Multimedia Retrieval

Seyed Hamidreza Mohammadi

CSE 506/606 - Topics in Information Retrieval

Oregon Health and Science University

Fall 2012

Paper 1

- Paper 1

- Müller, H., Michoux, N., Bandon, D., & Geissbuhler, A. (2004). **A review of content-based image retrieval systems in medical applications-clinical benefits and future directions**. *International journal of medical informatics*, *73*(1), 1-24.

Paper 1

- The paper reviews the literature for Content-Based Image Retrieval (CBIR)
- There is a growing amount of visual and multimedia data.
- This is also the case for Medical field.
  - Radiology, Cardiology, Endoscopy.

Overview

1- Intro. to Image Retrieval

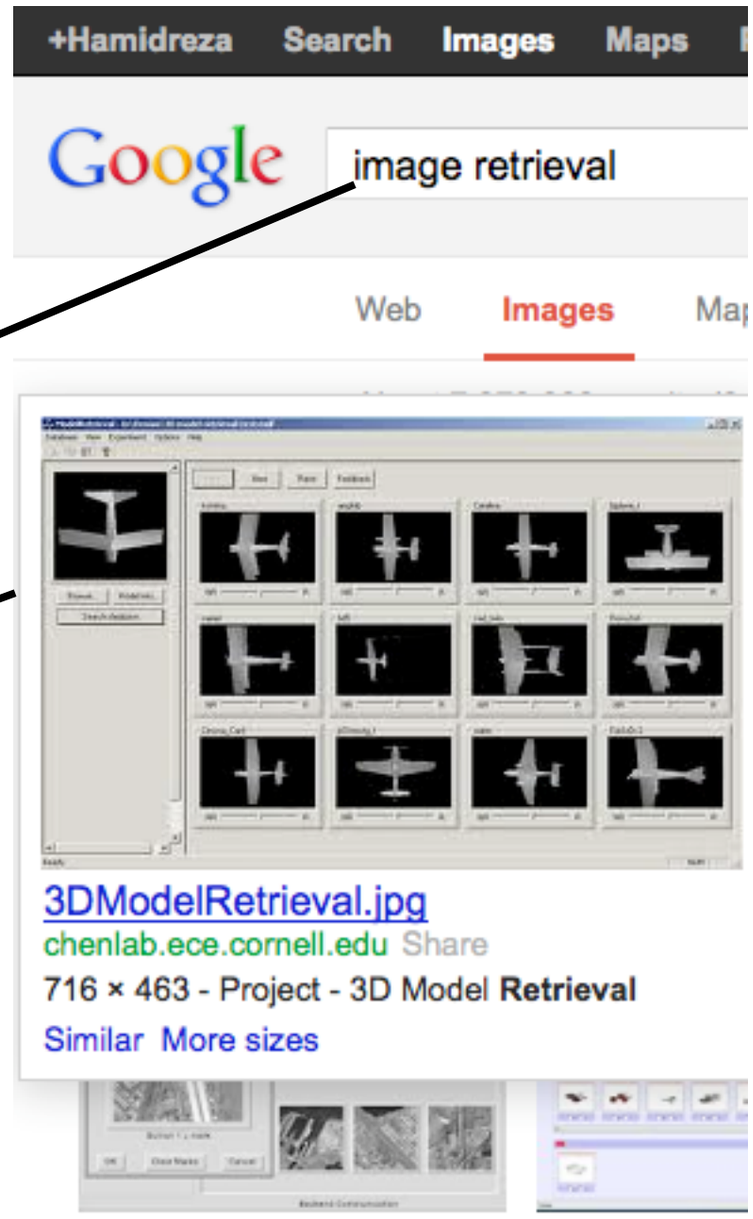
2-Use of Image Retrieval in Medical apps.

3- Techniques used

Image Retrieval

- Image Search:

IR Query: text



CBIR Query: image

- Content-based IR:

Some CBIR Systems

- IBM QBIC (Query by Image Content)
- Virage: used by CNN
- Candid
- Photobook
- Netra
- BlobWorld
- PicHunter
- GNU Image Finder Tool
- Viper, WIPE, Compass

