey of Shape Feature Extraction Techniques. Peng-Yeng Yin. Pattern Recognition, IN-TECH, pp.43-90, 2008.

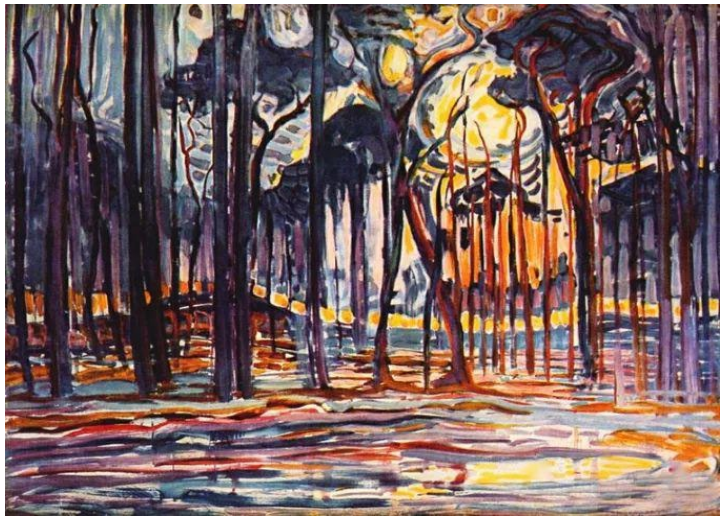
+ Representation Learning by Neural Networks*

- Neural networks: a mapping from image to a class (label), layer by layer with inner layers as hidden layers
- Learned representation: hidden layer output

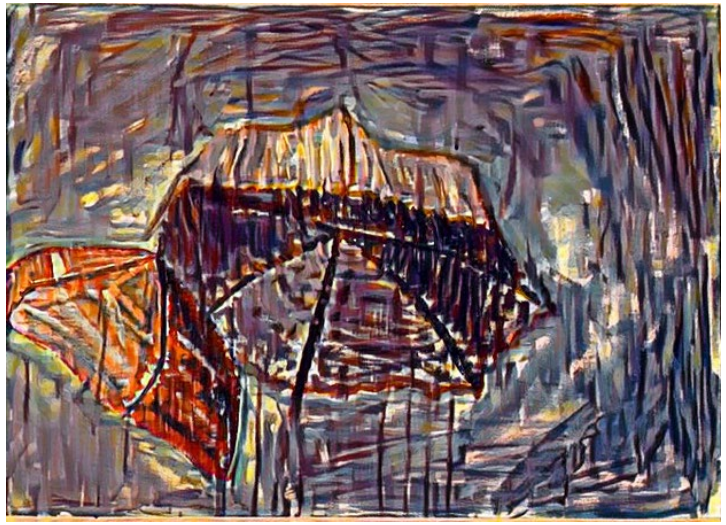


+ Example: Style Transfer*





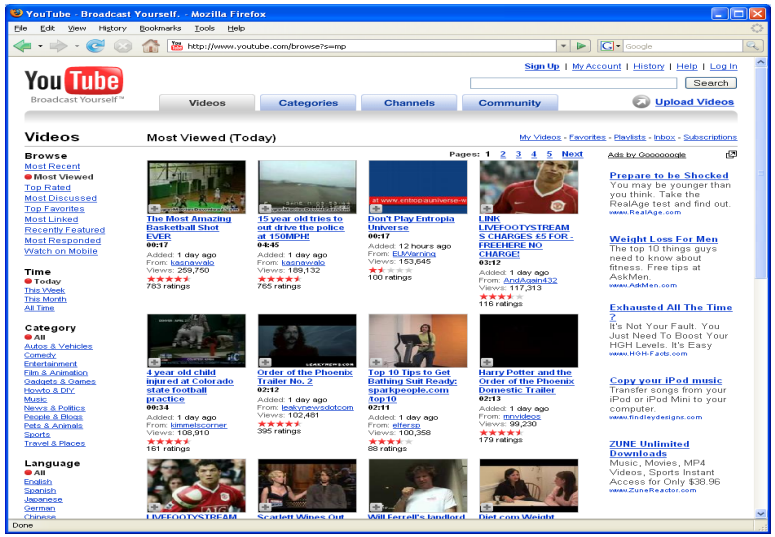
Woods near Oele, Piet Mondrian, 1908



Leon A. Gatys, Alexander S. Ecker, Matthias Bethge. Image Style Transfer Using Convolutional Neural Networks. CVPR 2016

+ Video

- Video is a sequence of pictures …
  - Analog
  - Some with rich information such as sound, text, annotations and metadata…
- Compression
  - 30 frames/sec
  - Resolution: 640 × 480
  - 24-bit RGB
  - Storage: ~26 Mbytes/sec (uncompressed)
  - Most likely in compressed forms



