Knowledge Based Semantic Indexing

SSMS 2010, Amsterdam

Yiannis Kompatsiaris
Giorgos Papadopoulos, Stamatia Dasiopoulou, Spyros Nikolopoulos

CERTH - Informatics and Telematics Institute

http://mklab.iti.gr

SSMS'10
Summer School on Multimedia Semantics

Outline

- Introduction
- Content – Applications
- Problem Definition
- Context and Reasoning
- Combined Approaches
  - Visual + Context
  - Visual + Fuzzy DL Reasoning
  - Visual + Probabilistic Inference
- Conclusions

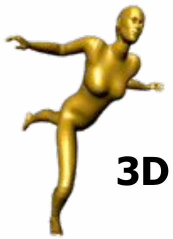
Content - Applications

Content → **Knowledge Extraction** → **Applications**

Personal

Sports - News

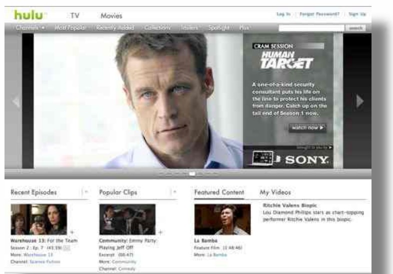
3D



Industrial

News

Web 2.0



Premium - Movies

Semantic Desktop

Personalization

Retrieval

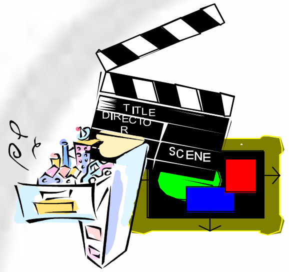
Mobile

Need for annotation + metadata

"The value of information depends on how easily it can be found, retrieved, accessed, filtered or managed in an active, personalized way…

…matching user needs"

otherwise...
we are LOST in content

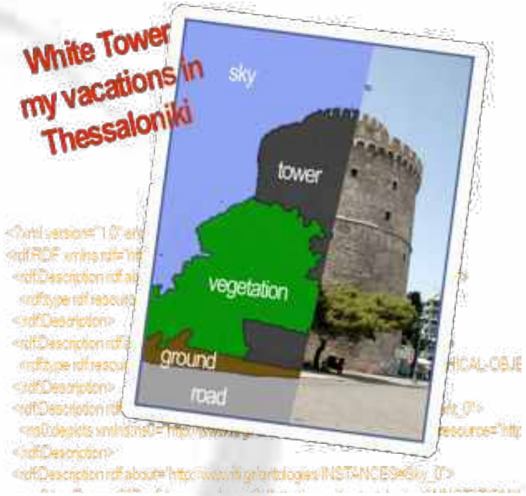


Multimedia Content

Networks

Storage & Devices

Web 2.0 photo - video applications



Segmentation KA Analysis

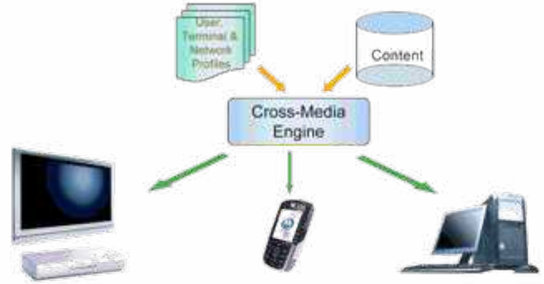
Labeling

Cross-media analysis

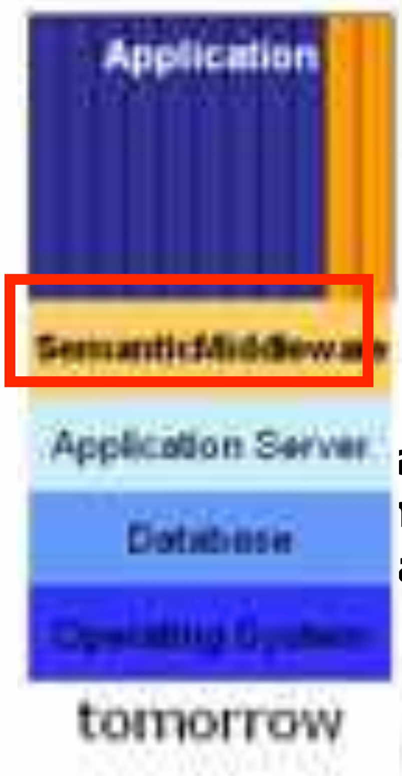
Context

Reasoning

Metadata Generation & Representation



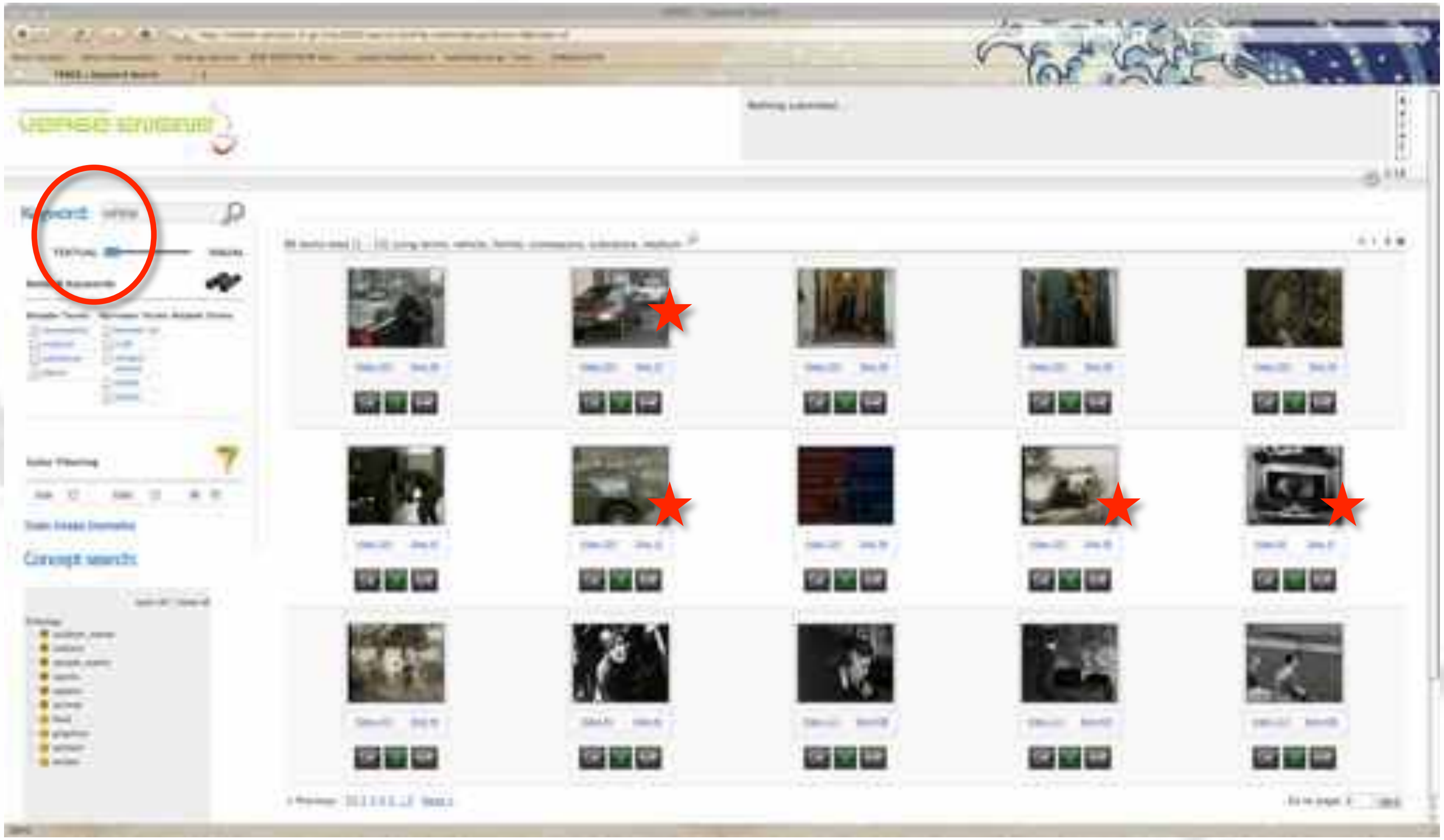
Content adaptation and distribution - Multiple Terminal &