Components of CBIR

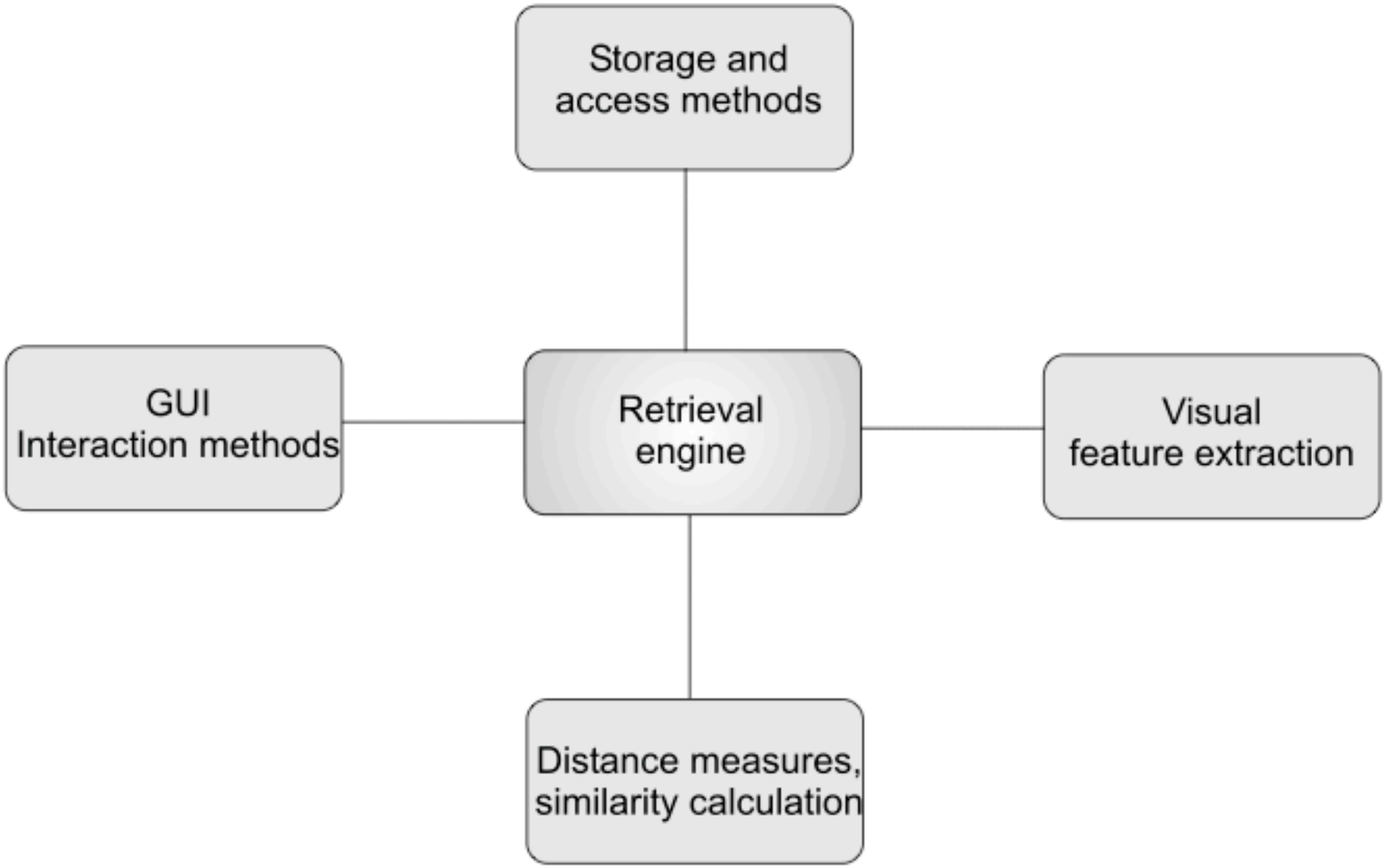
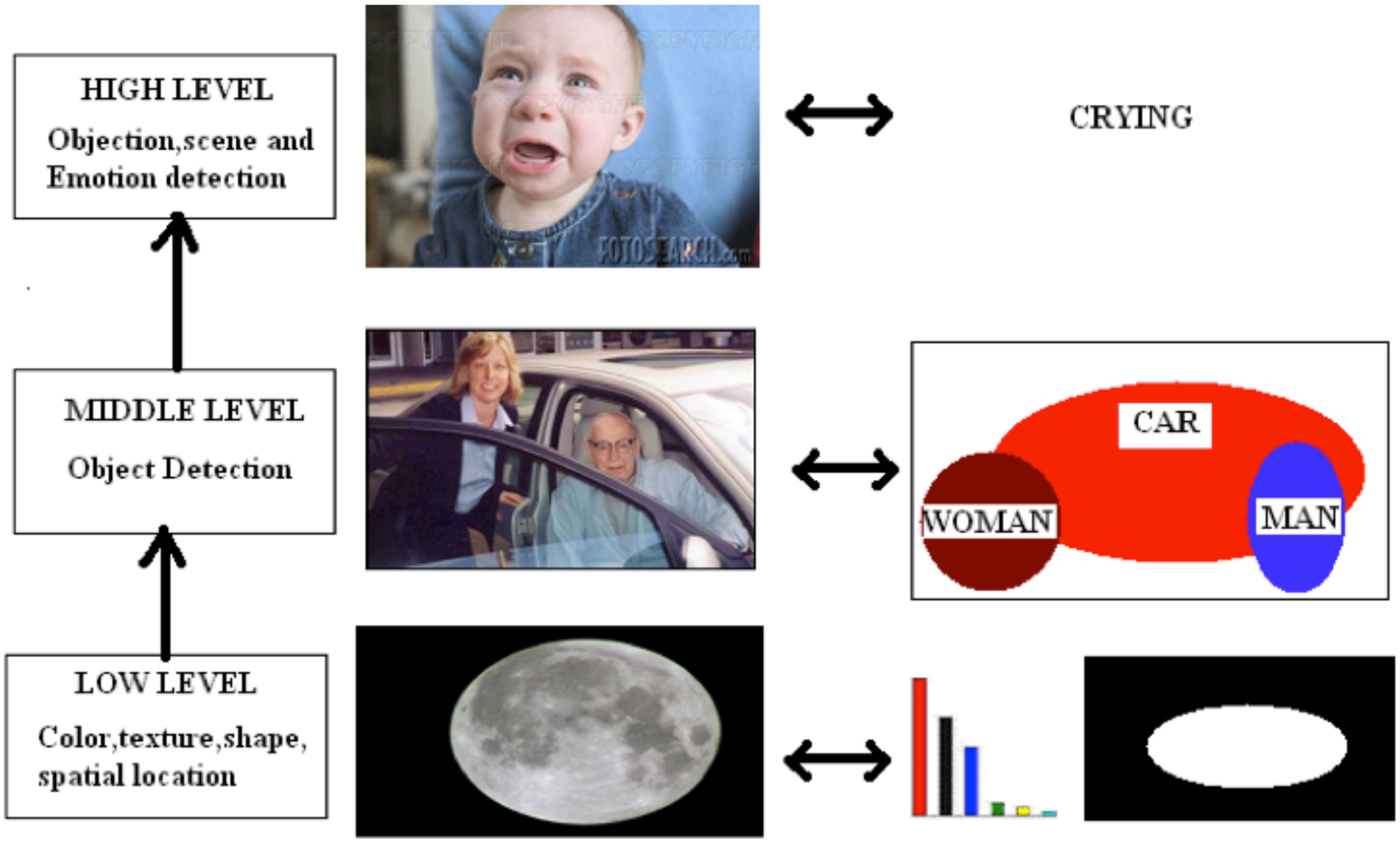


Figure 1: The principal components of all content–based image retrieval systems.

Features

- Primitive: Pixel Color, Texture, Shape
- Logical: Identity of the shape shown
- Abstract: Significance of the scenes detected

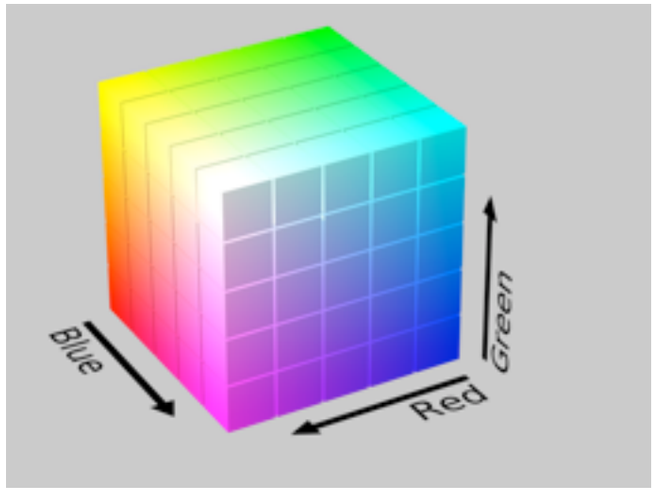
Features


HIGH LEVEL
Objection,scene and
Emotion detection
CRYING
MIDDLE LEVEL
Object Detection
CAR
WOMAN
MAN
LOW LEVEL
Color,texture,shape,
spatial location

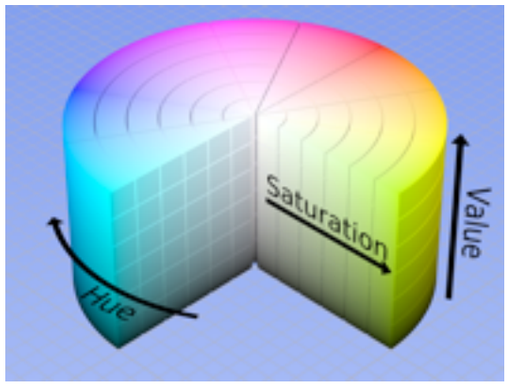
Features

- Currently most of the systems use Primitive Features (Color+Texture+Shape) unless manual annotation is coupled with features.
- **semantic gap**: The loss of info. from actual image to a representation by features.

