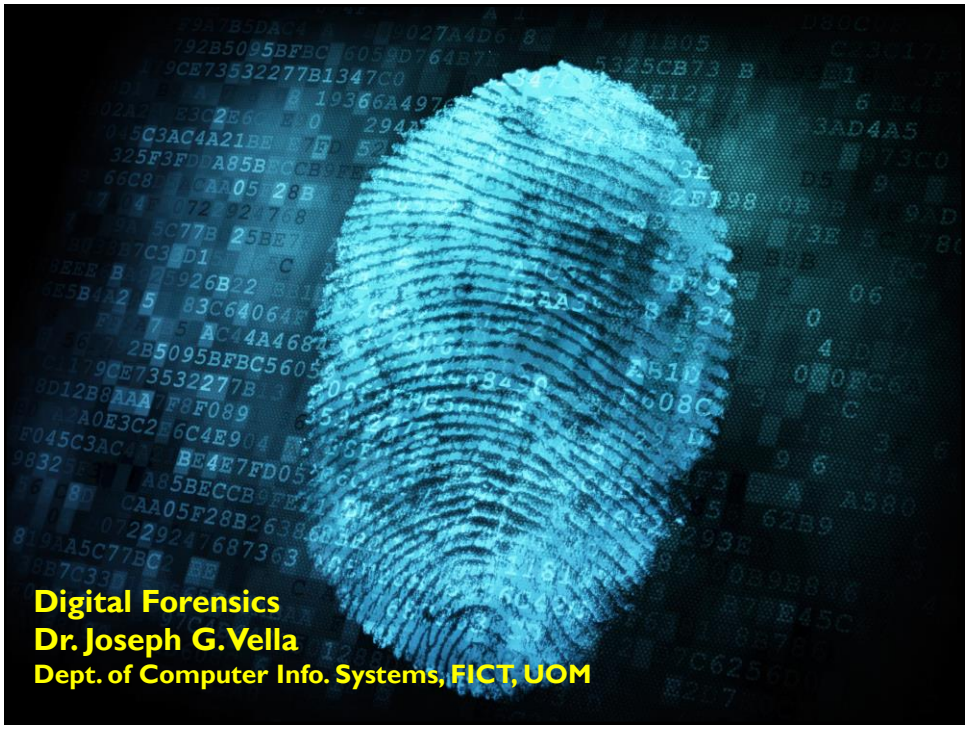


IMAGE PROCESSING FOR CONTENT-BASED VISUAL INFORMATION RETRIEVAL

CBIR, QBIC, CBVIR

J Vella – Digital Forensics

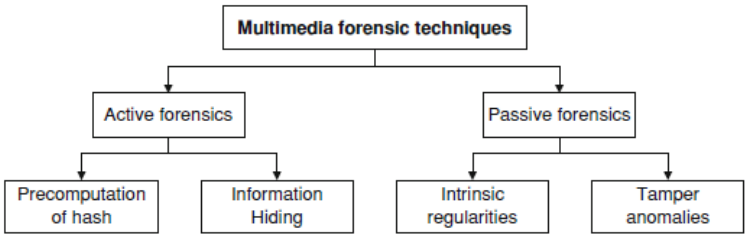
Image Processing

Introduction

- A digital forensic investigation might deal with an examination of numerous number of pictures.
  For example:
  - Finding, or find traces of, illicit images;
  - Case specific images:
    - Non-disclosable prints (even harder as no reference database can be built);
    - Finger prints;
    - Foot and tire prints;
    - Face recognition;
    - Location finder;
    - Injuries.
- Scope of this session:
  - Number of images in a data store might be large;
  - Images can be stored in personal devices (PC, smart phones etc.), company's hardware, cloud storage, and ISP.
  - The security of the reference database is usually of very high requirement.

Image forgery

Forensic techniques classification



Design Requirements

- **Performance**:
  - Quick response and good levels of performance.
  - What is a good match?

DEFINITION: **Rank** of an input image, is the ordinal number of the first image hit. In many practices it's labelled a good hitter if it finds a corresponding image in the first 100.

