ONTOLOGY-BASED ANNOTATION AND RETRIEVAL SYSTEM FOR DIGITAL MAMMOGRAPHY IMAGES

Adil ALPKOCAK

Dokuz Eylül University, Department of Computer Engineering, Turkey

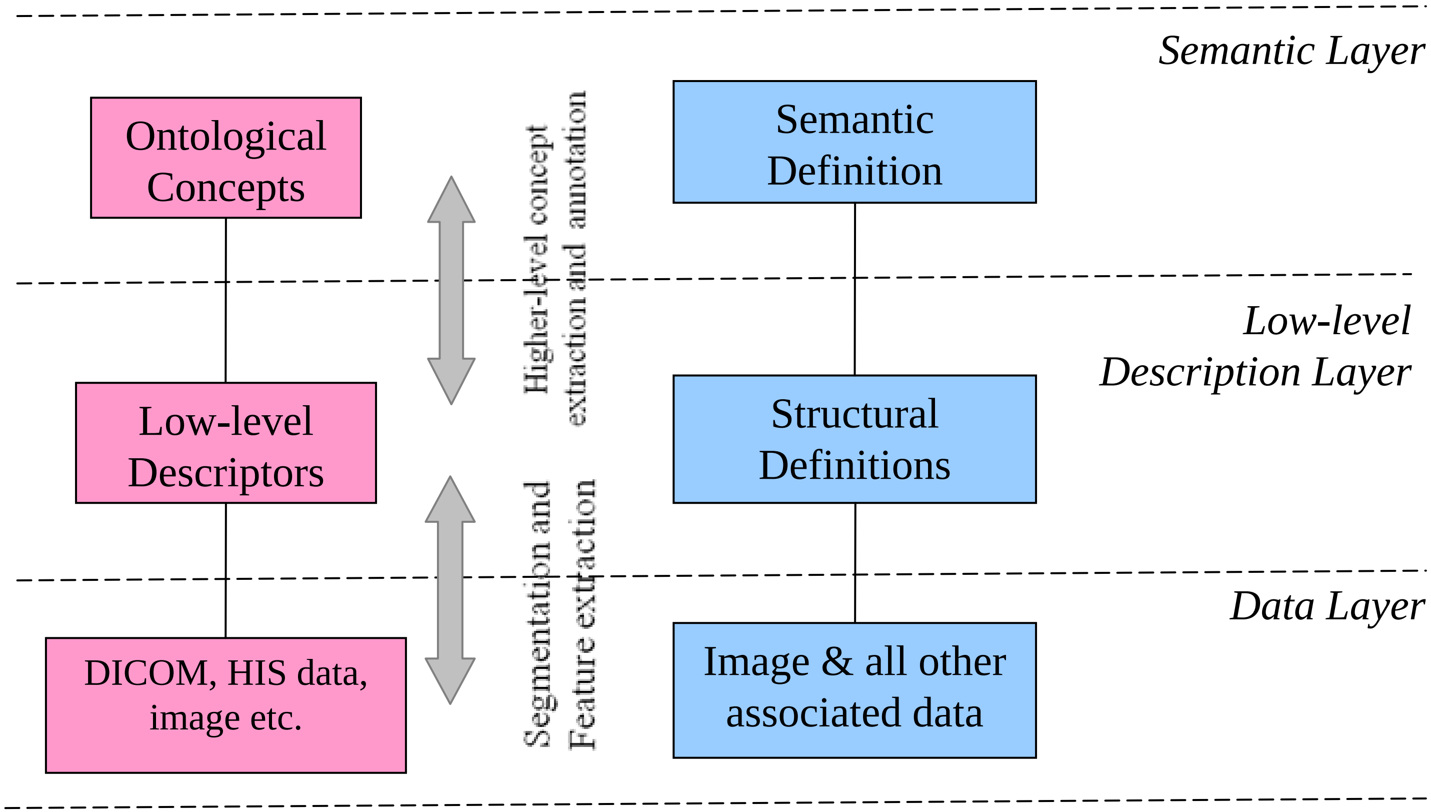
Outline

- About Project
- System Arcitecture
- ImageCLEF MM
  - Document & Query Expansion
  - Re-ranking
- ImageCLEF Med
  - Integrated Retrieval Model: text & image
- Ontology development for Mammography
- Annotated Mammography *test-bed* development
- Conclusion

About Project

- *Ontology-based Annotation* and Retrieval system for Digital Mammography.
- Three years long project
- Supported by Turkish National Science Foundation
- Involves researchers from computer engineering, electronic engineering and medical school.
- Aims to provide tools;
  - for evidence-based medicine to physicians,
  - Bridging semantic gap.
- Already completed one and a half year.

System Architecture


Semantic Layer
Ontological Concepts
Semantic Definition
Higher-level concept extraction and annotation
Low-level Description Layer
Low-level Descriptors
Structural Definitions
Segmentation and Feature extraction
Data Layer
DICOM, HIS data, image etc.
Image & all other associated data

Data Layer

- Representation of data in well known image format such as DICOM, jpeg, tiff etc.
- DICOM images are not optimized for content representation and extraction and, stored digitally together with external attributes such as date of acquisition, category, anatomical part, patient id and name etc.
- Interpretational attributes or annotations in DICOM describing image content and disease code are not considered in this layered.

Low-level Desciptor Level

- The middle level,
- Also defined as symbolic abstraction level,
- Contains description of multimedia content in forms of low level features like texture, color and shape using a well defined data format.

Semantic Layer

- The highest level,
- Also named as conceptual abstraction layer,
- Provides semantic interpretation of lower levels and aims the mapping between structural information resources and information representation of the related fields.