/ Content-based
al
mendations and
alization

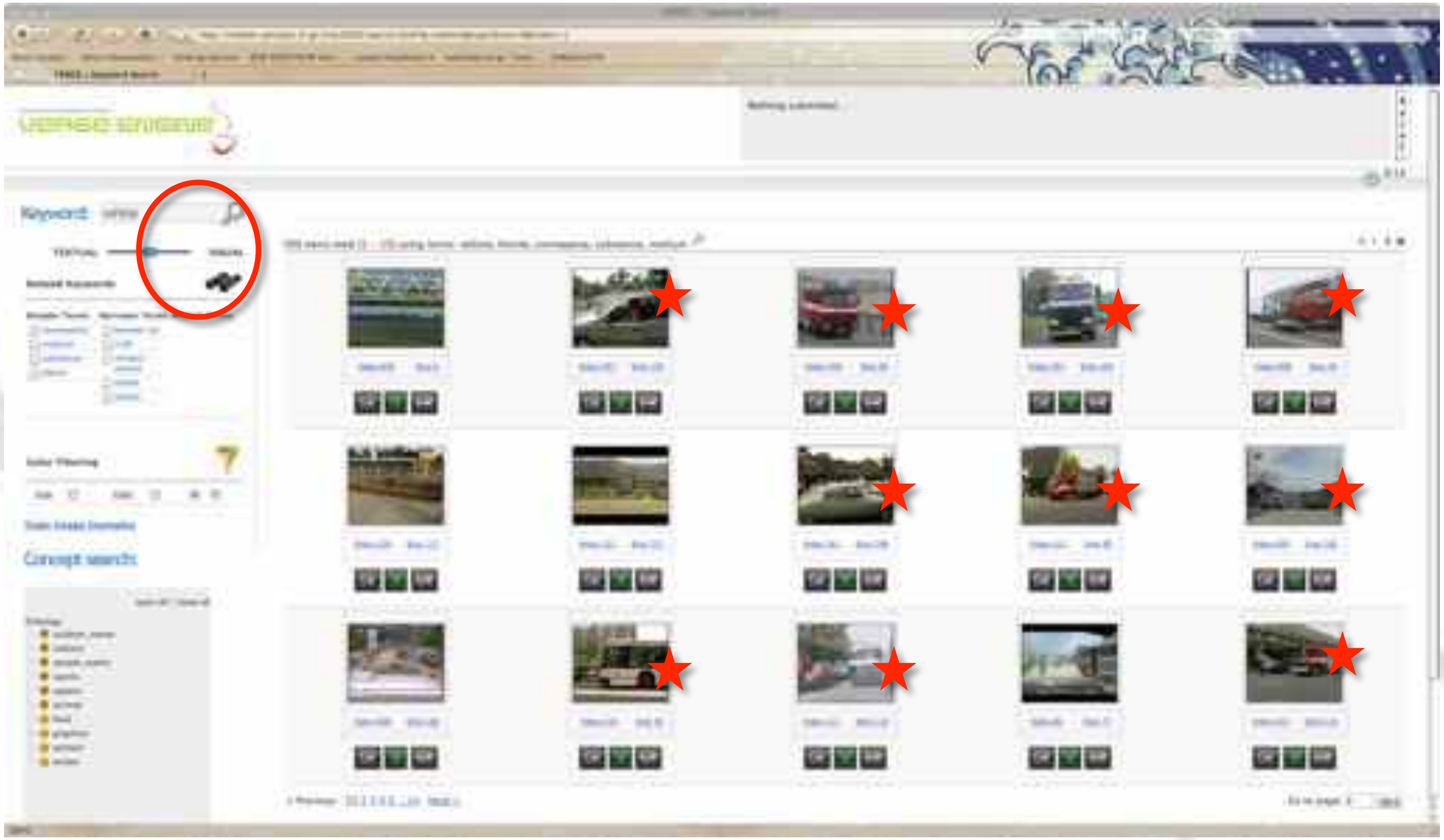
Semantic technology in Markets

Text is not always enough...



http://mklab.iti.gr/verge/

Text is not always enough…



http://mklab.iti.gr/verge/

Addressing the *Semantic Gap*

- ***Semantic Gap** for multimedia: To map automatically generated numerical low level-features to higher level human-understandable*


```
<?xml version='1.0' encoding='ISO-8859-1' ?>
<Mpeg7 xmlns…>
  <DescriptionUnit xsi:type = "DescriptorCollectionType">
    <Descriptor xsi:type = "DominantColorType">
      <SpatialCoherency>31</SpatialCoherency>
      <Value>
        <Percentage>31</Percentage>
        <Index>19  23  29 </Index>
        <ColorVariance>0  0  0 </ColorVariance>
      </Value>
    </Descriptor>
  </DescriptionUnit>
</Mpeg7>
```


This image contains a sky region and is a holiday image

Dominant Color Descriptor of a sky region

Visual features based Classification

Segment's hypothesis set

Natural-Person: 0.456798
Sailing-Boat: 0.463645
Sand: 0.476777
Building: 0.415358
Pavement: 0.454740
Road: 0.503242
Body-Of-Water: 0.489957
Cliff: 0.472907
Cloud: 0.757926
Mountain: 0.512597
Sea: 0.455338
Sky: 0.658825
Stone: 0.471733
Waterfall: 0.500000
Wave: 0.476669
Dried-Plant: 0.494825
Dried-Plant-Snowed: 0.476524
Foliage: 0.497562
Grass: 0.491781
Tree: 0.447355
Trunk: 0.493255
Snow: 0.467218
Sunset: 0.503164
Car: 0.456347
Ground: 0.454769
Lamp-Post: 0.499387
Statue: 0.501076