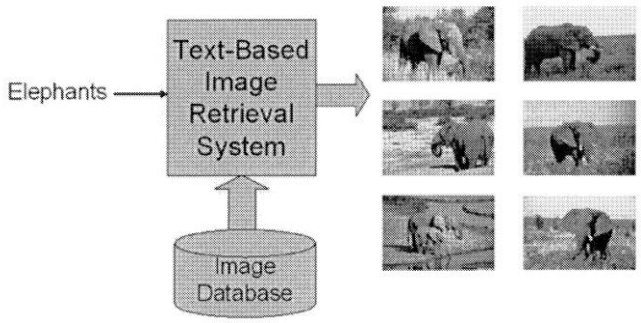
- *Scalability*
  - To photo size, photo colour depth, etc.
- *Security*
  - Most datasets are confidential. Query response might need to be rethought – i.e. not necessary an image is returned.
- *Flexible deployment*
  - Image capture, image format, subject content, query definition (by example, by labels, by a sketch)

Generic Techniques (i) – text based

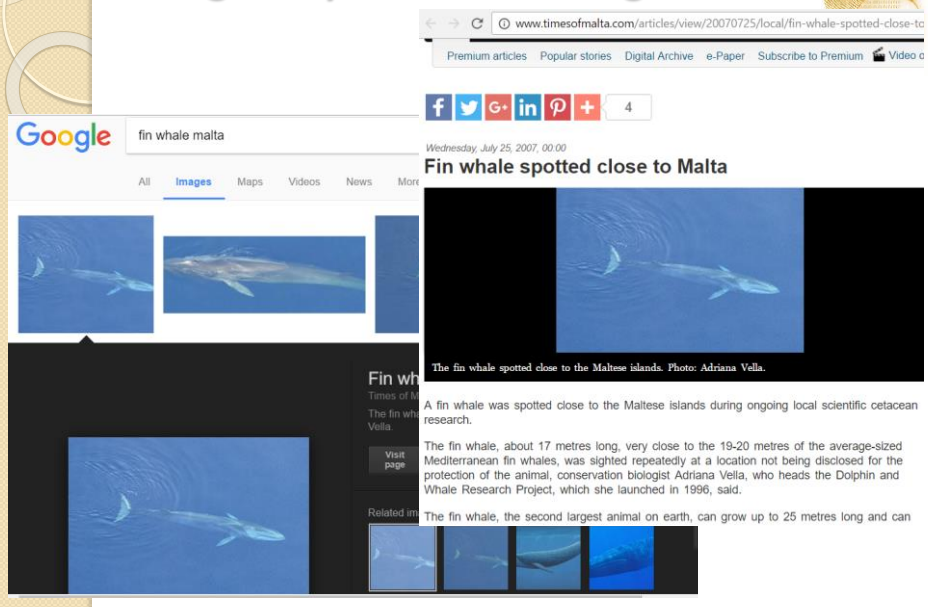
- Pre-work: Encode each picture with text based labels or tokens.

Either:
The query picture is to be encoded and its best matching returned;

Or:
A list of labels are inputted and best matching pictures are returned.

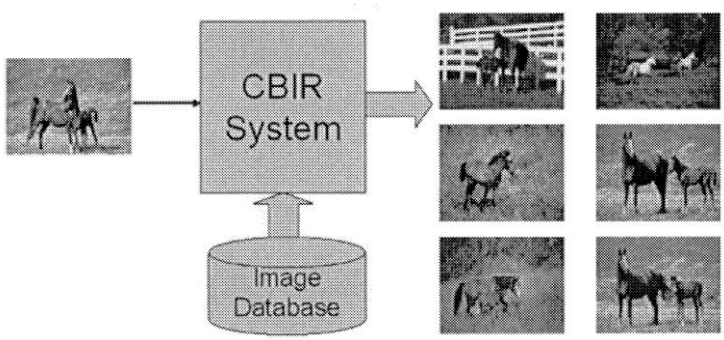


Google keywords in Images



Gen. Techniques (ii) – content based

- Pre-work:
  Pre-process each picture (e.g. clean and standardise size and colour depths);
  Calculate features (at least a few are needed) for each picture's content.
  - Features include: shape, texture, colour intensities.
- Pre-process input picture and calculate features.
  - Find matches based on some **similarity** measure.