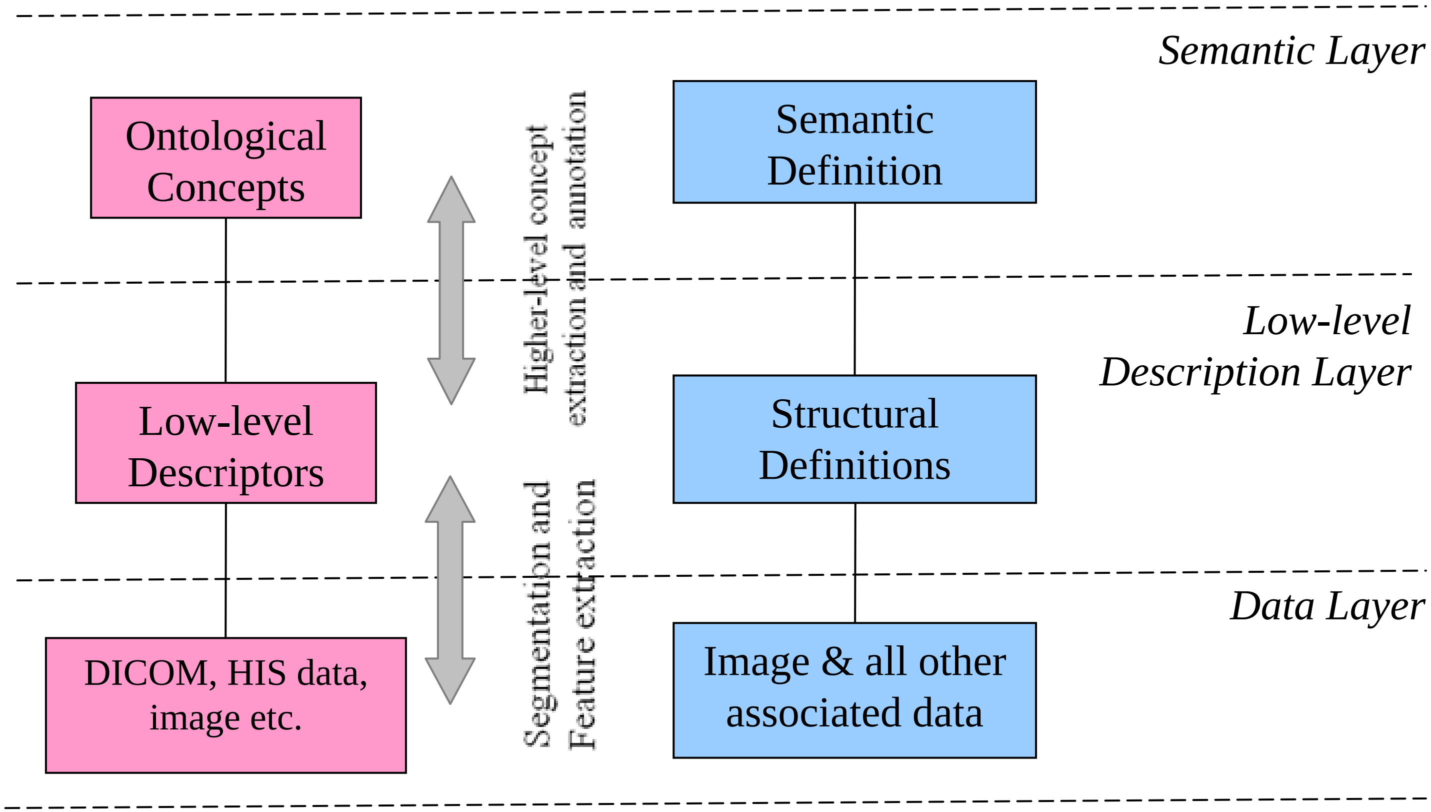
Transition between layers

- Transition from lowest layer to mid-layer
  - Relatively easy.
  - Processing whole image: it is done in automatic manner, without human intervention.
  - Processing a part of image: it is considered requires user to select region of interest with proper interfaces.
  - Once region of interest is defined, extraction of low level descriptors is generally fast, automatic and systematic.

Transition between layers

- Transition from Mid-layer to top
- All descriptors in mid-layer are abstracts and, does not directly map to real world concepts which is known as “semantic gap” problem in literature.
- provide a solution to this problem, we propose the top layer which includes semantic interpretations of law level descriptors.

System Architecture


Semantic Layer
Ontological Concepts
Higher-level concept extraction and annotation
Semantic Definition
Low-level Description Layer
Low-level Descriptors
Structural Definitions
Segmentation and Feature extraction
Data Layer
DICOM, HIS data, image etc.
Image & all other associated data

ImageCLEF2009: WikipediaMM

- Document and Query Expansion
- Re-ranking

Baseline Retrieval

- Stop-words elimination.
- The phase is Lemmatization reducing an inflected spelling to its lexical root or lemma form.
- Performed document expansion,
- Pivoted Unique Normalization, which is a modified version of classical cosine normalization.

$$R_{base} = \frac{(R_{original} \times \mu) + (R_{expanded} \times \partial)}{2}$$

Document and Query Expansion

- The aim of expanding both documents and the queries is to push document and queries into each other.
- Expanding the queries and widening the search terms, increase the quality of ranking by bringing relevant documents not matching literally with the original user query.
- Expanding the poorly defined documents and adding new terms or term-phrases, results in higher ranking performance.

Term Phrase Selection

- If the two successive terms exist in WordNet as a noun-phrase, they are accepted as term-phrases, added to dictionary.
- In this work, 6,808 term-phrases are generated and added into dictionary for Wiki dataset.
- For example, “hunting”, “dog”
- If this phrase exists in WordNet, the document or query is expanded with the term “hunting-dog”.
- And finally the term phrase is added to the term phrase dictionary.

Document and Query Expansion

- Consider as an example document that includes the term “*sea lavender*” *and, a query* “*blue flower*”,
- Without expansion, they are not matching literally and they seem irrelevant.
- Expand document “*sea lavender*”, *add new* terms “*blue flower*”.
- So, expanding both query and the document results
  - same terms in both document and query,
  - an increase in ranking score.

Reranking

- Reranking is a methodical technique to reorder the initial retrieved documents for better results by increasing the precision.
- relevant documents that have low ranking weights are reweighted and reordered in a retrieved resultset.

Reranking

- We propose a new reranking approach in two phases.
- Base retrieval results are generated, the result sets of each query and the base ranking scores () are saved for the reranking phases.
- The first phase comprises reranking and reordering with the Boolean retrieval approach.
- Boolean retrieval is performed first.
- Second step is about reranking with the clustering based on C3M clustering algorithm.