Semantics goes beyond perceptual manifestations

Discrepancy between semantic expressiveness

Discrepancy between intended and learned semantics

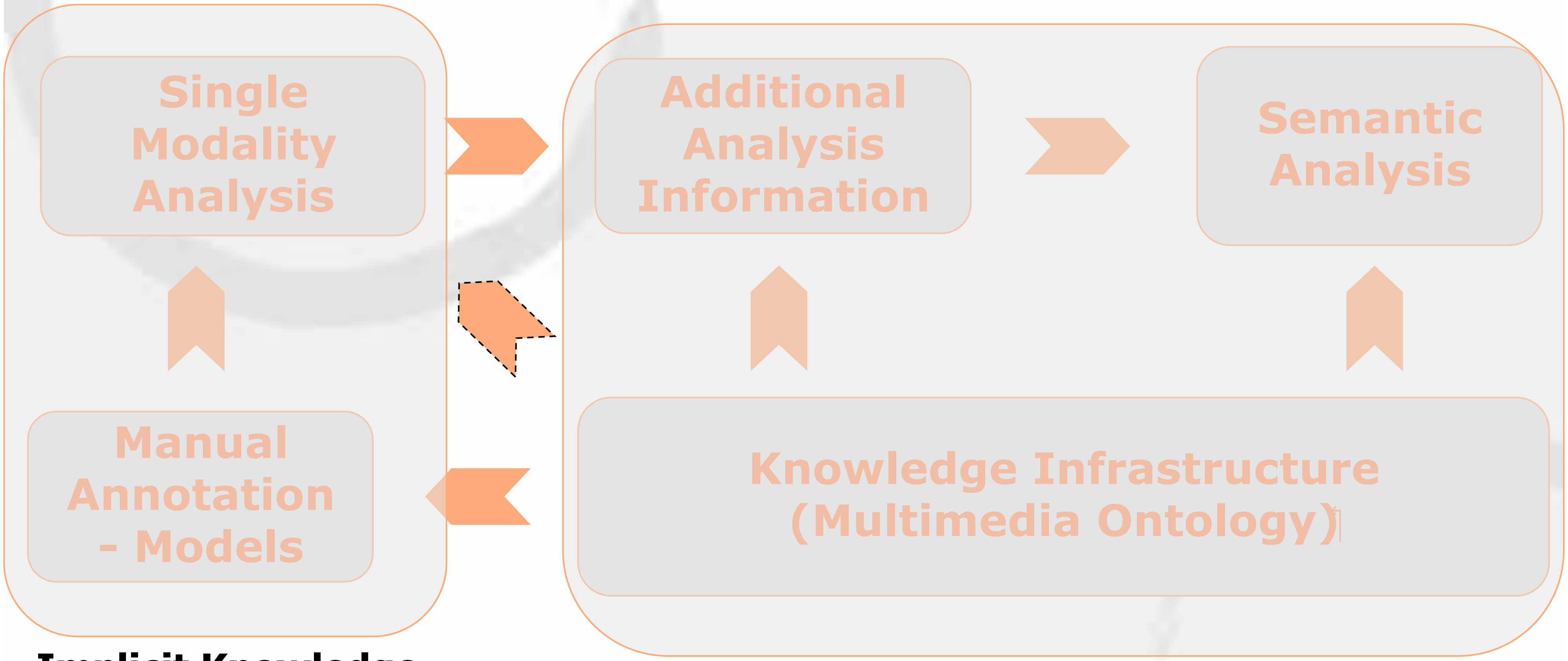
Search Topic	Best Possible: *Best Detector*	Best Possible: *AP*
Two visible tennis players on the court	Athlete	0.6501
A goal being made in a soccer match	Stadium	0.3429
Basketball players on the court	Indoor Sports Venue	0.2801
A meeting with a large table and people	Furniture	0.1045
People with banners or signs	People Marching	0.1013
One or more military vehicles	Armored Vehicles	0.0892
Helicopter in flight	Helicopters	0.0791
A road with one or more cars	Car	0.0728
An airplane taking off	Classroom	0.0526
A tall building	Office Building	0.0469
A ship or boat	Cloud	0.0427
George Bush entering or leaving vehicle	Rocket Propelled Grenades	0.0365
Omar Karami	Chair	0.0277
Graphic map of Iraq, Baghdad marked	Graphical Map	0.0269
Condoleeza Rice	US National Flag	0.0237
One or more palm trees	Weapons	0.0225

Snoek et al., "Adding Semantics to Detectors for Video Retrieval", IEEE Multimedia, 2007

Problem definition

- **Semantic image and video analysis**: *how to translate the automatically extracted visual descriptions into human like conceptual ones*
- **Low-level features** provide ***cues*** *for strengthen/weaken evidence based on visual similarity*
- ***Prior knowledge*** *is needed to support semantics disambiguation / enforce coherent interpretations*

Knowledge Extraction
A common view



Implicit Knowledge
Signal Level

Explicit Knowledge – Logic - Semantics
& Hybrid Level

Knowledge Extraction
A common view

Feature extraction
Text, Image analysis
Segmentation, SVMs
Evidence generation
"Vehicle", "Building"

Analysis

Classifiers fusion
Global vs. Local
Modalities fusion
Context
"Ambulance"

Information

Reasoning
Fusion of annotations
Consistency checking
Higher-level concepts/
events
"Emergency scene"