+ Video Feature Levels

Level	Granularity	Description
Video	Meta	Concept, Producer, Director
Scene	Macro	Event description
Shot	Mini	Action, talk, goal
Frame	Micro	Objects & spatial relationships

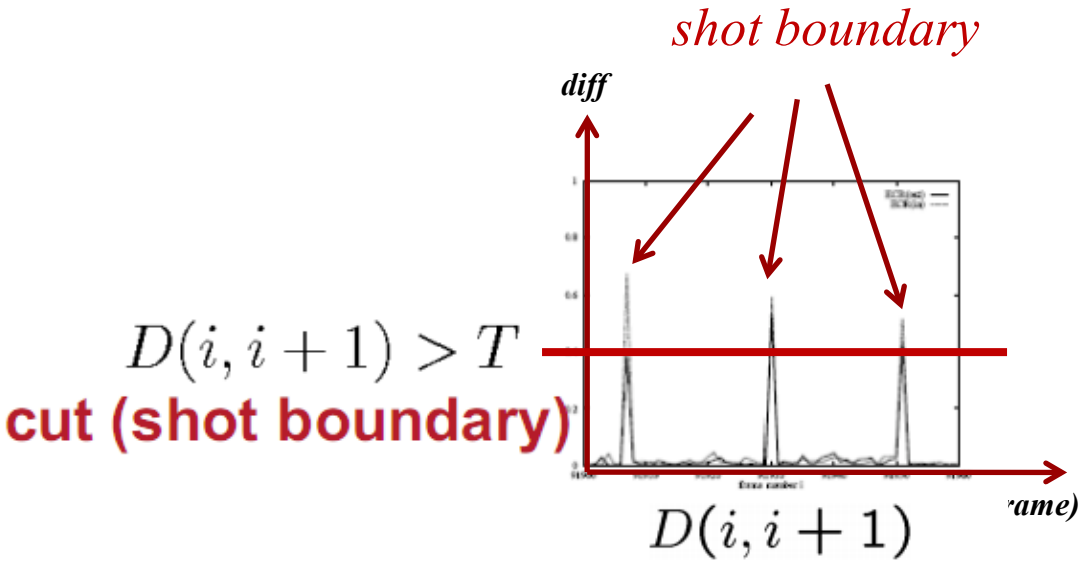
https://en.wikipedia.org/wiki/One-shot_film
https://en.wikipedia.org/wiki/List_of_one-shot_music_videos

Shot Boundary Detection

- Shot
  - A **collection** of contiguous video frames depicting the same action in time and space
  - A basic unit for further video applications
    - Classification – high-level concept detection
    - Search – index on keyframes
- Key Frame
  - **Representative** (image) frame of the shot
    - The first frame of the shot
    - The clearest one
    - The most common one
    - …
- Shot boundary detection
  - By detecting **frame difference** based on colour, motions, etc.

+ Shot Boundary Detection: Example

- For example – colour distribution feature comparison
  - Evaluating the differences on colour features of two successive frames
  - e.g., Euclidean Distance
    - $\text{frame}_i = (x_1, \ldots, x_n)$, $\text{frame}_{i+1} = (y_1, \ldots, y_n)$, $D(i,i+1) = (\sum_{k=1}^{n} (x_k - y_k)^2)^{\frac{1}{2}}$



+ Example: Video Feature Space

video image image feature space

https://www.youtube.com/watch?v=5bpYV-YvcN8

Other Types of Features: VSM*

- The vector space model (VSM) for text data
  - Each word in the vocabular is a dimension
    - Stemming can be applied
    - The value for a dimension can be 0/1, or tf-idf weighted
  - Cosine similarity to measure document similarity
- Simple, and supports similarity ranking and partial matching
- However
  - Suffers from the curse of dimensionality
  - Has no sequence information
  - All terms are assumed statistically independent (synonyms and corelated terms not recognized)