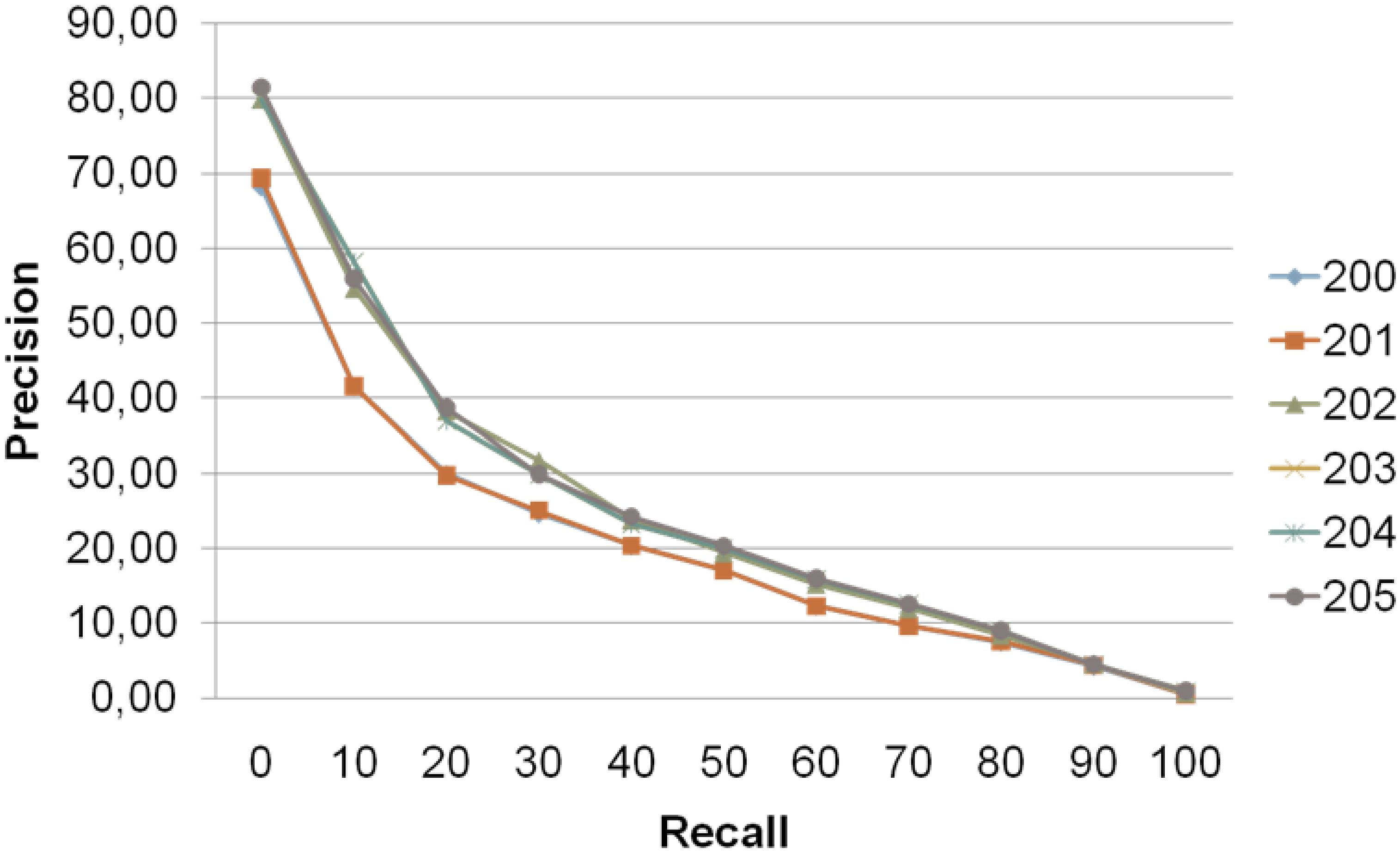
Reranking with Boolean Retrieval

- **Reranking with Boolean Retrieval**

Our runs in WikipediaMM Task

ID	MAP	P@5	P@10	R-Precision	Retrieved	Rel.Ret.	Relevant
200	0.1861	0.3244	0.2956	0.2133	41242	1283	1622
201	0.1865	0.3422	0.2978	0.2146	41242	1283	1622
202	0.2358	0.4844	0.3933	0.2708	43052	1352	1622
203	0.2375	0.4933	0.4000	0.2692	43053	1351	1622
204	**0.2375**	**0.4933**	**0.4000**	0.2692	39257	1351	1622
205	**0.2397**	**0.5156**	**0.4000**	0.2683	43052	1351	1622