Analysis

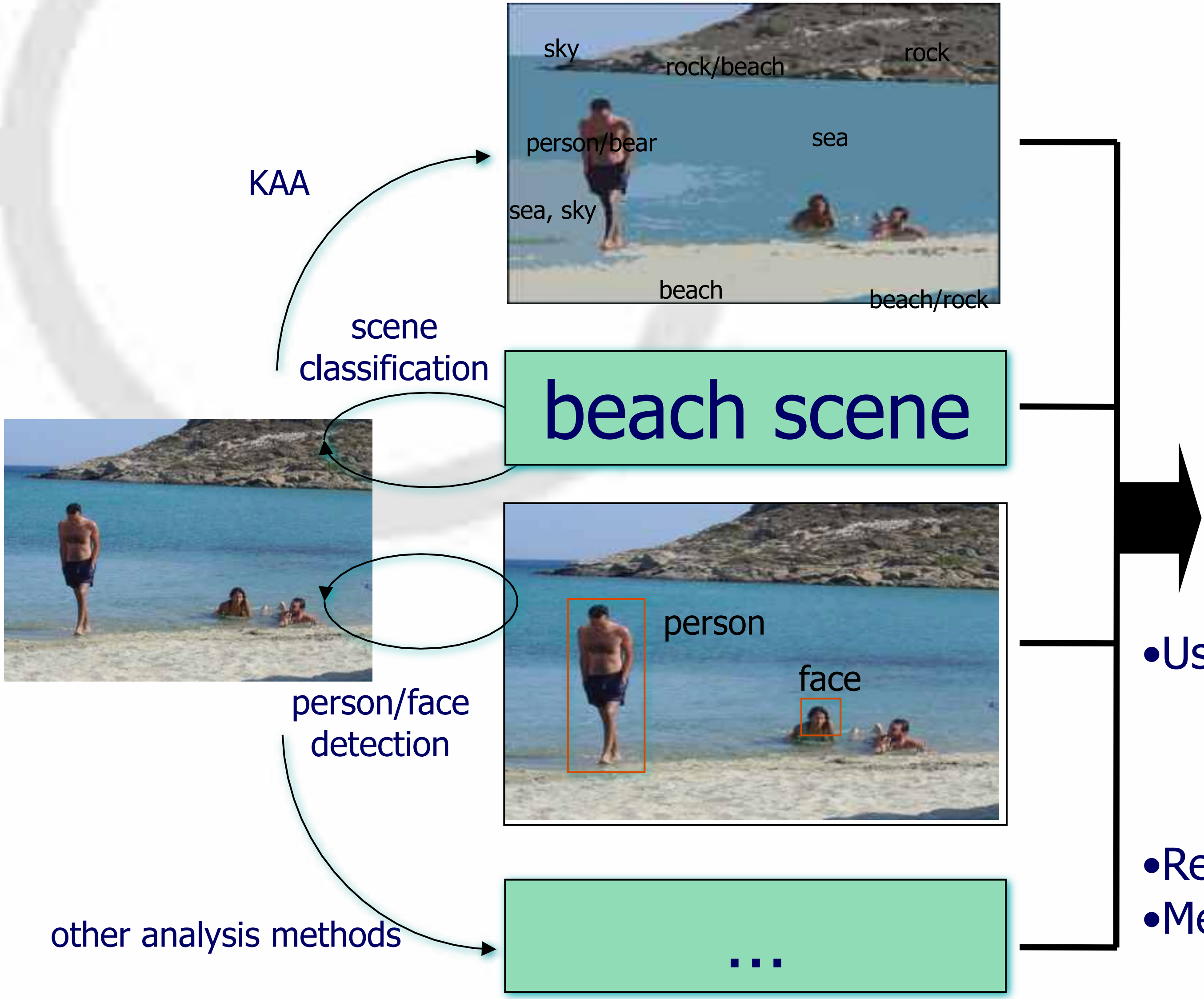
Manual

Multimedia content
annotation tools
Training
(Statistical)
Modeling

Knowledge Infrastructure
(Multim

Domain
Multimedia content
Annotations
Algorithms - Features
Context

Context and Reasoning for Analysis



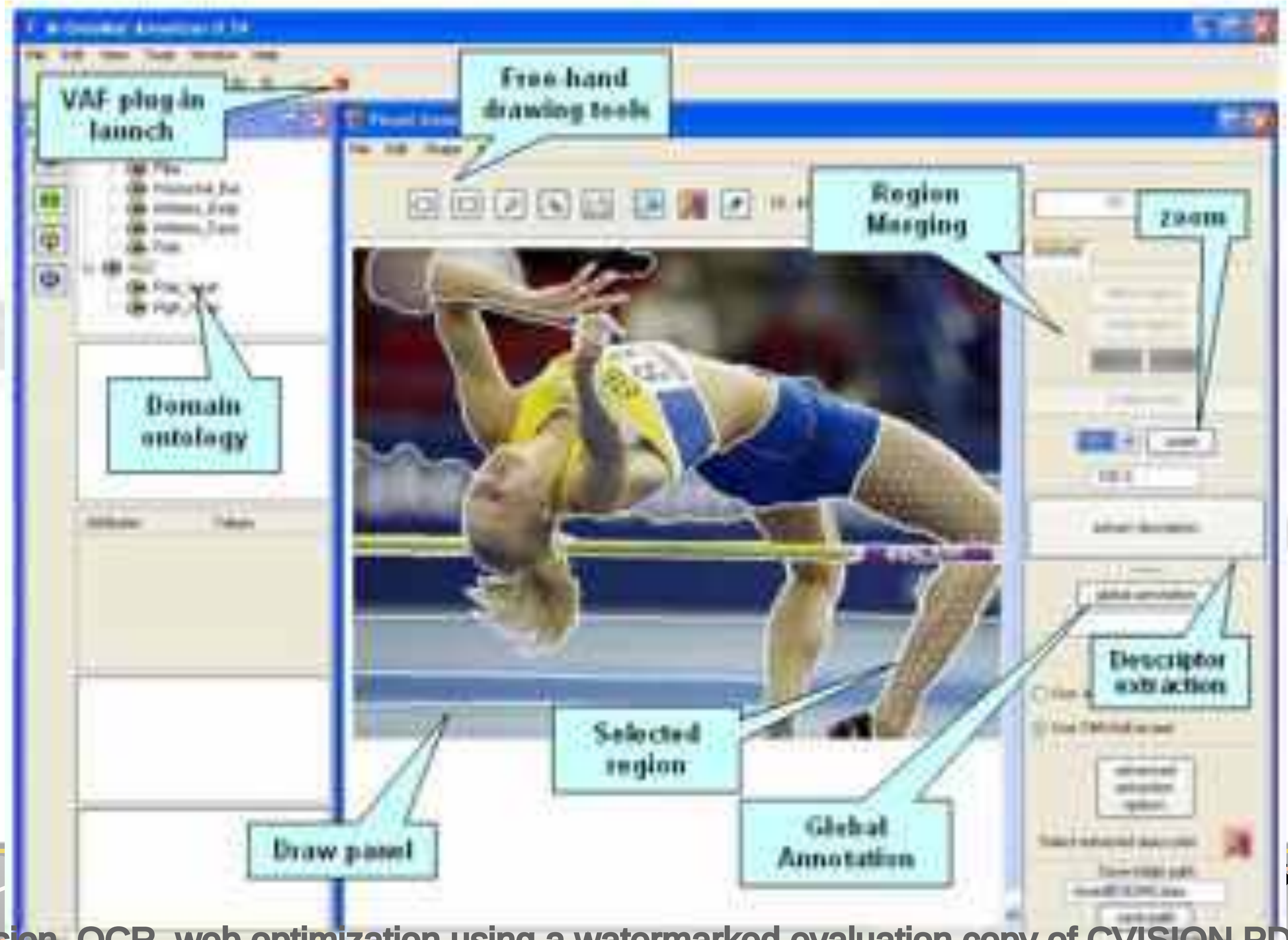
Creation of contextual information

multimedia reasoning

- Use of contextual information
    - From metadata layer
    - spatio-temporal relations
    - Domain knowledge
- Reduction of label sets
- Merging of segments

Multimedia Content Annotation

VIA: http://mklab.iti.gr/via/



Multimedia Content Analysis (Implicit)

- MPEG-7, SIFT, ... widely used for LL features
- Segmentation and feature extraction tools
- Well-known classifiers applied and developed
  - SVMs, EM, HMM
  - Bio-inspired approaches
- Increasing use of context
  - Spatial, Frequency, EXIF
- Fusion
  - Classifiers (global+local)
  - Modalities
    - Text+Image+1D data
    - Text+Speech+Video
    - Tags+Image (Web 2.0)
- Mostly statistical and machine learning (implicit) based but also
  - Hybrid (implicit + explicit)

Support Vector Machines

- Widely used in semantic image analysis tasks due to their reported generalization ability
- Receive as input the estimated region-level descriptors
- An individual SVM introduced for every defined high-level semantic concept
- ‘one-against-all’ approach followed for training
- Each SVM estimates degree of confidence for region–concept association
- Every region evaluated by all trained SVMs

Approaches

- Spatial Context
  - Optimisation: genetic algorithm
- Fuzzy DL Reasoning
  - Imprecise ontology reasoning: fuzzy DLs
- Probabilistic inference
  - Bayesian network