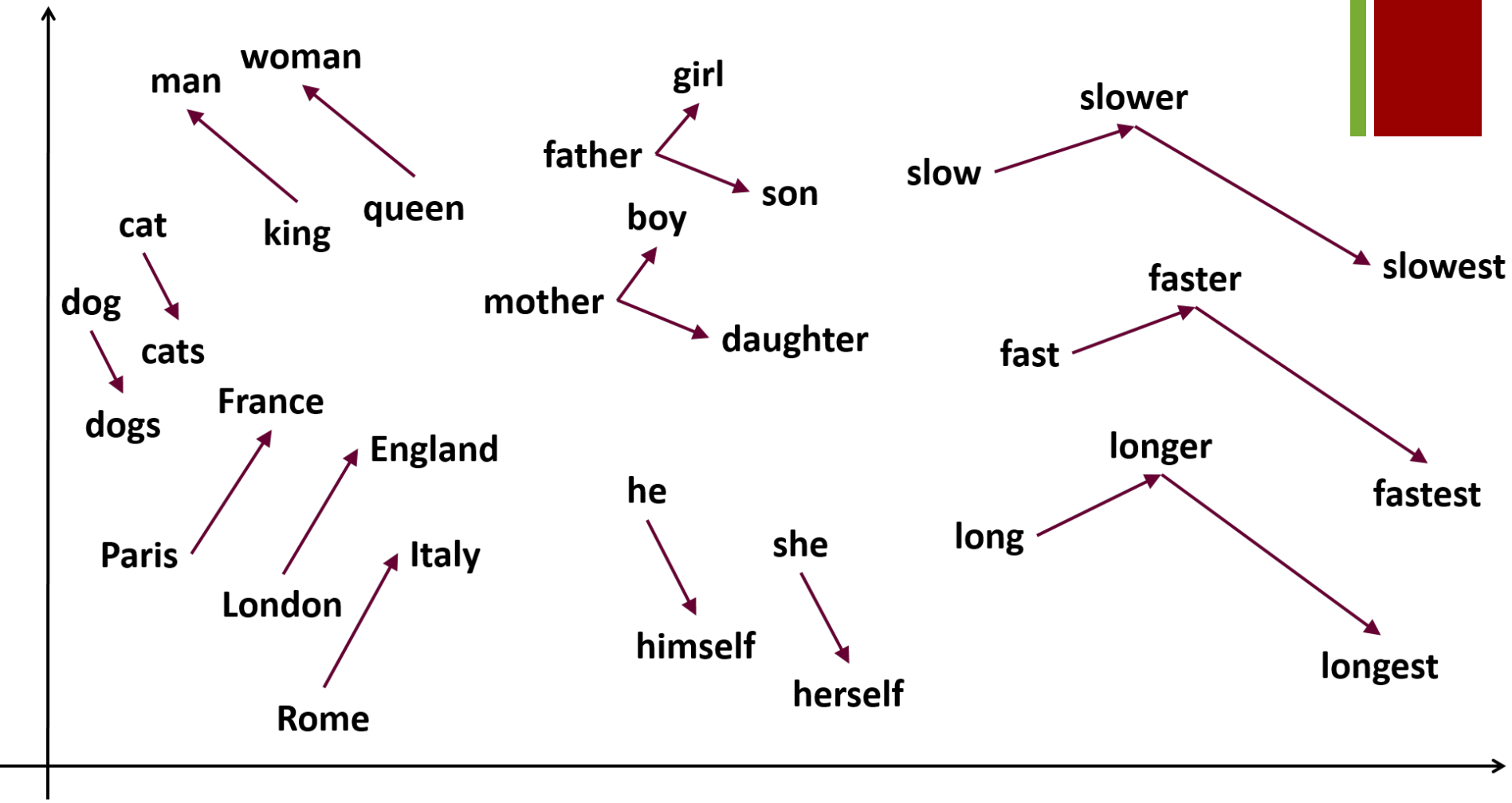
Other Types of Features: Word2Vec*

- Used to produce word embeddings, with two-layer neural networks trained to reconstruct linguistic contexts of words
  - Input: a large corpus of text (words with their context)
  - Output: a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space
  - Words that share common contexts in the corpus are located in close proximity to one another in the space
- Node2vec
  - For network embeddings (random walks from each node)

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space. ICLR 2013

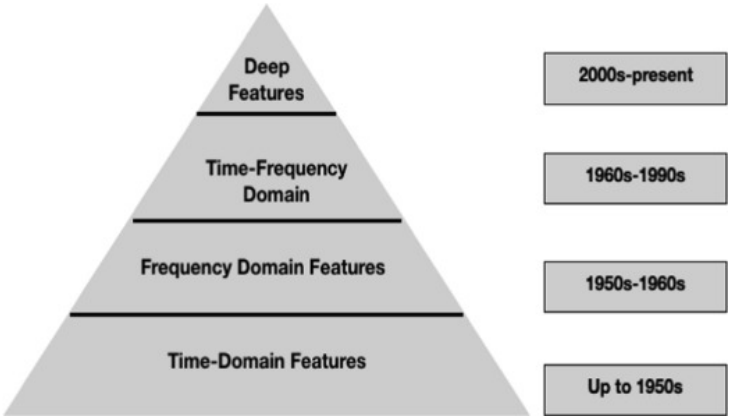
Aditya Grover, Jure Leskovec: Node2vec: Scalable Feature Learning for Networks, ACM SIGKDD, 2016

+ Word2Vec: Illustration



+ Abstraction for Audio/Speech*

- Audio data
  - Garima Sharma, Kartikeyan Umapathy, Sridhar Krishnan. Trends in audio signal feature extraction methods. Applied Acoustics 158, 2020.