Precision-Recall Graph of our runs



ImageCLEFMed

- Integrated Retrieval Model

Retrieval Model

- A typical Vector Space Model of Salton

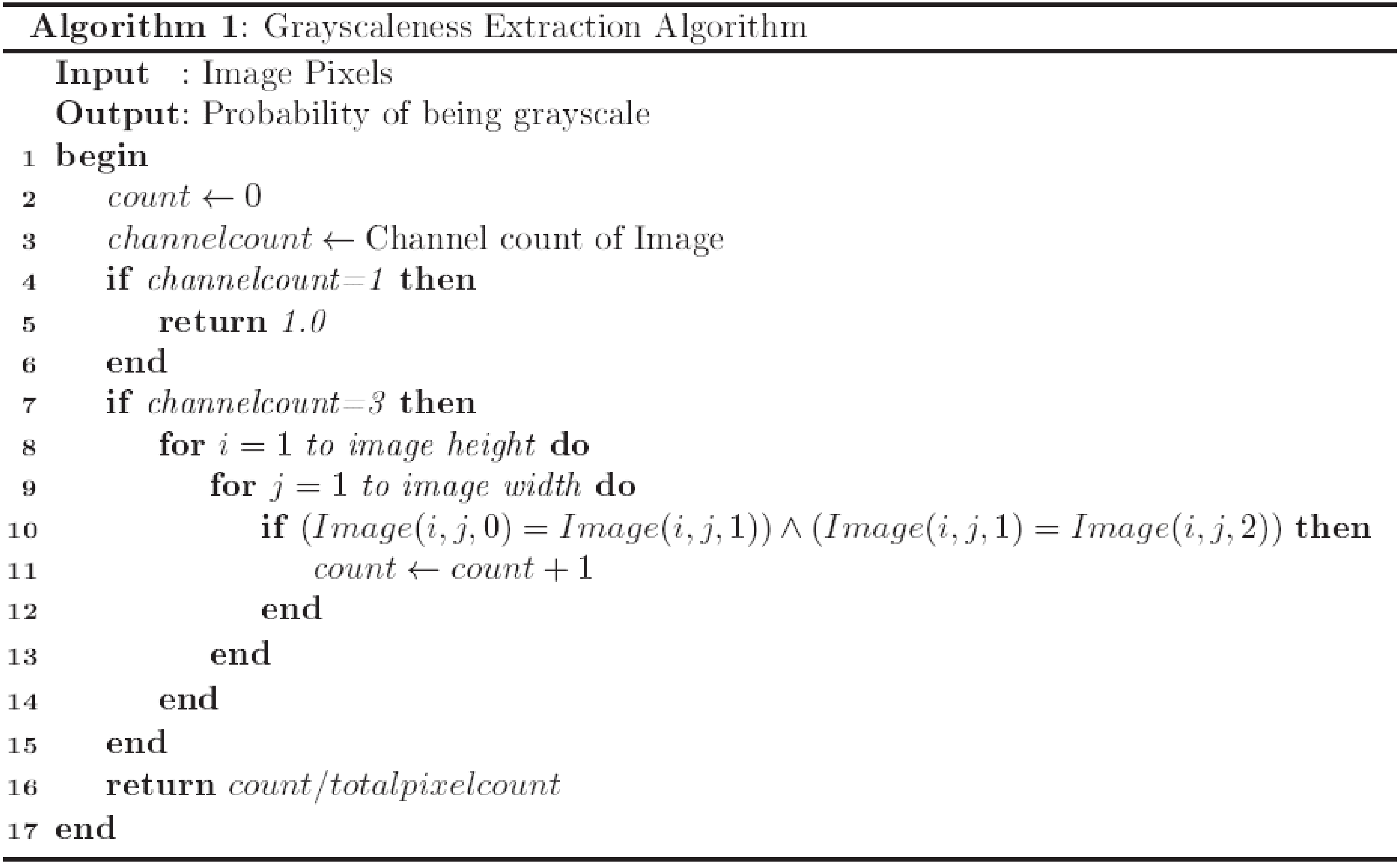
$$D = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{22} \\ \cdots & \cdots & \cdots & \cdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix}$$

Integrated Retrieval Model

□ It integrates both text and image in one model.

$$D' = \left[\begin{array}{cccc|cccc} w_{11} & w_{12} & \cdots & w_{1n} & i_{11} & i_{12} & \cdots & i_{1k} \\ w_{21} & w_{22} & \cdots & w_{2n} & i_{21} & i_{22} & \cdots & i_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} & i_{m1} & i_{m2} & \cdots & i_{mk} \end{array}\right]$$

Image Features used

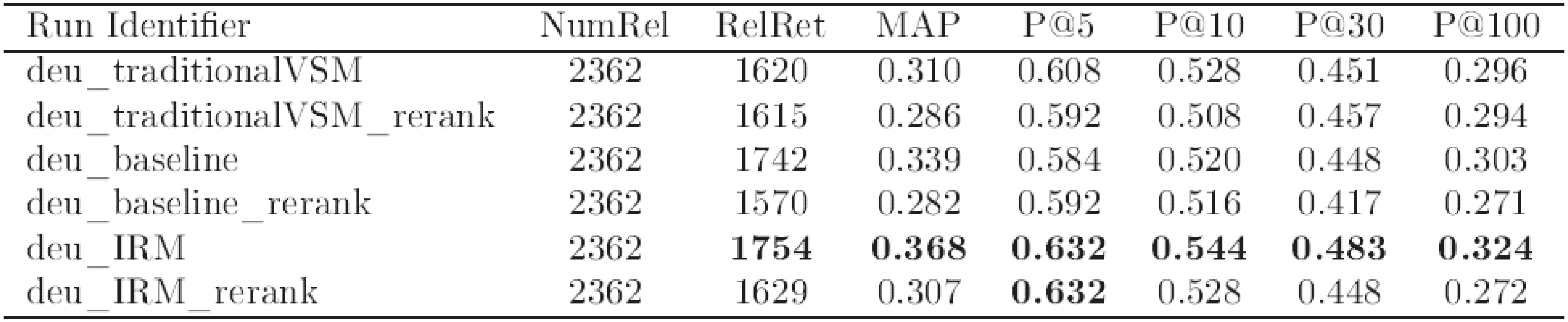


Algorithm 1: Grayscaleness Extraction Algorithm

```
Input  : Image Pixels
Output: Probability of being grayscale
begin
    $count \leftarrow 0$
    $channelcount \leftarrow$ Channel count of Image
    if $channelcount$=1 then
        return 1.0
    end
    if $channelcount$=3 then
        for $i = 1$ to image height do
            for $j = 1$ to image width do
                if $(Image(i,j,0) = Image(i,j,1)) \wedge (Image(i,j,1) = Image(i,j,2))$ then
                    $count \leftarrow count + 1$
                end
            end
        end
    end
    return $count/totalpixelcount$
end
```