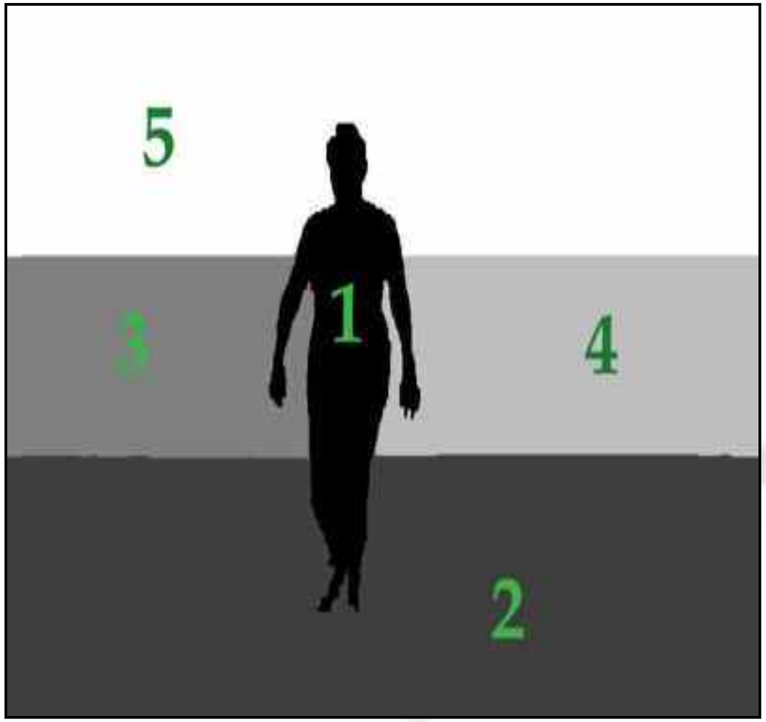
Spatial Context

Use of Context: Spatial Relations

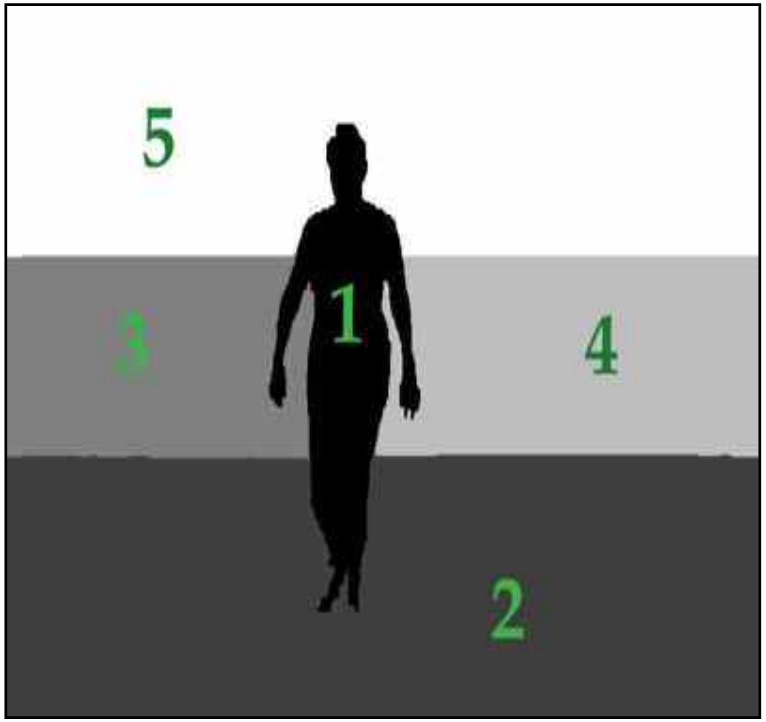
- Objects tend to be present in a scene within a particular spatial context
- Spatial information can assist in discriminating between objects exhibiting similar visual characteristics
- Directional relations: denote the order of objects in space
- Eight relations supported: Above, Above-Right, Above-Left, Right, Left, Below, Below-Right, Below-Left

Spatial Context Demonstration

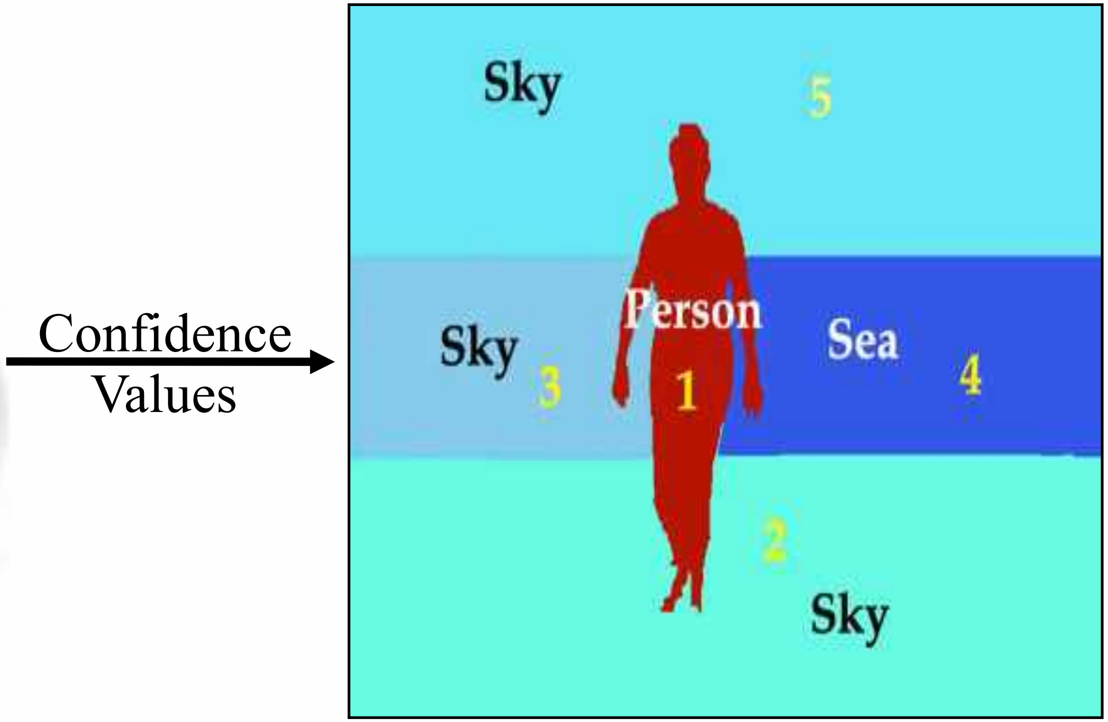
Initial image



Segmentation Mask



	Sea	Sky	Sand	Person
Region 1	0.05	0.03	0.07	**1.00**
Region 2	0.28	0.42	**0.30**	0.00
Region 3	**0.54**	0.74	0.32	0.00
Region 4	**0.79**	0.54	0.43	0.08
Region 5	0.00	**0.80**	0.03	0.09

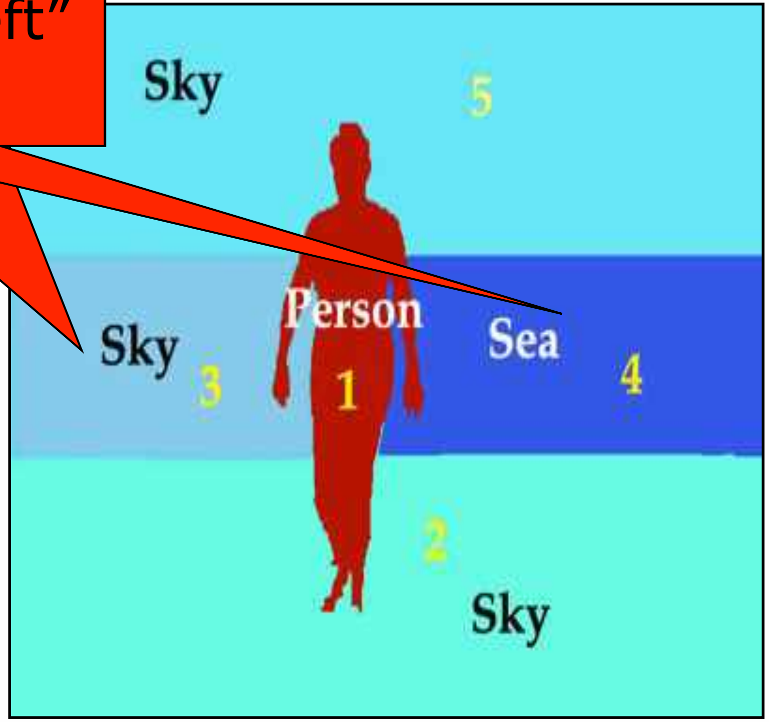


	Sea	Sky	Sand	Person
Region 1	0.05	0.03	0.07	1.00
Region 2	0.28	0.42	0.30	0.00
Region 3	0.54	0.74	0.32	0.00
Region 4	0.79	0.54	0.43	0.08
Region 5	0.00	0.80	0.03	0.09