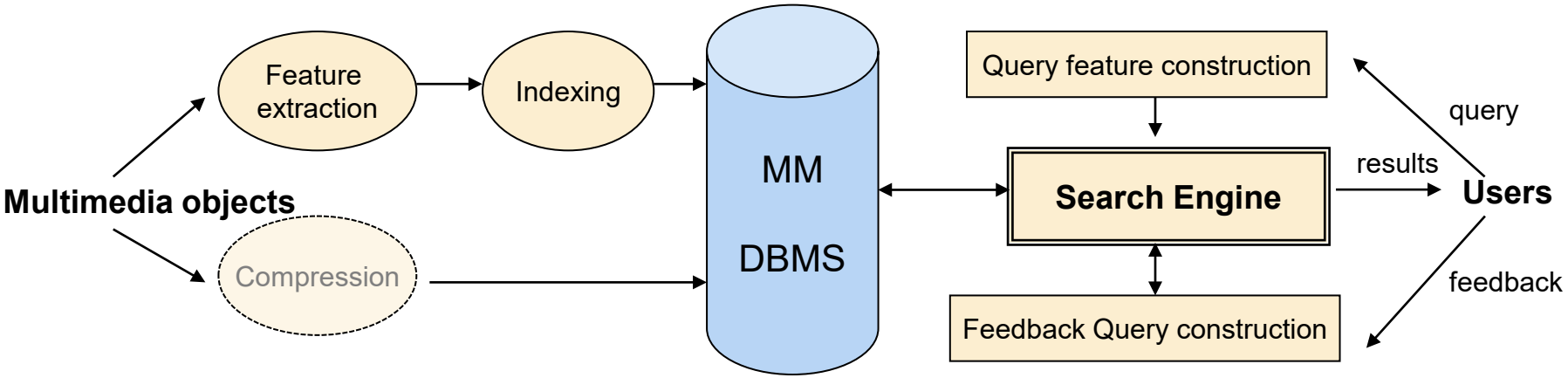


- Speech data
  - Daniel Jurafsky and James H. Martin. Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition 2ed, 2009. Chapter 9

Main Components of MMDBMS

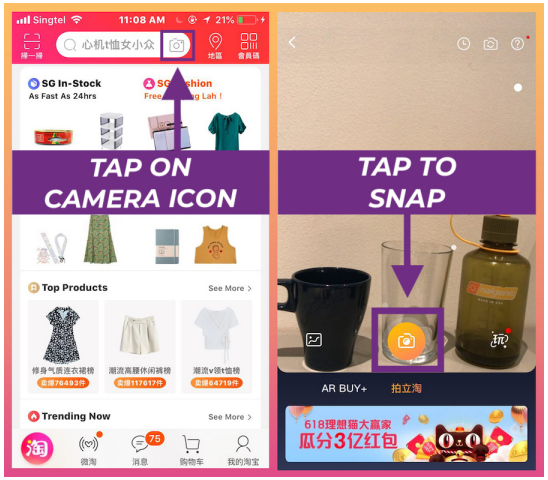
- MMDBMS **stores** and **organizes** both original data and their abstractions
- **Features/signatures** are extracted and indexed to facilitate the retrieval and search
- **Indexing** is the primary method in organizing data to achieve efficient data access
- The goal is to deploy effective indexing methods to achieve accurate and fast retrieval

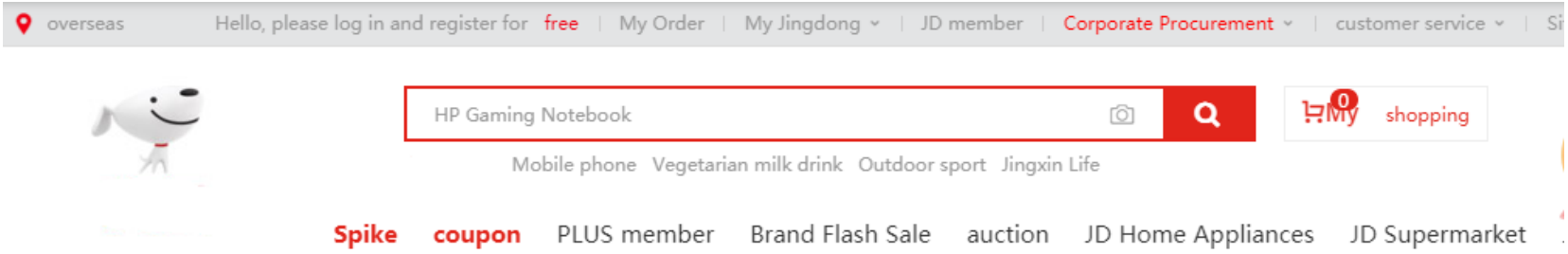
A Generic Architecture of MMDBMS



+ Similarity Search

- Definition
  - **Given** a set of multimedia objects in database, **find** the ones similar (or top *K* most similar) to a desirable query object quickly!
- Applications
  - Image similarity search
  - Video-clip near-duplicate detection