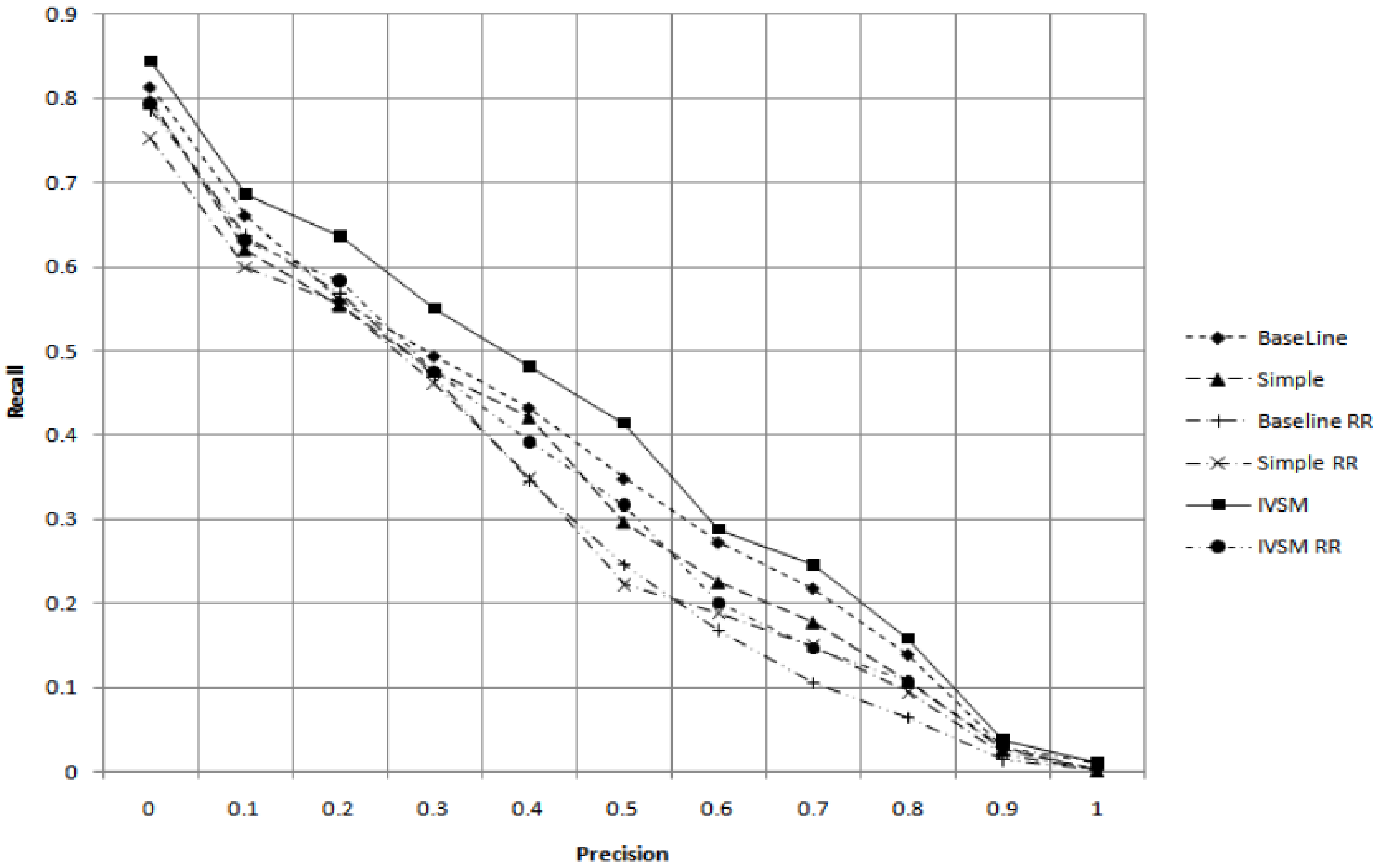
Integrated Retrieval Method

- Experimentation Results



Run Identifier	NumRel	RelRet	MAP	P@5	P@10	P@30	P@100
deu_traditionalVSM	2362	1620	0.310	0.608	0.528	0.451	0.296
deu_traditionalVSM_rerank	2362	1615	0.286	0.592	0.508	0.457	0.294
deu_baseline	2362	1742	0.339	0.584	0.520	0.448	0.303
deu_baseline_rerank	2362	1570	0.282	0.592	0.516	0.417	0.271
deu_IRM	2362	**1754**	**0.368**	**0.632**	**0.544**	**0.483**	**0.324**
deu_IRM_rerank	2362	1629	0.307	**0.632**	0.528	0.448	0.272