Primitive Features

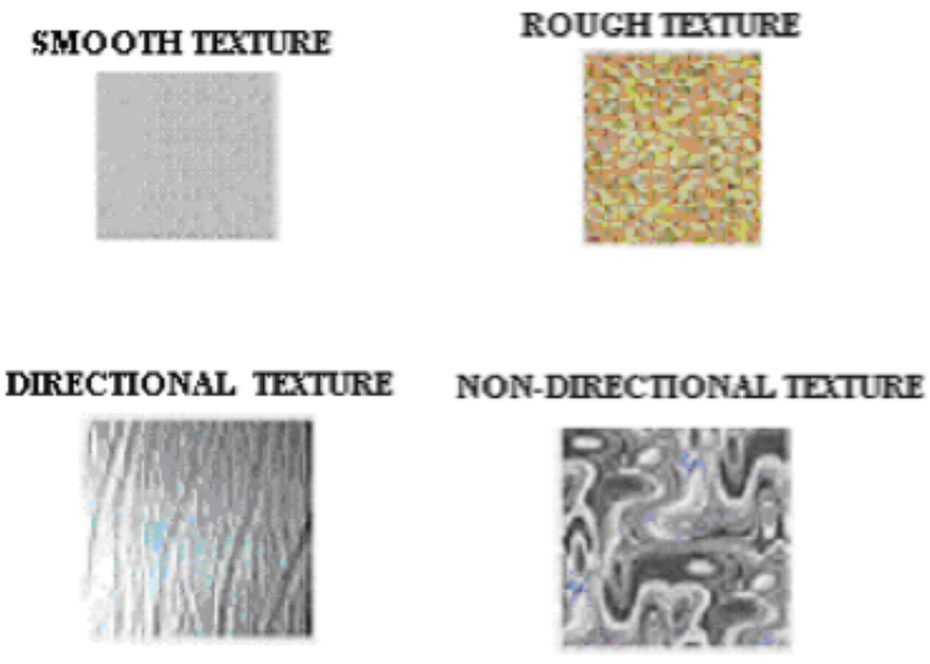


- Primitive Features:
  - **Pixel Color:**
    - RGB (Red, Green, Blue) for storing, but does not represent human perception, so not used for indexing.
    - HSV can be used. This space is much better with respect to human perception.



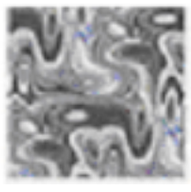
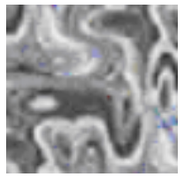
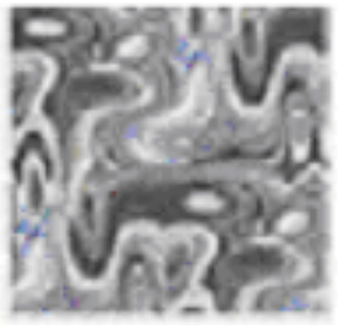
Primitive Features

- Primitive Features:
  - **Texture:**
  - imprecise definition.



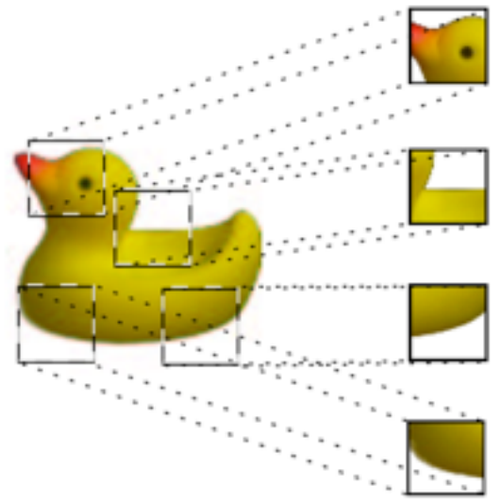
Primitive Features

- Use Gabor filters or Wavelets
- Texture try to capture characteristics of image parts with respect to changes in certain directions of scale of the changes.
- Can be rotation, shift, scale invariant.

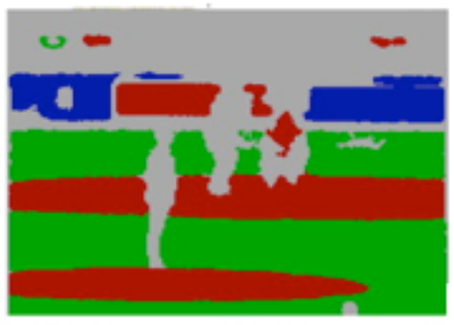
Primitive Features

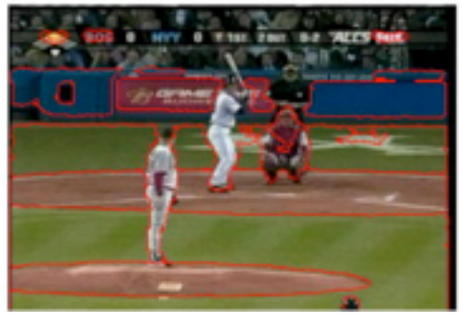
- **Local Features:**



- Fixed-size blocks
- Regions of Interest (by user)

- Segmentation

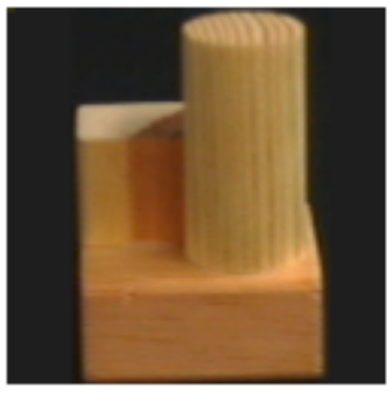


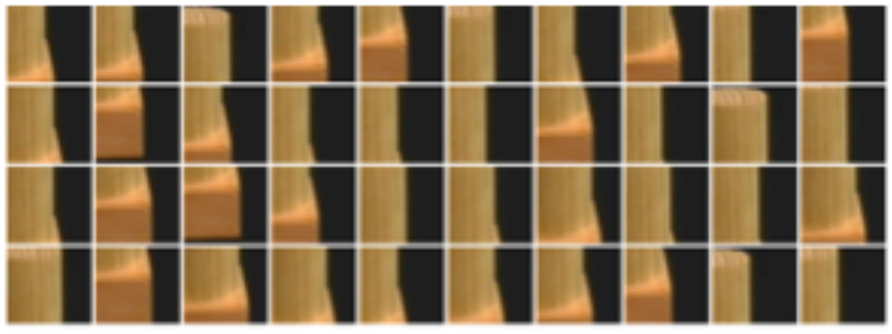


- **Segmentation and Shape:**
- unsolved problem
- After segmentation, the segmentation can be represented by shape features (rotation/ shift invariant)

Semantics

- Primitive features are very low-level.
- They do not correspond to objects in the image or semantics that the user is interested in.