Google search by image

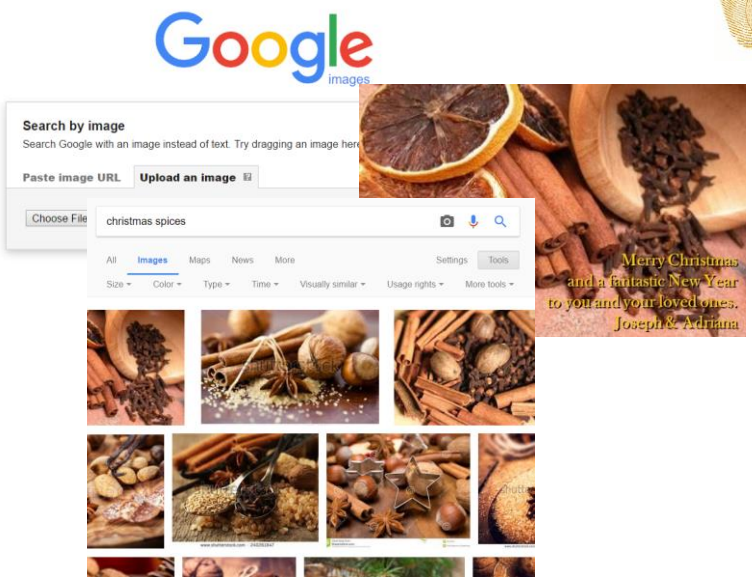
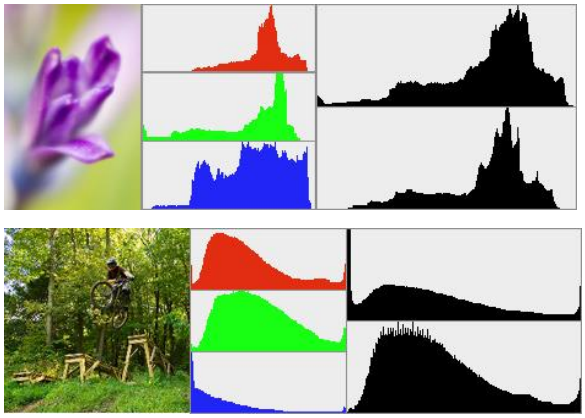


Image histograms



Similarity measures

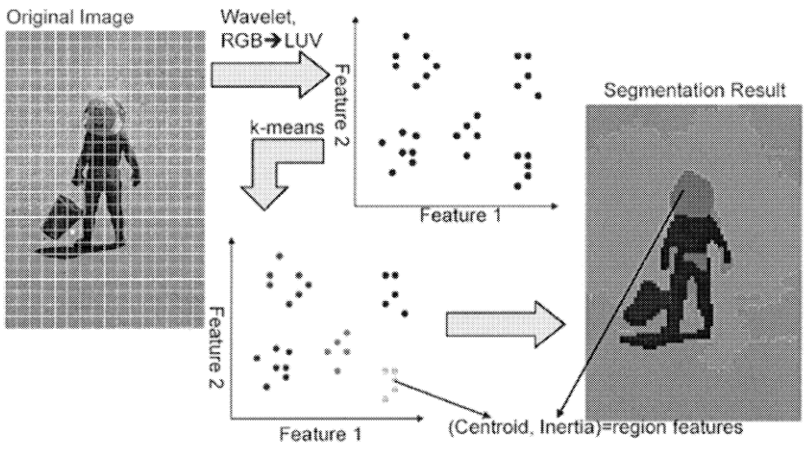
- Calculated on whole picture, as in histogram matching, or part of picture, as in object decomposition.
- Major issue with histogram matching is its sensitivity colour intensity, colour distortions and cropping of images.
  - Also histogram slippage …
- An advantage of decomposed methods is that one can get good hits, high similarity, on a fragment rather than average the similarities across the whole picture.
  - In such segmentation process one must balance the texture effectiveness vs computation.
    - Large segments compute quick, but lose texture signature.