Precision Recall graph test runs in İmageCLEF 2009



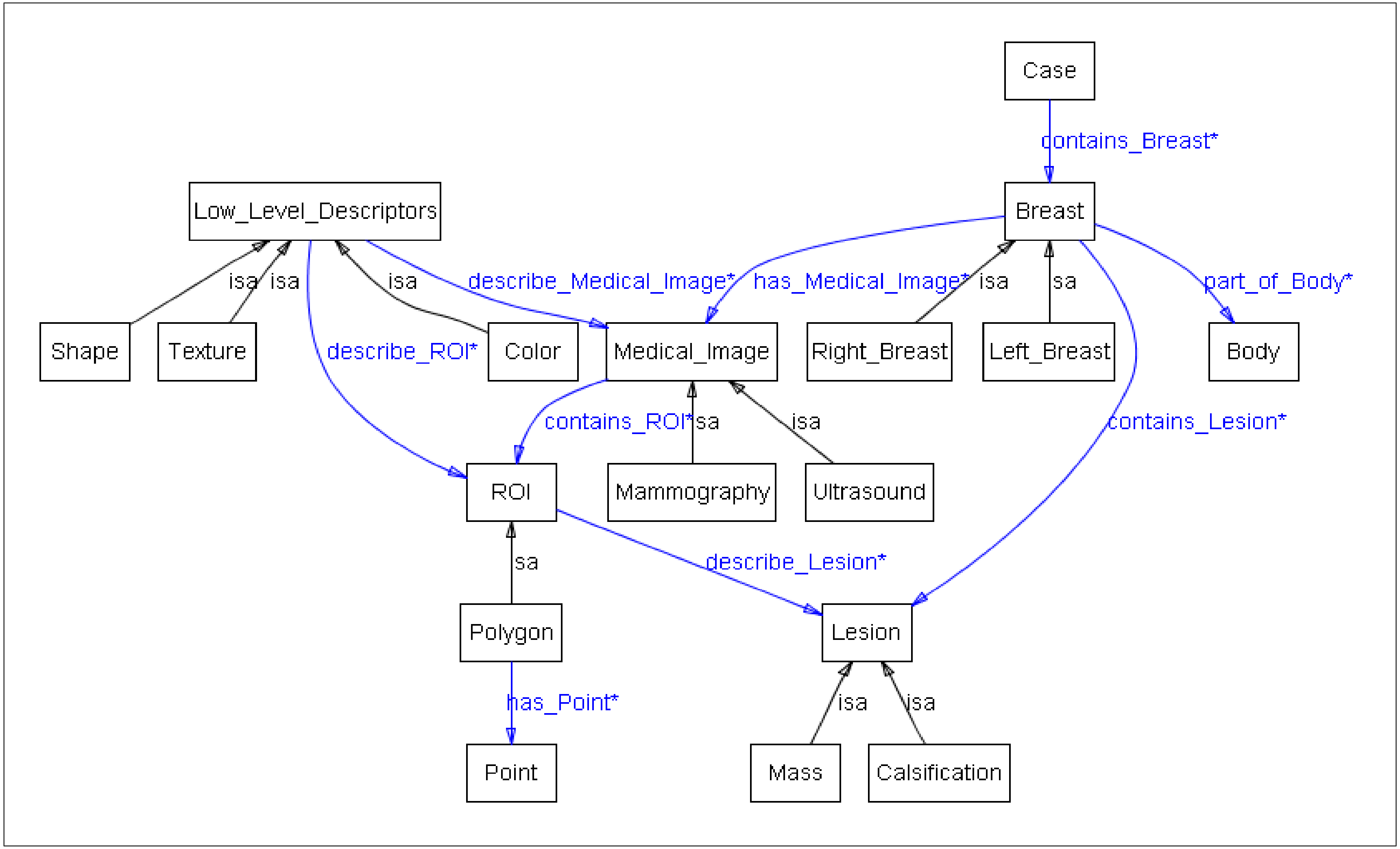
Ontology Construction

1. Ontology construction
2. Preperation of test-bed
3. Ontological Annotation

Ontology construction

- A Mammography ontology with domain expert
- An iterative method on construction.
- 48 top level classes
- In OWL-DL using protege
- Puplicly available
-