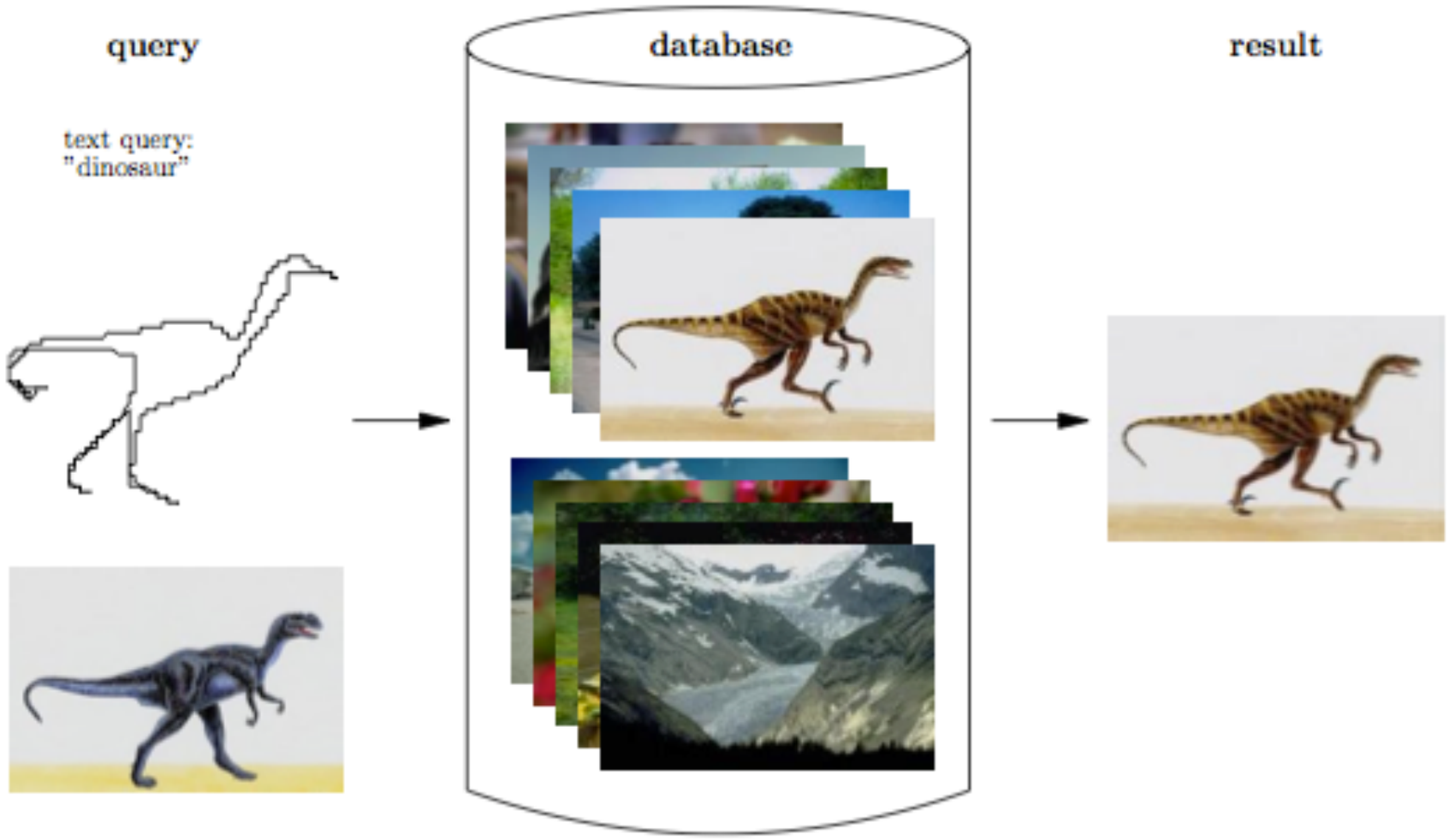




Comparison Methods

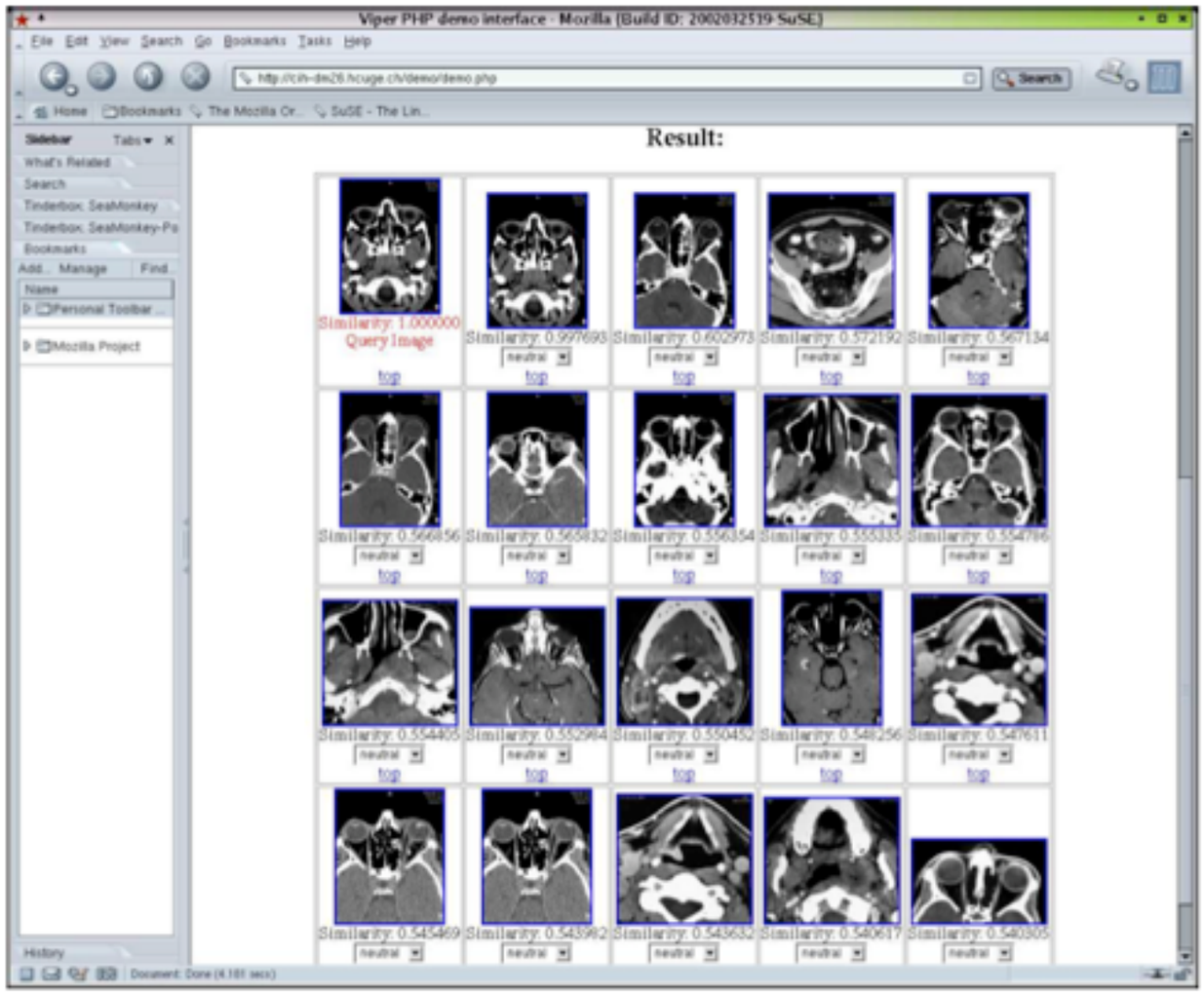
- **Vector-based**: Measure the similarity in feature space using Euclidian.
- **Probabilistic**: The probability that an image is relevant. Equivalent to above.
- **Text Retrieval based**: Apply text retrieval algorithms to visual features. features=words.

Overview



Medical

- CBIR in Medical application



Medical

- Common: Query by patient name, ID for images
- The author argues that it is beneficial to find other images of the same analytical region of the same disease.
- DICOM standard requires medical images to save these info. but it has some 16% error.