Content-Based Retrieval

- A major issue – CBR (Content-Based Retrieval)
  - Search the database based on ***similarity*** of media content (i.e., similarity search), e.g.,
    - Find similar images from the database
    - Identify the position of a short query video in a long video
    - Retrieve the most similar video clip w.r.t the query clip
- It should support multiple features/media retrieval
  - E.g., find a similar image in both colour and texture
- Object recognition techniques can help
  - Require human annotations
  - Semantics can be complex

+ Text vs. Content-based Search

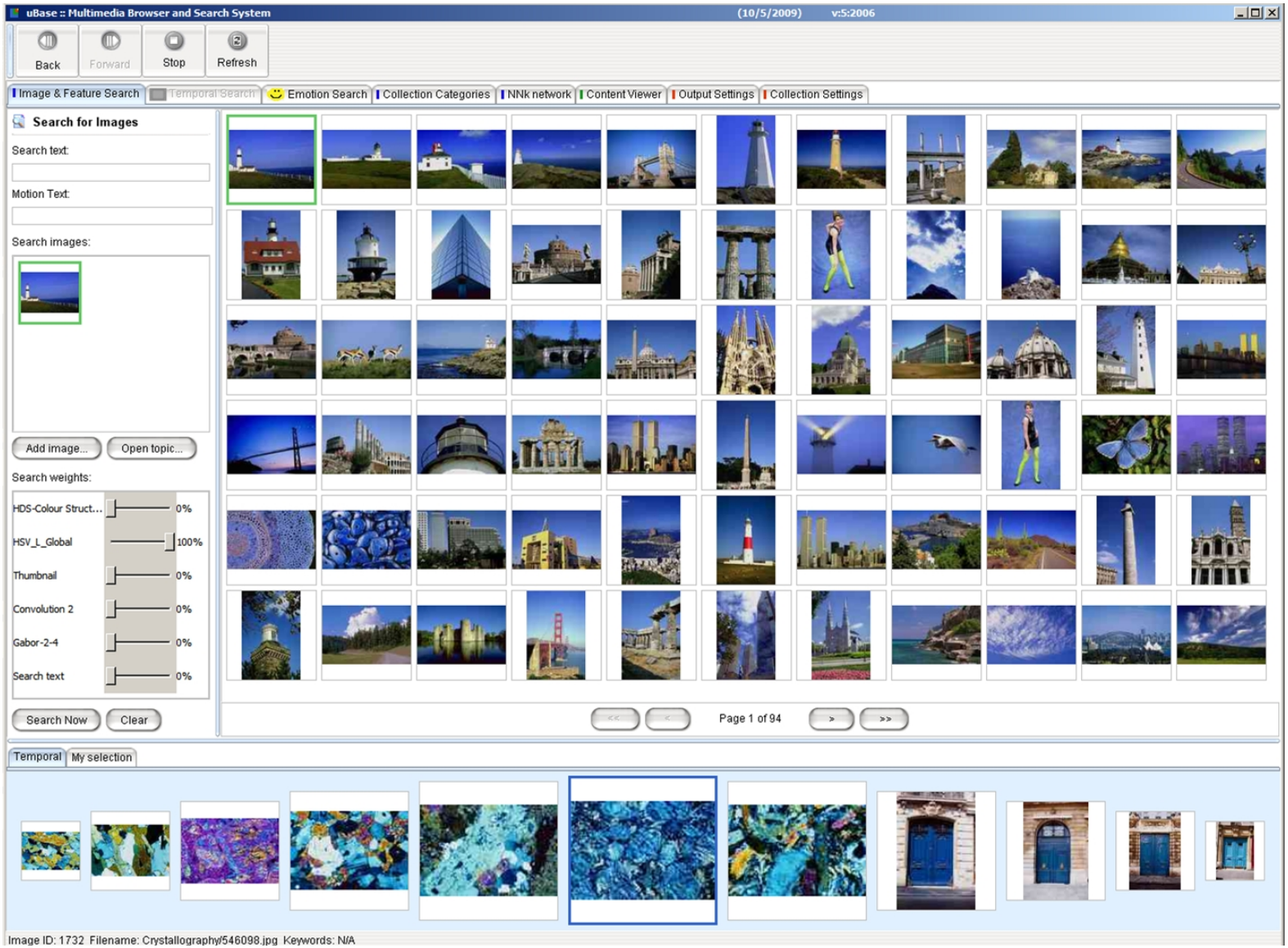
- Traditional text-based search
  - Keyword matching (e.g., exact search)
- Multimedia retrieval
  - **Text-based approach** for indexing and retrieval
    - Keywords/tags
    - Annotation—by human: labor intensive, subjective and partial, no support for visual properties
    - Some queries may be too complex by text
  - **Content-based Retrieval**
    - No exact match , but **similarity** match (evaluated by precision and recall)