DECOMPOSITION METHOD

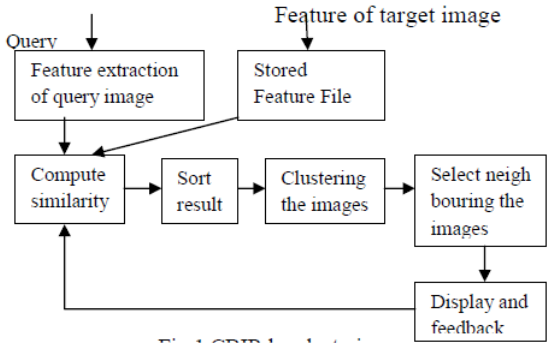
Similarity measure in SIMPLicity

- A pioneering approach (in SIMPLicity, Wang et al 1998).
- First it computes six features:
  - three average colour values (ie LUV – luminance and chrominance), and
  - other three from the energy in high frequency of the wavelet transform – these are known to represent image texture.
    - The wavelet transform allows for directional texture capture – horizontal (increasing and decreasing), and vertical (increasing and decreasing).
- Then the k-means algorithm is used to cluster the feature vectors. The system loops upwards on k until a criterion is reached.
  - When the criterion is met, k's 6 dim vectors are calculated from the k's centroids.
- After clustering, three extra features are added (ie called inertia of orders 1 to 3) to describe shape properties.
- The similarity is computed for two images is defined by a weighted sum of distances in the feature space and between each images regions.

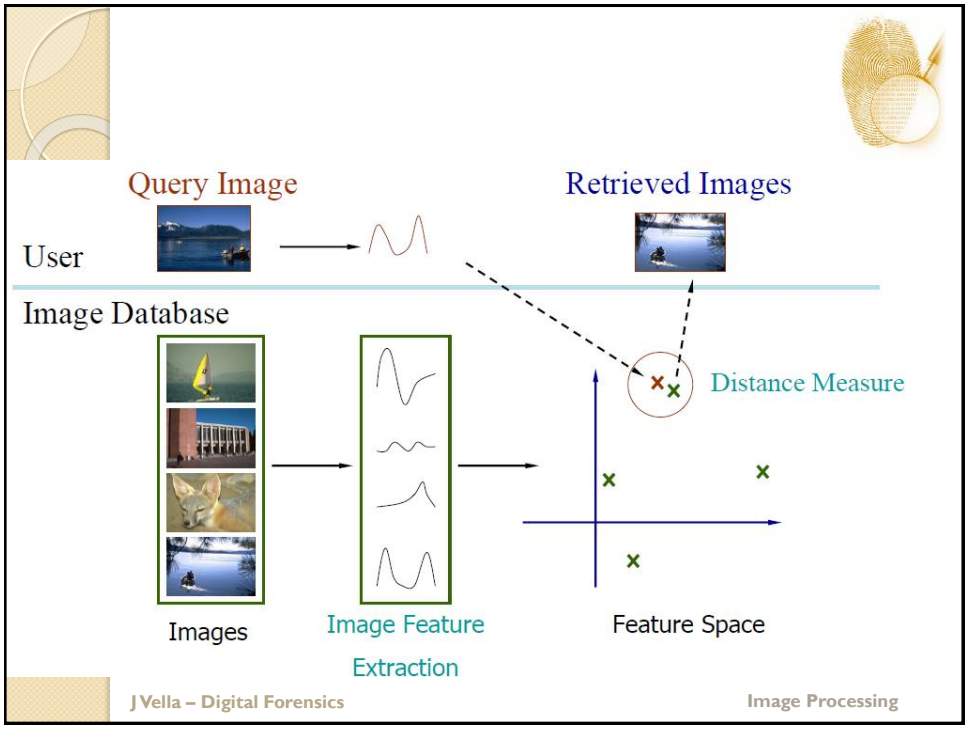
Similarity measure in SIMPLicity



Similarity measure in SIMPLicity



WHOLE IMAGE (HISTOGRAM)



The QBIC color histogram distance is:

$$d_{hist}(I,Q) = (h(I) - h(Q))\ A\ (h(I) - h(Q))$$

- h(I) is a K-bin histogram of a database image
- h(Q) is a K-bin histogram of the query image
- A is a K x K similarity matrix

Image retrieval in digital forensics

- General cases
- Particular investigations
  - Twirl paedophile