Medical

- Clinical decision support techniques like case-based reasoning or evidence-based medicine can benefit from images.
- Image-based reasoning? for diagnosis
- in addition to diagnosis, beneficial for research

Medical

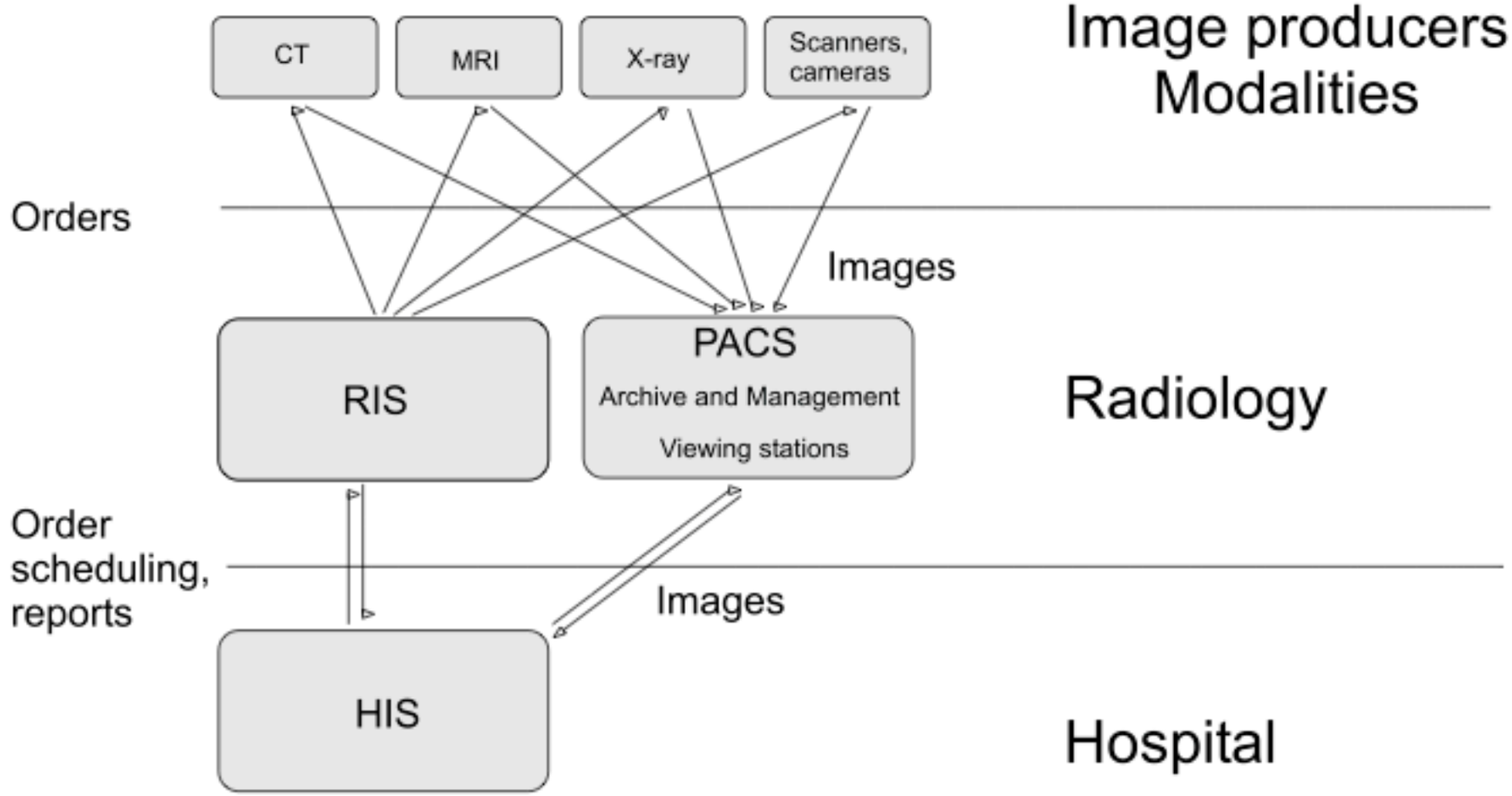


Figure 3: The basic position of a PACS within the information system environment in a hospital.

Medical

- Where CBIR is used?
  - radiology (mammography is the most application, reducing false positives)
  - dermatology
  - pathology

Medical

Images used	Names of the systems
HRCTs of the lung	*ASSERT*
Functional PET	*FICBDS*
Spine X–rays	*CBIR2, MIRS*
Pathologic images	*IDEM, I-Browse, PathFinder, PathMaster*
CTs of the head	*MIMS*
Mammographies	*APKS*
Images from biology	*BioImage, BIRN*
Dermatology	*MELDOQ, MEDS*
Breast cancer biopsies	*BASS*
Varied images	*I^2C, IRMA, KMed, COBRA, MedGIFT, ImageEngine*

Table 1: Various image types and the systems that are using these images.

Techniques

- Query Formation:
- TEXT: If the text is attached, use that as a starting query to find some initial images.
- HUMAN SKETCH: time-consuming, need skill
- IMAGE

Techniques

- Features:
  - Text: can we use that in a Content-based system?
  - Visual:
    - mostly grey level features
    - color less important
    - shape and texture more important

Techniques

- Methods:
  - Vector-space: Euclidian, city block, ...
  - Probabilistic: ANN, HMM, Bayesian Net.,...
- Compression:
  - PCA, MDL, ICA, ...

Evaluation of CBIR

- Evaluation

$$sensitivity = \frac{pos.\ items\ classified\ as\ pos.}{all\ positive\ items} \quad (1)$$

$$specificity = \frac{neg.\ items\ classified\ as\ neg.}{all\ negative\ items} \quad (2)$$

$$accuracy = \frac{items\ classified\ correctly}{all\ items\ classified}$$

$$precision = \frac{no.\ relevant\ items\ retrieved}{no.\ items\ retrieved} \quad (4)$$

$$recall = \frac{no.\ relevant\ items\ retrieved}{no.\ relevant\ items} \quad (5)$$

Paper 2

- Paper 2

- Zhou, X., Stern, R., & Müller, H. (2012). **Case-based fracture image retrieval**. *International journal of computer assisted radiology and surgery*, *7*(3), 401-411.

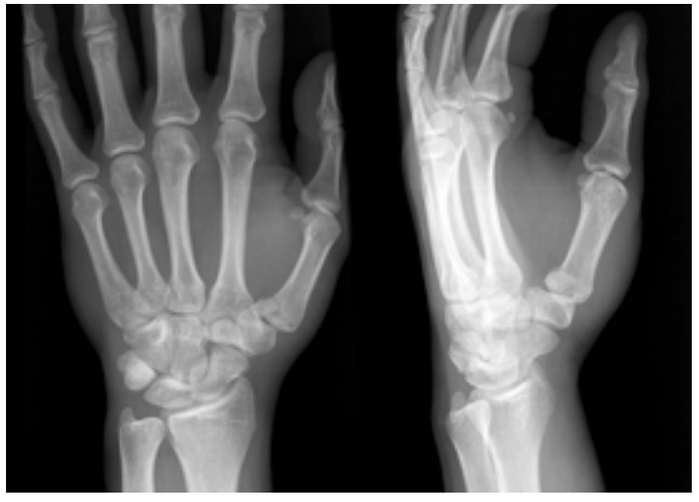
Case-based fracture IR

- CBIR can be used to assist surgeons by giving them past cases.