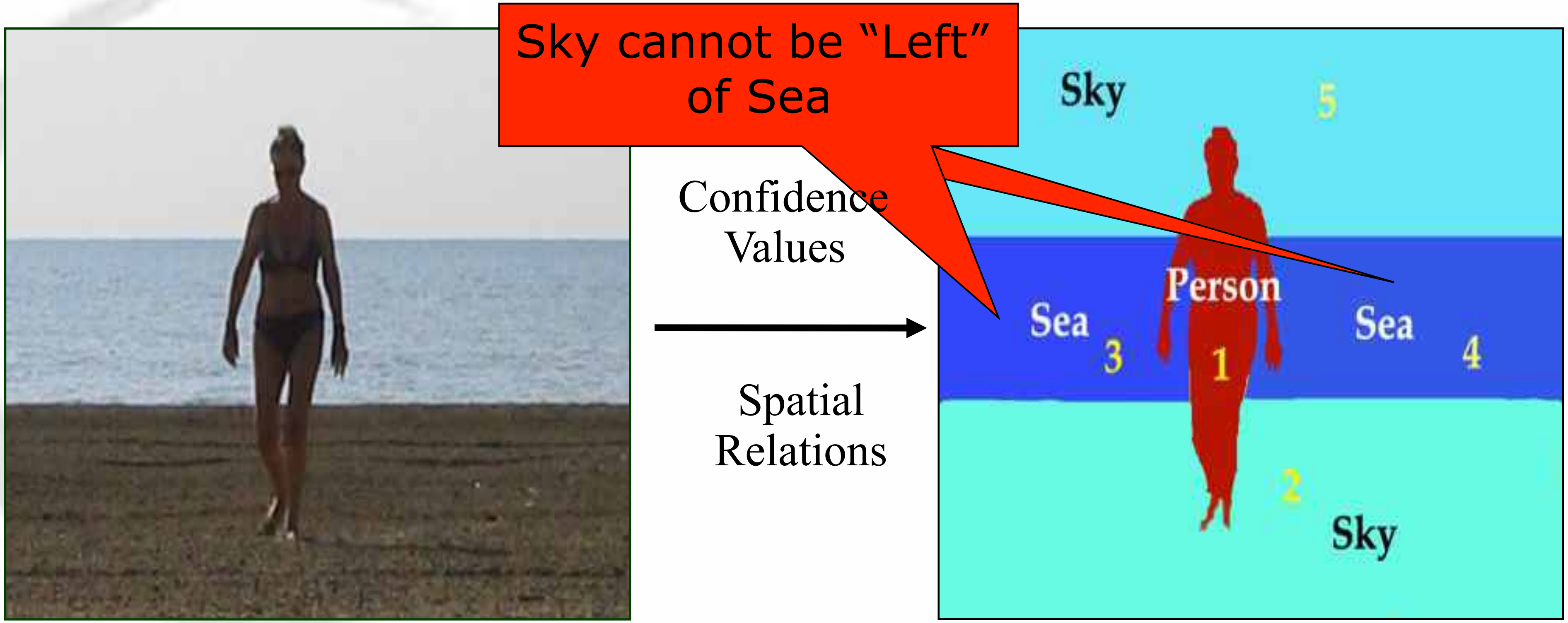
Sky cannot be "Left" of Sea

Confidence Values

Spatial Relations



	Sea	Sky	Sand	Person
Region 1	0.05	0.03	0.07	1.00
Region 2	0.28	0.42	0.30	0.00
Region 3	0.54	0.74	0.32	0.00
Region 4	0.79	0.54	0.43	0.08
Region 5	0.00	0.80	0.03	0.09

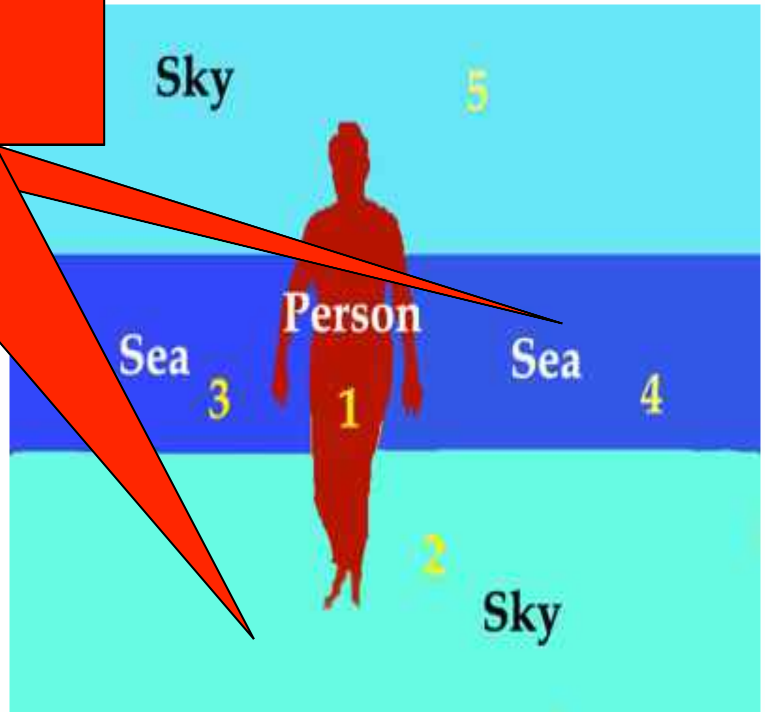


	Sea	Sky	Sand	Person
Region 1	0.05	0.03	0.07	**1.00**
Region 2	0.28	**0.42**	**0.30**	0.00
Region 3	**0,54**	~~0.74~~	0.32	0.00
Region 4	**0.79**	0.54	0.43	0.08
Region 5	0.00	**0.80**	0.03	0.09

Sky cannot be "Below" Sea

Confidence Values

Spatial Relations

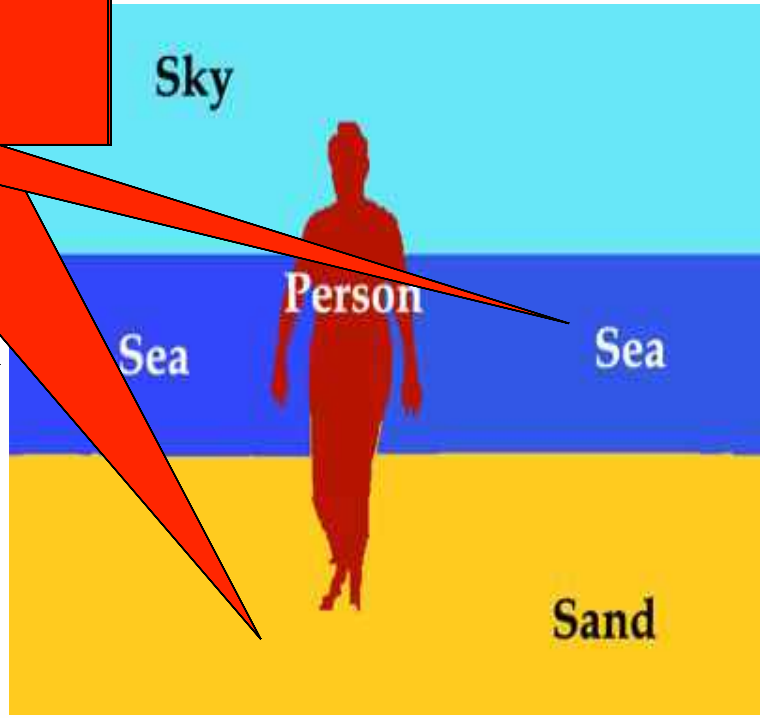


	Sea	Sky	Sand	Person
Region 1	0.05	0.03	0.07	**1.00**
Region 2	0.28	**0.42**	**0.30**	0.00
Region 3	**0.54**	**0.74**	0.32	0.00
Region 4	**0.79**	0.54	0.43	0.08
Region 5	0.00	**0.80**	0.03	0.09