Characteristics of MM retrieval

- Large amounts of complex data
- Use the **feature-based** approach
  - Use high-dimensional data
  - Need high-dimensional index technique for efficient query processing to achieve faster-than-sequential searching
- Use **similarity** search
- Integration of **multiple features** is common
  - E.g., an image database may contain colour histogram, shape, location, and texture

Case 1: Content-based Image Retrieval

+ Image Similarity Search



Similarity Measure

- **Database**
  - A set of multimedia objects (or simply points) in a *d*-dimensional data space (i.e., feature space)
- **Query**
  - A *d*-dimensional feature vector extracted from the query object
- **Measure**
  - Distance between two objects *p* and *q* is *f*(*p*,*q*), where *f* is the distance measure function
    - Small Distance = High Similarity

L_p Distance Functions

- Given two n-dimensional points $\mathbf{x} = [x_1, \ldots, x_n]$ and $\mathbf{y} = [y_1, \ldots, y_n]$

p	Minkowski Dist	$(\sum_1^n \lvert x_i - y_i \rvert^p)^{\frac{1}{p}}$
$p = 1$	Manhattan Dist	$\sum_1^n \lvert x_i - y_i \rvert$
$p = 2$	Euclidean Dist	$(\sum_1^n \lvert x_i - y_i \rvert^2)^{\frac{1}{2}}$
$p = \infty$	Chebyshev Dist	$\max(\lvert x_1 - y_1 \rvert, \ldots, \lvert x_n - y_n \rvert)$

- Weighted distance $(\sum_1^n w_i \cdot \lvert x_i - y_i \rvert^p)^{\frac{1}{p}}$
  - Adjusted based on the feedback

+ Results Evaluation

- Similarity value may not 100% imply 'real similarity' between two objects
  - Features may not represent objects accurately
  - Similarity measure may not be able to capture the real closeness of two feature vectors
  - Hence retrieved results may not be 100% correct
- Correct (relevant) results are usually manually-identified by users
  - Correct results could be subjective since different people have different perception
  - Retrieved results can then be evaluated

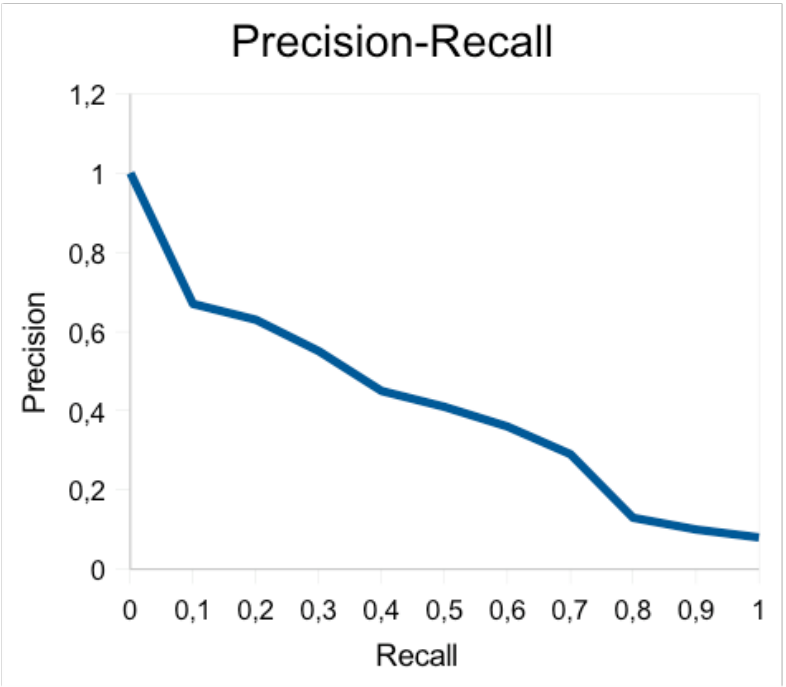
+ Precision and Recall

- Retrieval effectiveness evaluation
  - Precision and Recall
  - Similar to Information Retrieval (Search Engines)

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

+ Precision vs. Recall



$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

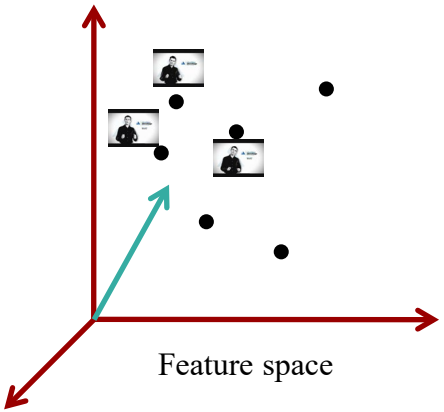
- Higher Recall
  - Return more retrieved files, but may suffer from lower precision
- Higher Precision
  - May be increased by reducing the number of retrieved files, but may suffer from lower recall