- Baseline: GNU Image Finding Tool
- Feature: Scale Invariant Feat. Trans.
- Evaluation: Mean Average Precision

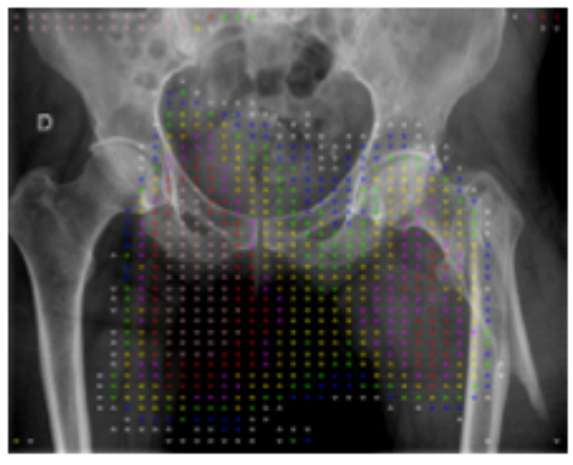
database

- A set of fracture cases pulled from Geneva Hospital database.
- 23,970 images, 2,690 cases, 43 fracture classes



Retrieval Techniques

- Features:
- 40*40 pixel grid sampling
- The standard SIFT (Scale Invariant Feat. Trans.) detector

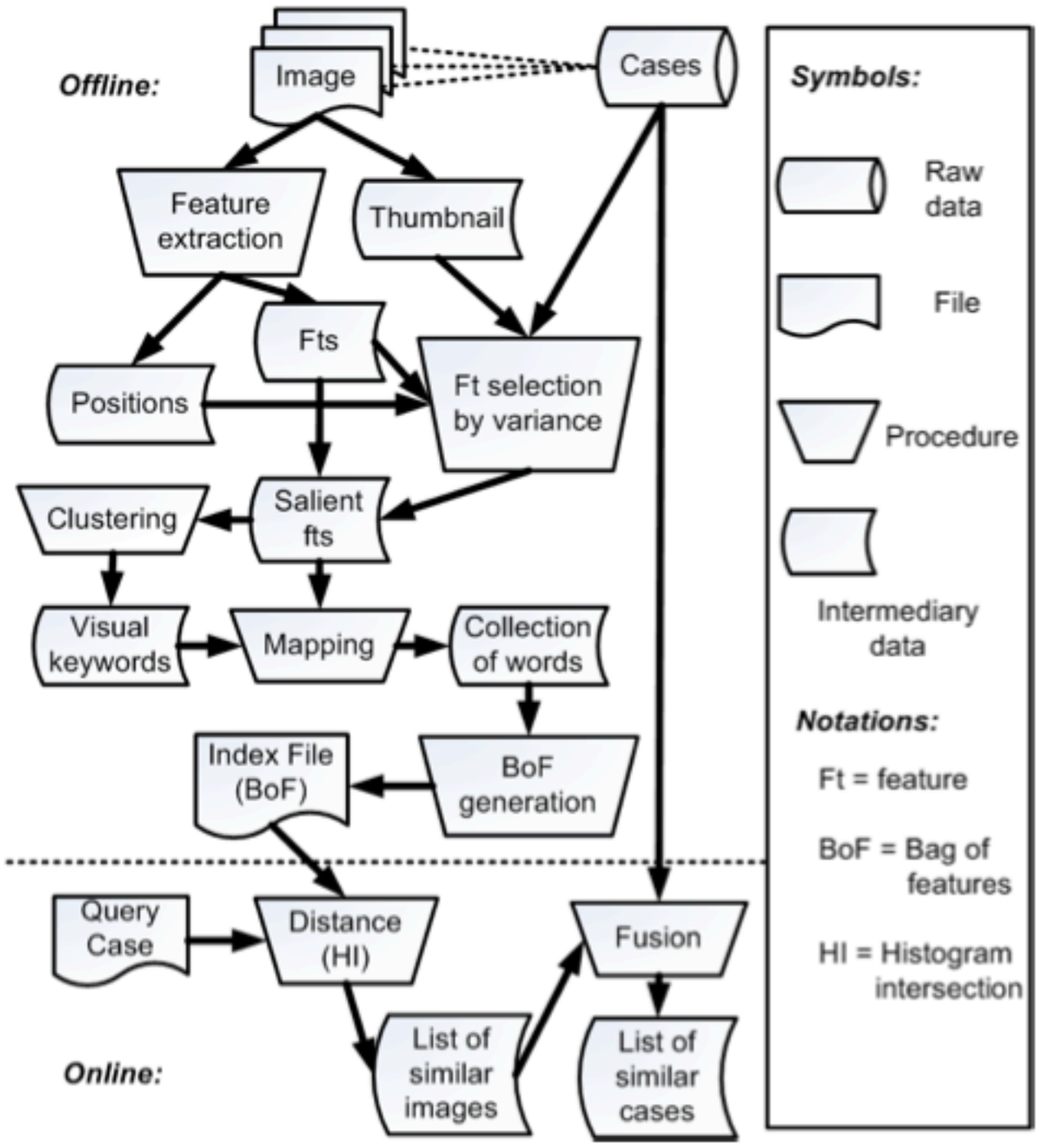


Method

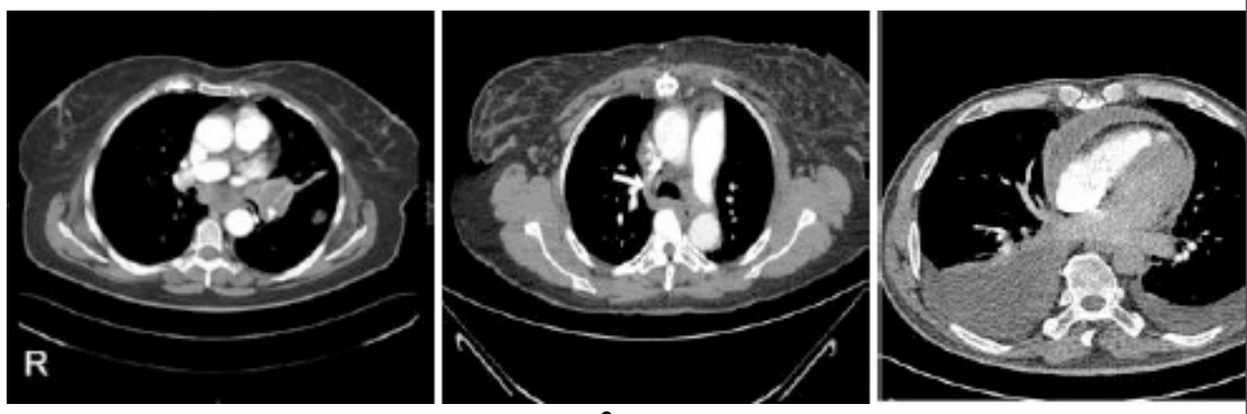
- Bag of key Words approach
- features are continuos -> quantize them using clustering -> visual codebook
- The optimal number of clusters kd = 1000

Method

- Each image is represented as a BoW or histogram of visual codebook
- distance between two images: calculated using HI (histogram Intersection)


Offline:
Image
Cases
Feature extraction
Thumbnail
Fts
Positions
Ft selection by variance
Clustering
Salient fts
Visual keywords
Mapping
Collection of words
Index File (BoF)
BoF generation
Query Case
Distance (HI)
Fusion
List of similar images
List of similar cases
Online:
Symbols:
Raw data
File
Procedure
Intermediary data
Notations:
Ft = feature
BoF = Bag of features
HI = Histogram intersection