Sky cannot be
“Below” Sea

Confidence
Values

Spatial
Relations



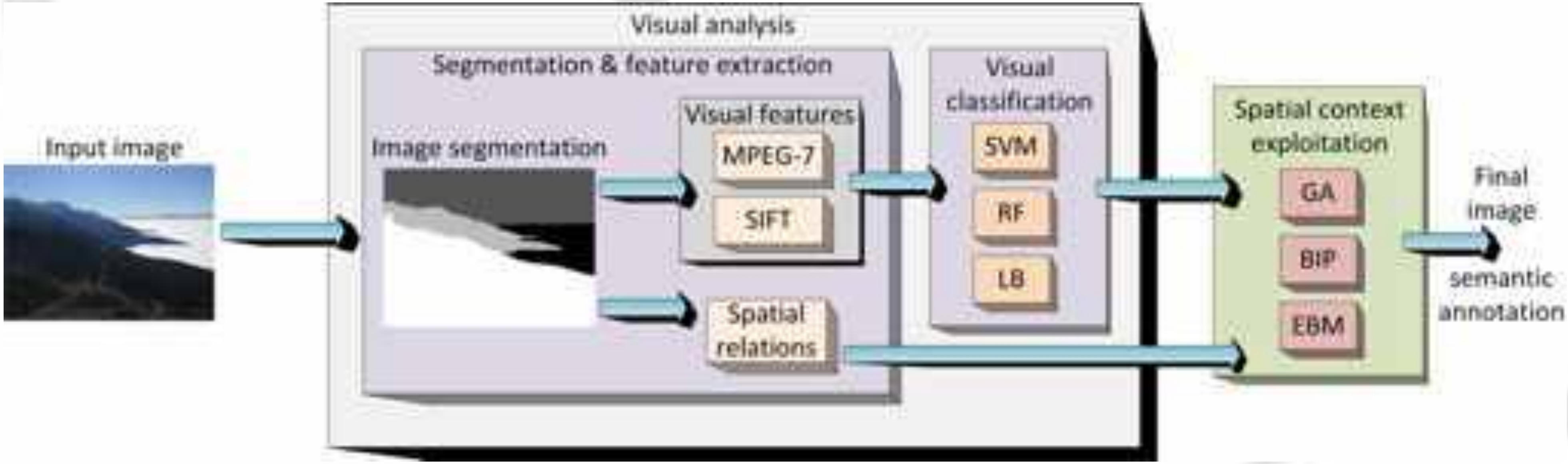
	Sea	Sky	Sand	Person
Region 1	0.05	0.03	0.07	**1.00**
Region 2	0.28	~~0.42~~	**0,30**	0.00
Region 3	**0.54**	**0.74**	0.32	0.00
Region 4	**0.79**	0.54	0.43	0.08
Region 5	0.00	**0.80**	0.03	0.09

Spatial context comparative evaluation

- Aim:
  - In-depth investigation of the advantages of different spatial context techniques
    - The selected techniques cover the main categories of the approaches proposed in the literature
  - Gain of a better insight on the use of spatial context
- Developed framework:
  - Techniques: Genetic Algorithm, Energy-based Model, Binary Integer Programming
  - Datasets: Coastal scenes (D1) – 7 concepts, SCEF[1] (D2) – 10 concepts, Personal collection (D3) – 17 concepts, MSRC (D4) – 21 concepts
  - Features: MPEG-7, SIFT
  - Classifiers: Support Vector Machines, Random Forest, Logitboost
  - Spatial relations: Fuzzy directional relations
    - Above, Above-right, Above-left, Below, Below-right, Below-left, Right, Left