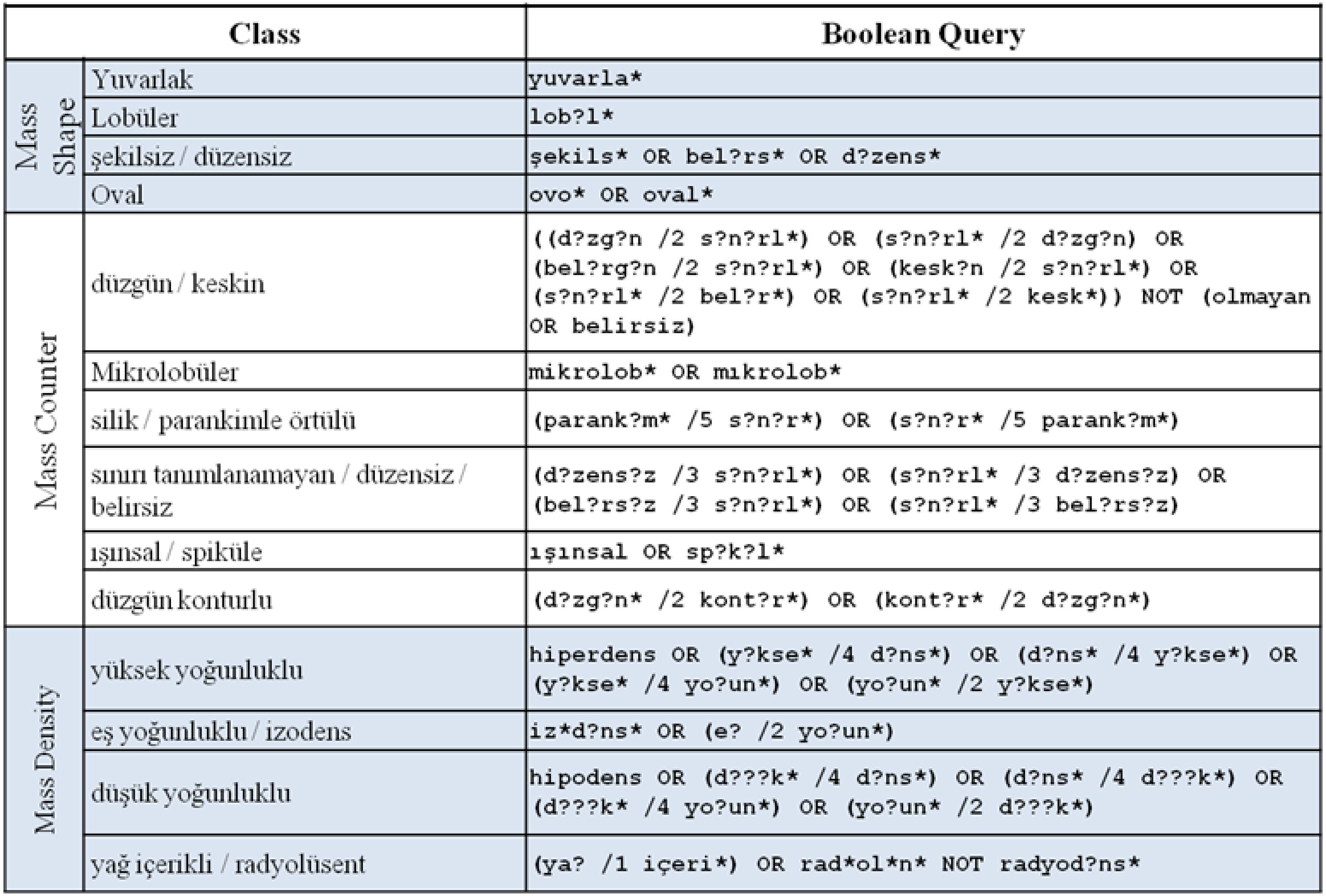
Part of Mammography Ontology



Preperation of Test-bed

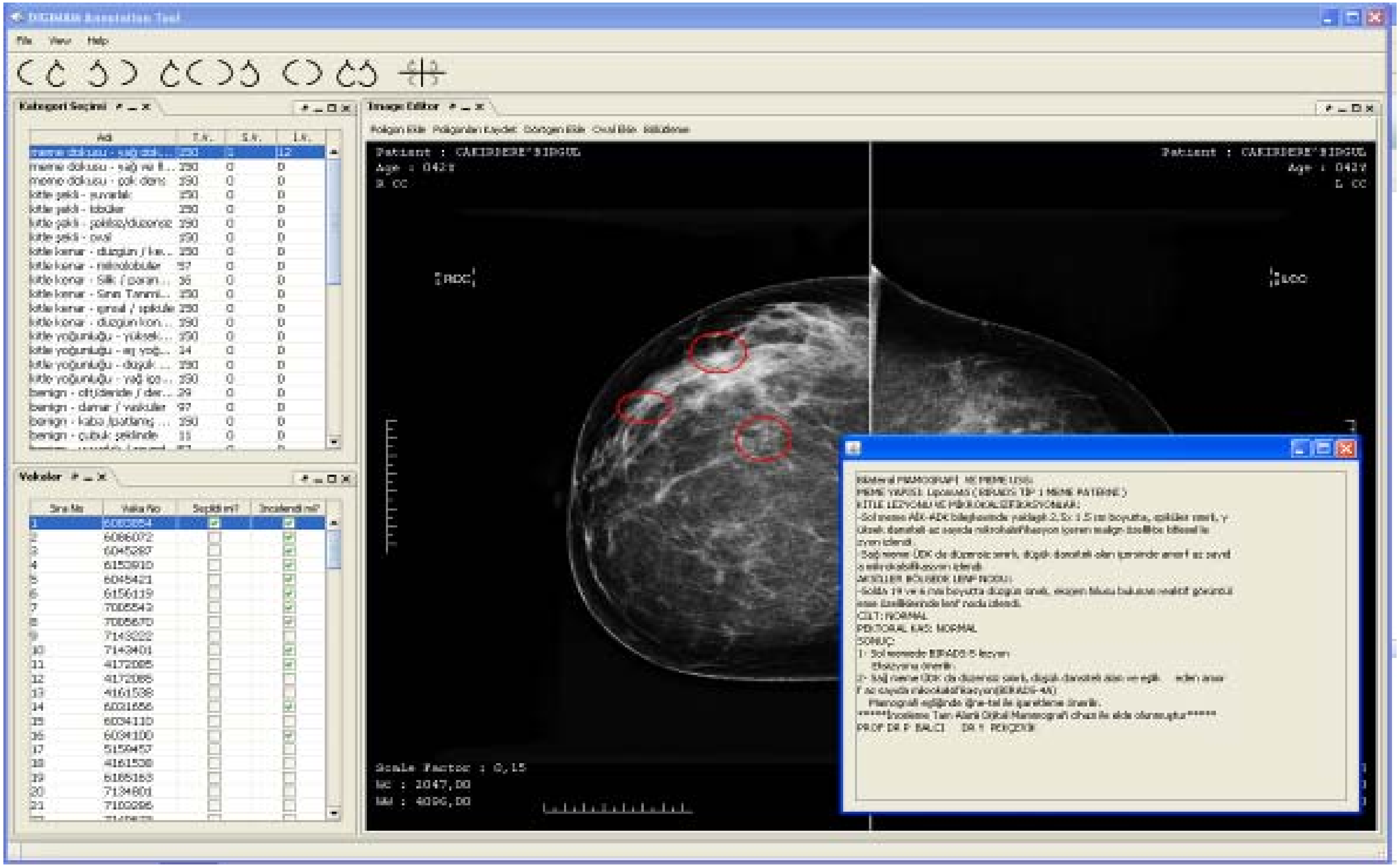
- University hospital PACS system has ~50,000 mammography cases.
- Max 150 candidate was selected automatically for each class.
- Some class has less example !!

Classes and Boolean Query



Class		Boolean Query
Mass Shape	Yuvarlak	yuvarla*
	Lobüler	lob?l*
	şekilsiz / düzensiz	şekils* OR bel?rs* OR d?zens*
	Oval	ovo* OR oval*
Mass Counter	düzgün / keskin	((d?zg?n /2 s?n?rl*) OR (s?n?rl* /2 d?zg?n) OR (bel?rg?n /2 s?n?rl*) OR (kesk?n /2 s?n?rl*) OR (s?n?rl* /2 bel?r*) OR (s?n?rl* /2 kesk*)) NOT (olmayan OR belirsiz)
	Mikrolobüler	mikrolob* OR mıkrolob*
	silik / parankimle örtülü	(parank?m* /5 s?n?r*) OR (s?n?r* /5 parank?m*)
	sınırı tanımlanamayan / düzensiz / belirsiz	(d?zens?z /3 s?n?rl*) OR (s?n?rl* /3 d?zens?z) OR (bel?rs?z /3 s?n?rl*) OR (s?n?rl* /3 bel?rs?z)
	ışınsal / spiküle	ışınsal OR sp?k?l*
	düzgün konturlu	(d?zg?n* /2 kont?r*) OR (kont?r* /2 d?zg?n*)
Mass Density	yüksek yoğunluklu	hiperdens OR (y?kse* /4 d?ns*) OR (d?ns* /4 y?kse*) OR (y?kse* /4 yo?un*) OR (yo?un* /2 y?kse*)
	eş yoğunluklu / izodens	iz*d?ns* OR (e? /2 yo?un*)
	düşük yoğunluklu	hipodens OR (d???k* /4 d?ns*) OR (d?ns* /4 d???k*) OR (d???k* /4 yo?un*) OR (yo?un* /2 d???k*)
	yağ içerikli / radyolüsent	(ya? /1 içeri*) OR rad*ol*n* NOT radyod?ns*