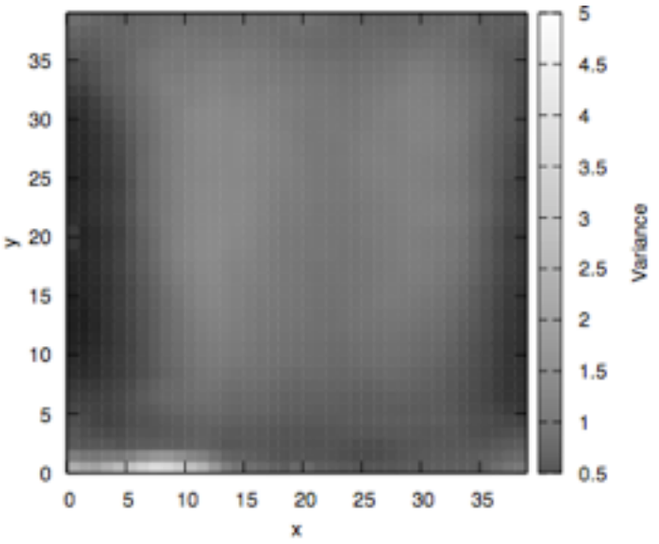
Fusion

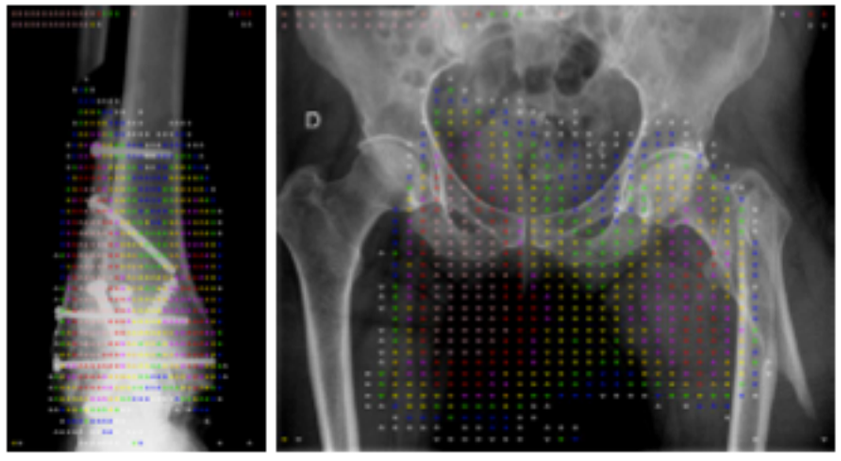


- Fusion:
  - We have multiple images as queries
  - We want to merge them
  - Two level merge
    - first keep best of each case
    - then combine DcomMNZ??

Feature Selection

- Feature Selection
- pixels with higher variance





Discussion

- Points about Paper 2:
  - high-dimensional sampling, not good for modeling purposes
  - even this high dimensional sampling does not seem to represent fractures perfectly
  - Use of segmentation/shape features?

Paper 3

- Paper 3

- Mitchell, Margaret, et al. "Midge: Generating image descriptions from computer vision detections." *Proceedings of EACL*. 2012.

- Paper 3 includes a novel generation system that composes human-like descriptions of images from computer vision detections.
- Produces in present-tense, declarative sentences as a naive viewer with no prior knowledge

The bus by the road with a clear blue sky

stuff:	*sky*	.999	
	id:	1	
	atts:	clear:0.432,	blue:0.945
		grey:0.853,	white:0.501 ...
	b. box:	(1,1 440,141)	
stuff:	*road*	.908	
	id:	2	
	atts:	wooden:0.722	clear:0.020 ...
	b. box:	(1,236 188,94)	
object:	*bus*	.307	
	id:	3	
	atts:	black:0.872,	red:0.244 ...
	b. box:	(38,38 366,293)	
preps:	id 1, id 2: by	id 1, id 3: by	id 2, id 3: below

Figure 2: Example computer vision output and natural language generation input. Values correspond to scores from the vision detections.

output of computer vision

- A_i is the set of object/stuff detections with bounding boxes and associated "attribute" detections within those bounding boxes.
- B_i is the set of action or pose detections associated to each $a_i \in A_i$.
- C_i is the set of spatial relationships that hold between the bounding boxes of each pair $a_i, a_j \in A_i$.

image description

- A_d is the set of nouns in the description with associated modifiers.
- B_d is the set of verbs associated to each $a_d \in A_d$.
- C_d is the set of prepositions that hold between each pair of $a_d, a_e \in A_d$.

- Problem:
  - How to filter out wrong detections?
  - Order objects so they are in a natural way
  - Connect these ordered objects in syntactically/semantically well-formed trees
- Use A_d as description anchors.

Learning from Text

- 700,000 Flicker images + associated descriptions
- normalize text
- run Berkeley parser

- relationship between two object nouns:

1. prepositional (a boy *on* the table)
2. verbal (a boy *cleans* the table)
3. verb with preposition (a boy *sits on* the table)

- The process of generation is approached as a problem of generating a semantically and syntactically well-formed tree based on object nouns.

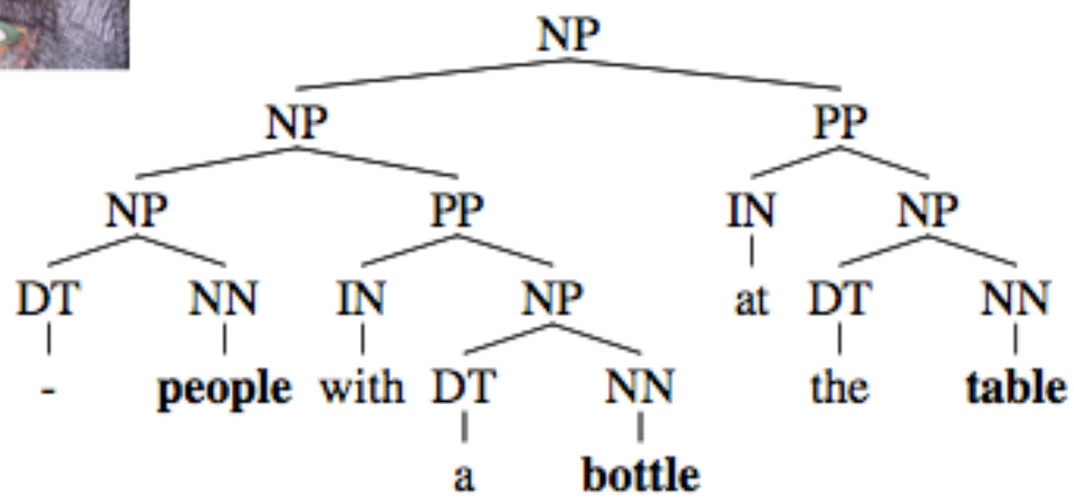


Figure 6: Tree generated from tree growth process.

- Group names: no more than 3 names
- Order names:

Unordered		Ordered
bottle, *table*, *person*	→	*person*, *bottle*, *table*
road, *sky*, *cow*	→	*cow*, *road*, *sky*