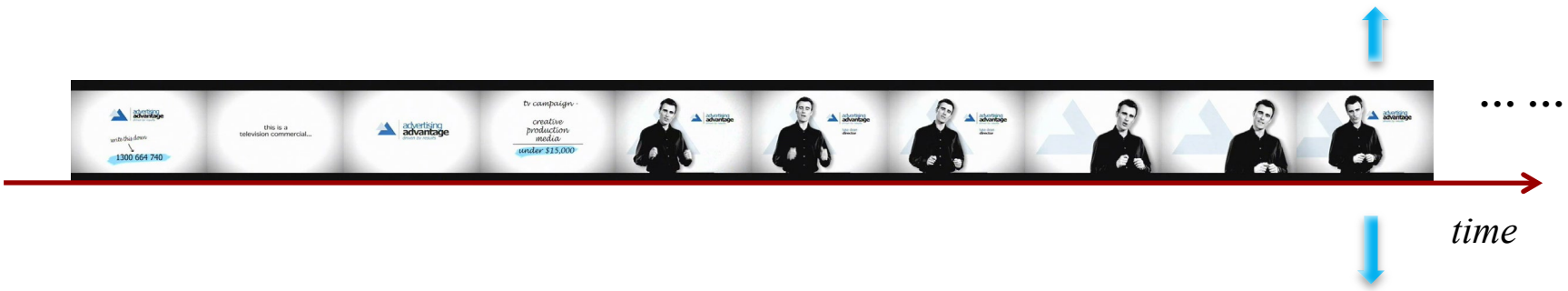
Case 2: Content-based Video Retrieval

+ Video Data Representation

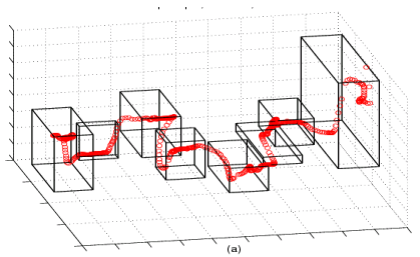
- Videos -> Set of points
  - Loss of sequence information

When neglect temporal information





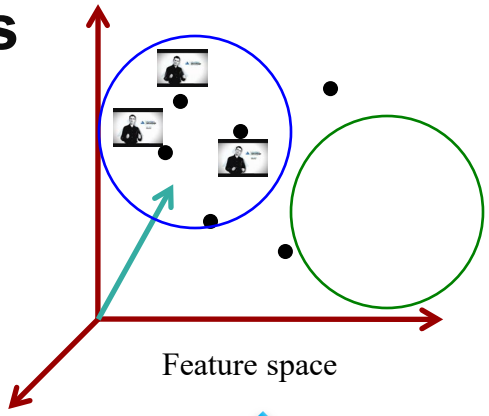
When incorporate temporal information



+ Video Triplet (VITRI)

[Shen et al. , 2005]

- Basic idea: summarize each video into a set of clusters, each of which is modelled as a hypersphere described by a **video triplet** *(position, radius, and density)*
- Each video is then represented by a much smaller number of **hyperspheres**



Heng Tao Shen, Beng Chin Ooi, Xiaofang Zhou. Towards Effective Indexing for Very Large Video Sequence Database. SIGMOD Conference 2005: 730-741

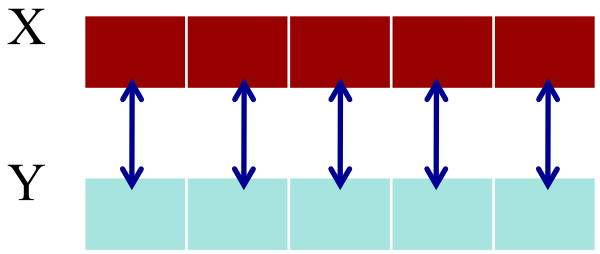
When Incorporate Temporal Info.

- Sequence of frame feature vectors
  - Similar to **time series** …
    - High-d time series
  - Various time series measures can be utilized
    - Mean distance
    - Dynamic Time Warping (DTW)
    - …

+ Mean Distance

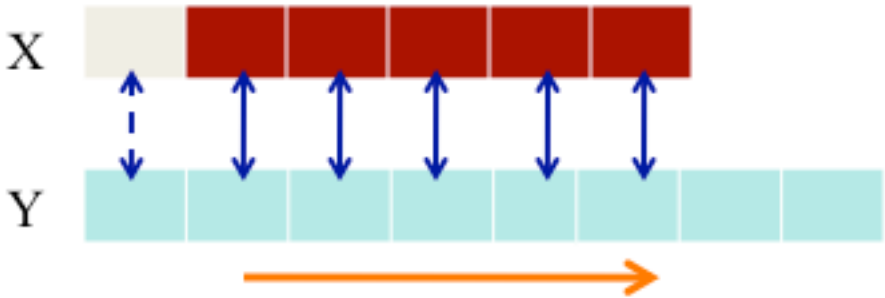
- Given two video sequences $X = x_1 \ldots x_n$ and $Y = y_1 \ldots y_m$
- For equal length sequences $(n=m)$