Ontology-based Annotation

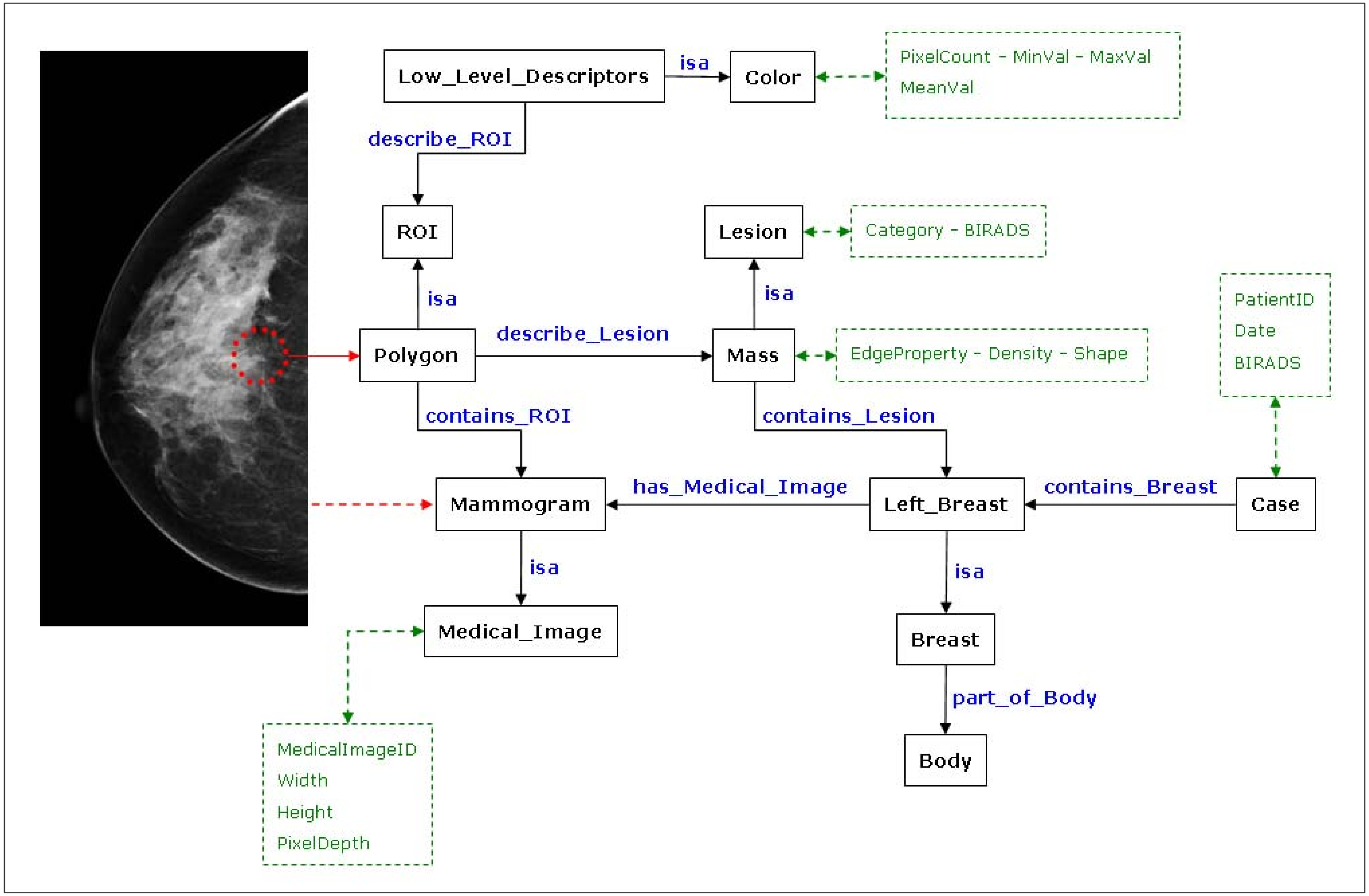


Adı	T.V.	S.V.	İ.V.
meme dokusu - yağ dok...	250	0	12
meme dokusu - yağ ve fi...	250	0	0
meme dokusu - çok dens	250	0	0
kitle şekli - yuvarlak	250	0	0
kitle şekli - lobüler	250	0	0
kitle şekli - şekilsiz/düzensiz	250	0	0
kitle şekli - oval	250	0	0
kitle kenar - düzgün / ke...	250	0	0
kitle kenar - mikrolobüler	57	0	0
kitle kenar - Silik (param...	36	0	0
kitle kenar - Sınır Tanımlı...	250	0	0
kitle kenar - ışınsal / spiküle	250	0	0
kitle kenar - düzgün kon...	250	0	0
kitle yoğunluğu - yüksek...	250	0	0
kitle yoğunluğu - eş yoğ...	14	0	0
kitle yoğunluğu - düşük ...	250	0	0
kitle yoğunluğu - yağ içe...	250	0	0
benign - cilt/deride / der...	29	0	0
benign - damar / vasküler	97	0	0
benign - kaba /patlamış ...	250	0	0
benign - çubuk şeklinde	11	0	0

Sıra No	Vaka No	Seçildi mi?	İncelendi mi?
1	6083054	☑	☑
2	6086072	☐	☑
3	6045287	☐	☑
4	6153910	☐	☑
5	6045421	☐	☑
6	6156119	☐	☑
7	7085543	☐	☑
8	7085670	☐	☑
9	7143222	☐	☐
10	7143401	☐	☑
11	4172085	☐	☑
12	4172086	☐	☐
13	4161538	☐	☐
14	6021656	☐	☑
15	6034110	☐	☐
16	6034100	☐	☑
17	5159457	☐	☐
18	4161536	☐	☐
19	6185163	☐	☐
20	7134801	☐	☐
21	7102286	☐	☐

Bilateral MAMOGRAFİ VE MEME USG:
MEME YAPISI: Lipomatöz (BIRADS TİP 1 MEME PATERNİ)
KİTLE LEZYONU VE MİKROKALSİFİKASYONLAR:
-Sol meme AİK-AÜK bileşkesinde yaklaşık 2.5x 1.5 cm boyutta, spiküler konturlu, yüksek dansiteli az sayıda mikrokalsifikasyon içeren malign özellikte kitlesel lezyon izlendi.
-Sağ meme ÜDK da düzensiz sınırlı, düşük dansiteli alan içerisinde amorf az sayıda mikrokalsifikasyon izlendi.
AKSİLLER BÖLGEDE LENF NODU:
-Solda 14 x 6 mm boyutta düzgün sınırlı, ekojen hilusu bulunan reaktif görünümlü özelliklerde lenf nodu izlendi.
CİLT: NORMAL
PEKTORAL KAS: NORMAL
SONUÇ:
1- Sol memede BIRADS 5 lezyon
Biyopsi önerilir.
2- Sağ meme ÜDK da düzensiz sınırlı, düşük dansiteli alan ve eşlik eden amorf az sayıda mikrokalsifikasyon(BIRADS 4A)
Mamografi eşliğinde tel ile işaretleme önerilir.
*****İnceleme Tam Alanlı Dijital Mamografi cihazı ile elde olunmuştur*****
PROF DR F BALCI DR [illegible]

Ontology-based annotation



Summary

- New Re-ranking approach was tested in Wikipedia MM task.
- Integrated Retrieval Model was evaluated in ImageCLEFmed
- Selection of mammography cases for evaluation set is already completed.
- But annotation works, which is very labor intensive activity, still in progress.