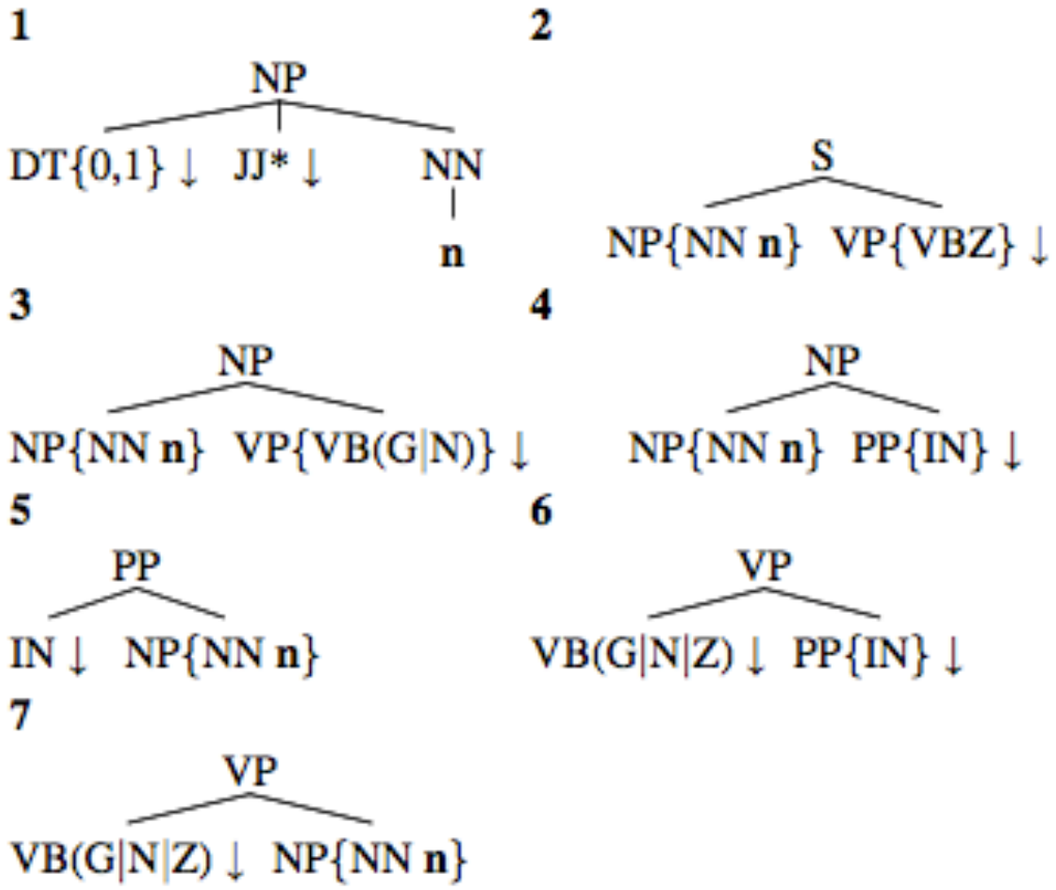
Figure 8: Example nominal orderings.

- Group adjectives in broader class:

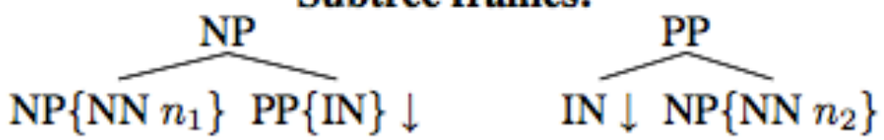
COLOR	purple blue green red white ...
MATERIAL	plastic wooden silver ...
SURFACE	furry fluffy hard soft ...
QUALITY	shiny rust dirty broken ...

Table 2: Example attribute classes and values.

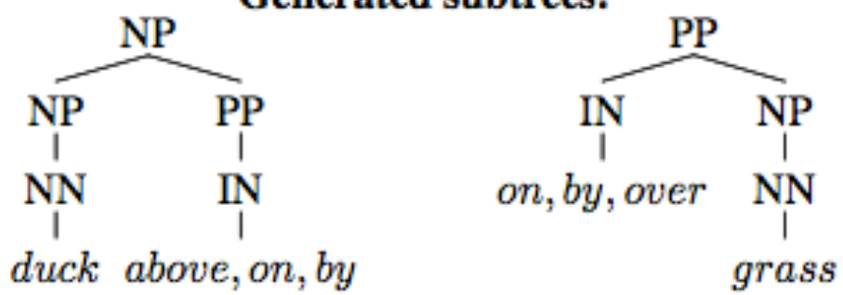
- Init. subtrees:



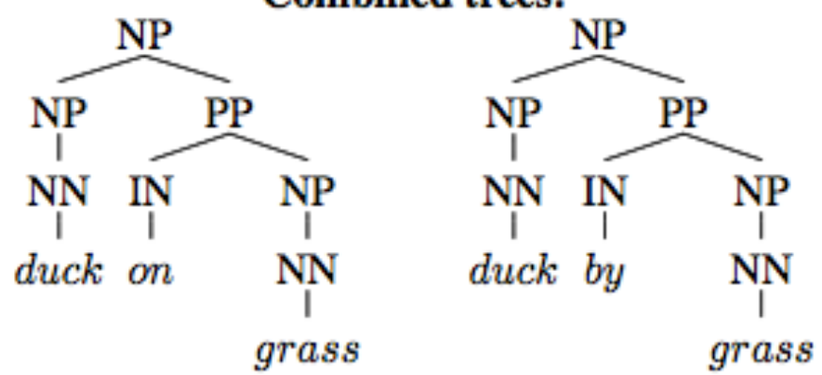
Subtree frames:



Generated subtrees:



Combined trees:



- Evaluation

	Grammaticality	Main Aspects	Correctness	Order	Humanlikeness
Human	4 (3.77, 1.19)	4 (4.09, 0.97)	4 (3.81, 1.11)	4 (3.88, 1.05)	4 (3.88, 0.96)
Midge	3 (2.95, 1.42)	3 (2.86, 1.35)	3 (2.95, 1.34)	3 (2.92, 1.25)	3 (3.16, 1.17)
Kulkarni et al. 2011	3 (2.83, 1.37)	3 (2.84, 1.33)	3 (2.76, 1.34)	3 (2.78, 1.23)	3 (3.13, 1.23)
Yang et al. 2011	3 (2.95, 1.49)	2 (2.31, 1.30)	2 (2.46, 1.36)	2 (2.53, 1.26)	3 (2.97, 1.23)

A person sitting on a sofa

Cows grazing

Airplanes flying

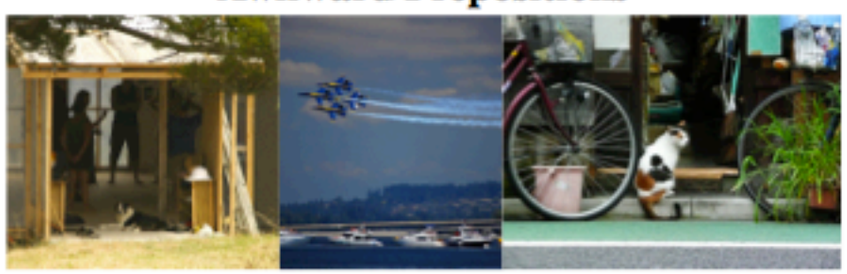
A person walking a dog

- Hallucinating: Creating likely actions. Straightforward to do, but can often be wrong.

a cow with green grass | sheep with a gray sky by the road | people with boats | a brown cow | people at a wooden table

Figure 4: Example generated outputs.

Awkward Prepositions



a person *on* the dog | boats *under* the sky | a black bicycle *at* a green potted plant

Incorrect Detections

the sky | a *yellow* bus with people | cows by black *sheep* by the road

Figure 5: Example generated outputs: Not quite right

Kulkarni et al.: This is a picture of three persons, one bottle and one diningtable. The first rusty person is beside the second person. The rusty bottle is near the first rusty person, and within the colorful diningtable. The second person is by the third rusty person. The colorful diningtable is near the first rusty person, and near the second person, and near the third rusty person.

Yang et al.: Three people are showing the bottle on the street

Midge: people with a bottle at the table

Kulkarni et al.: This is a picture of two pottedplants, one dog and one person. The black dog is by the black person, and near the second feathered pottedplant.

Yang et al.: The person is sitting in the chair in the room

Midge: a person in black with a black dog by potted plants

- Questions?