$$D(X,Y) = D_{\text{mean}}(X,Y) = \frac{\sum_{1 \le i \le n} d(x_i, y_i)}{n}$$



- For variable length sequences $(n<m)$

$$D(X,Y) = \min_{1 \le i \le m-n+1} D_{\text{mean}}(X[1:n], Y[i:i+n-1])$$



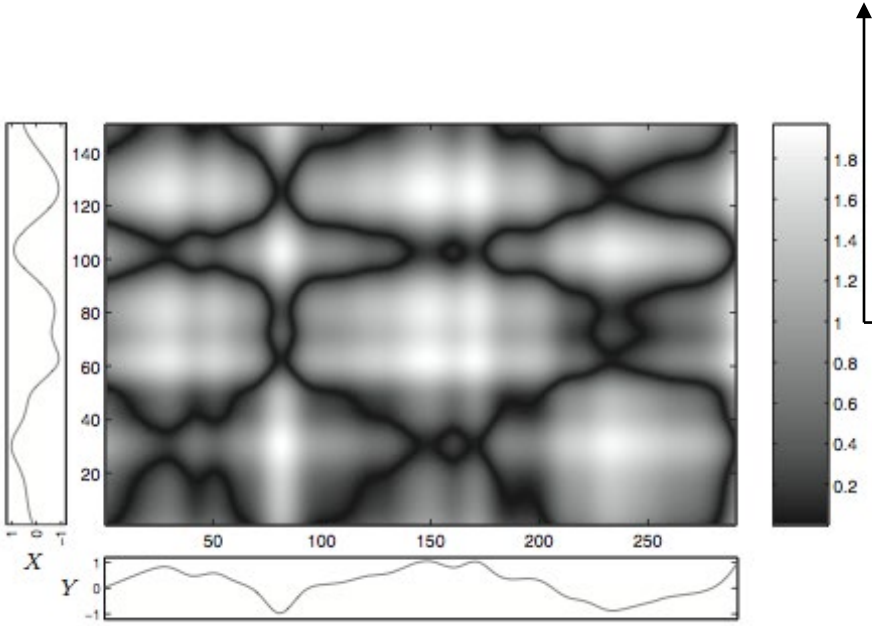
+ Mean Distance: Discussions

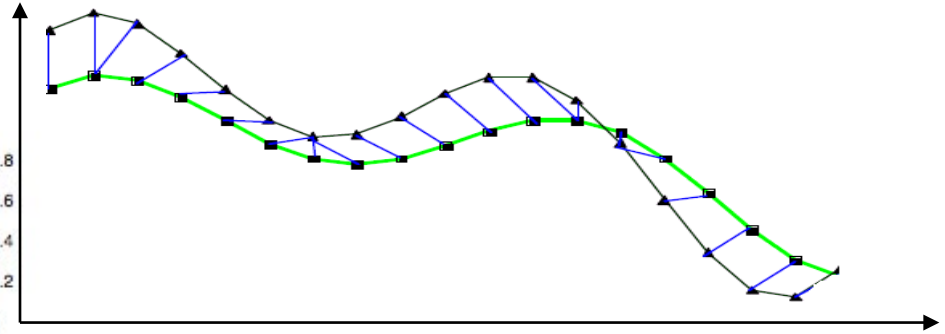
- Used for video similarity measure
  - **Advantages**
    - Consider the difference in sequence length
    - Explore the inter-frame similarity
  - **Disadvantages**
    - Does not support frame alignment or gap

+ DTW: Basic Idea

- Recursion function

$$\mathrm{DTW}(i,j) = \mathrm{Dist}(i,j) + \min[\mathrm{DTW}(i-1,j), \mathrm{DTW}(i,j-1), \mathrm{DTW}(i-1,j-1)]$$





+ DTW: Discussions

Used for video similarity measure

- **Advantages**
  - Handle frame alignment for video sequence matching
  - Preserve temporal order
- **Disadvantages**
  - Have to compare each frame pair, basic implementation $O(n^2)$, where n is the length of the sequences
  - No element can be skipped in a sequence, even it is just a noise frame

+ Readings

- Books:
  - *Principles of Multimedia Database Systems*, by V. S. Subrahmanian1997, Morgan Kaufmann Publishers, Inc.
- Papers:
  - Heng Tao Shen, Beng Chin Ooi, Xiaofang Zhou, "Towards Effective Indexing for Large Video Sequence Data", **SIGMOD** 2005
- Other Resources
  - Wikipedia Page: Content-based image retrieval
  - Wikipedia Page: Multimedia database

+ Summary

- Covered:
  - What is MM?
  - Feature representations of MM.
  - Abstraction of MM.
  - Architecture of MMDBMS.
  - CBIR and Video search
- Next:
  - Route Planning in Road Network (Mengxuan Zhang)