[1] http://mklab.iti.gr/project/scef

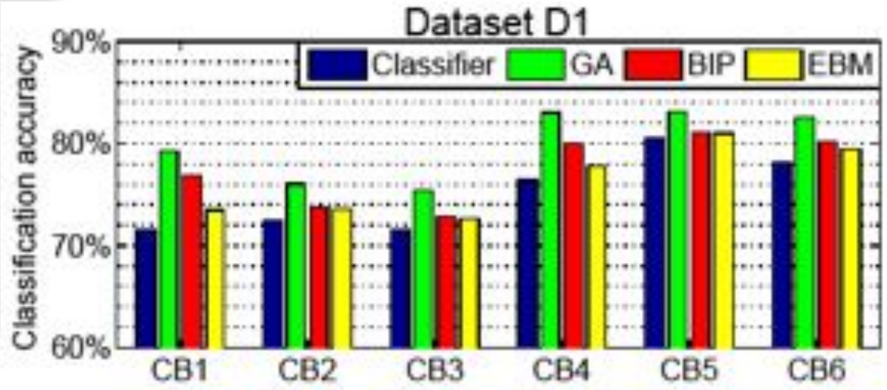
Developed evaluation framework

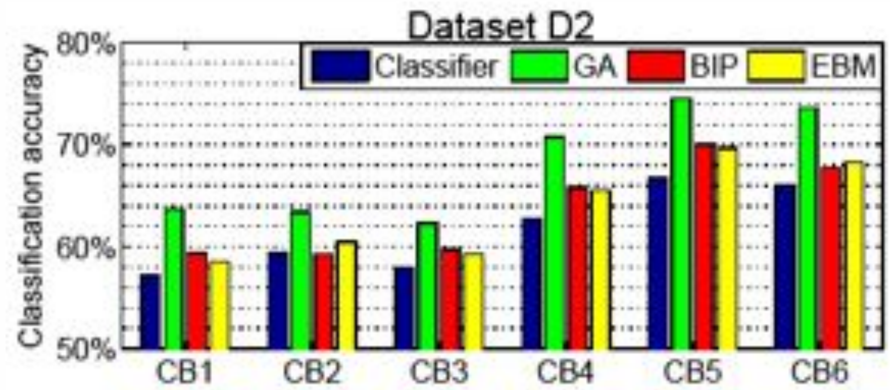


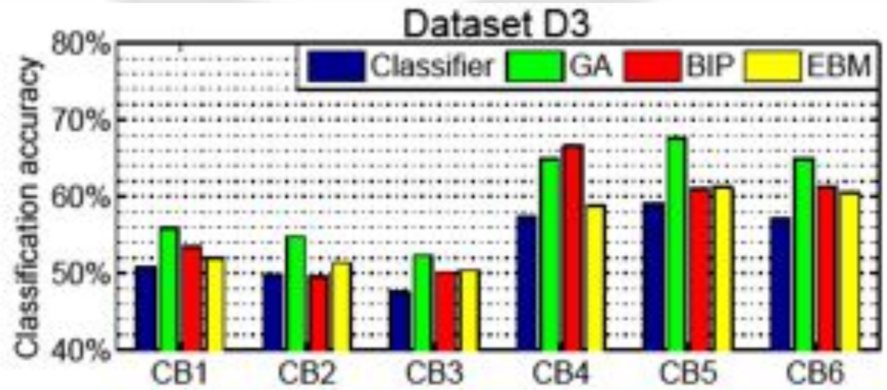
Spatial context techniques

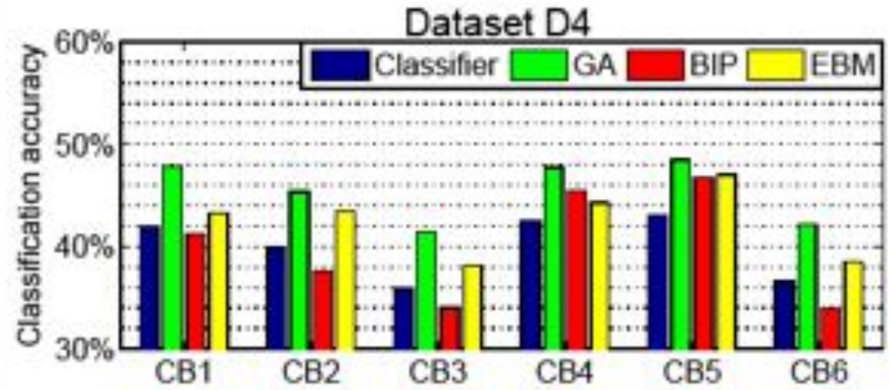
- Genetic Algorithm
  - Realizes image analysis as a global optimization problem
  - Makes use of complex fuzzy spatial constraints
  - Uses a set of Bayesian Networks for combining the spatial, visual and concept co-occurrence information
- Binary Integer Programming
  - Formalizes the spatial constraints enforcement as a binary integer problem
  - Uses binary constraints
  - Utilization of product operator for information fusion
- Energy-based Model
  - Reduces the region labeling problem to that of minimizing an energy function of a graphical model
  - Uses fuzzy constraints
  - Assigns a global weight factor to each information source

Overall spatial context results





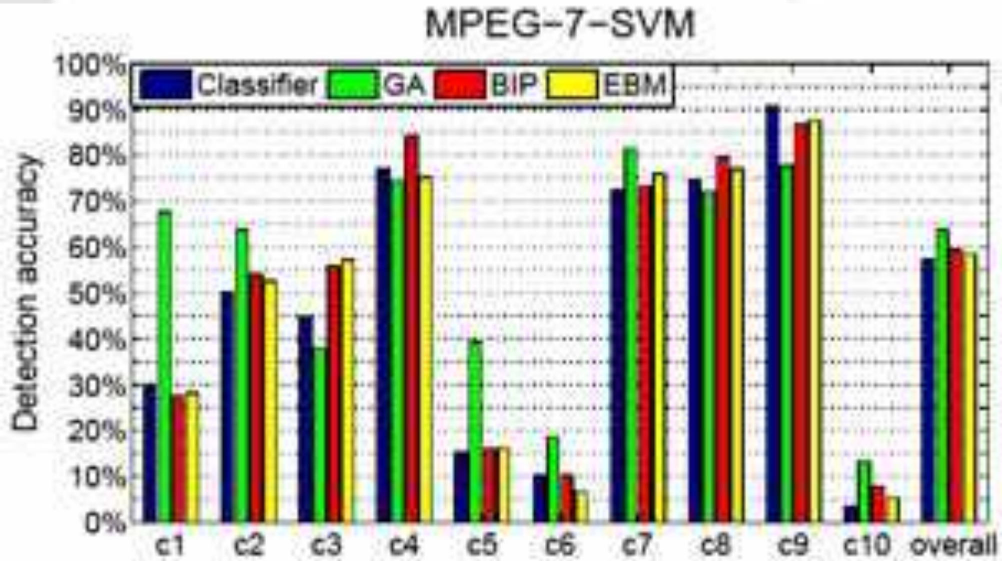


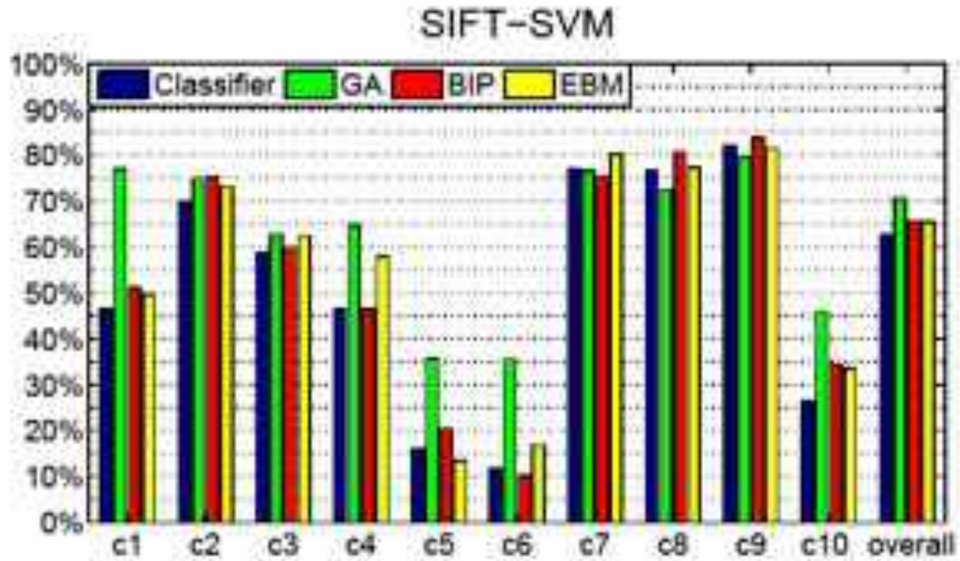


Combinations: CB1: MPEG-7-SVM, CB2: MPEG-7-RF, CB3: MPEG-7-LB, CB4: SIFT-SVM, CB5: SIFT-RF, CB6: SIFT-LB

Indicative concept-level results

D2 dataset

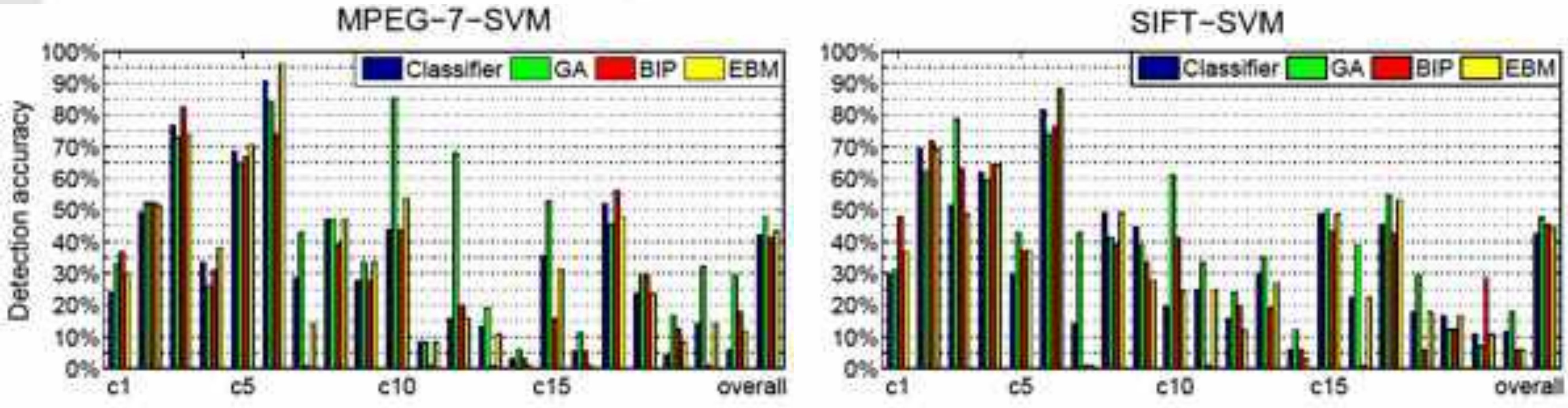




Supported concepts: c1: Building, c2: Foliage, c3: Mountain, c4: Person, c5: Road, c6: Sailing-boat, c7: Sand, c8: Sea, c9: Sky, c10: Snow

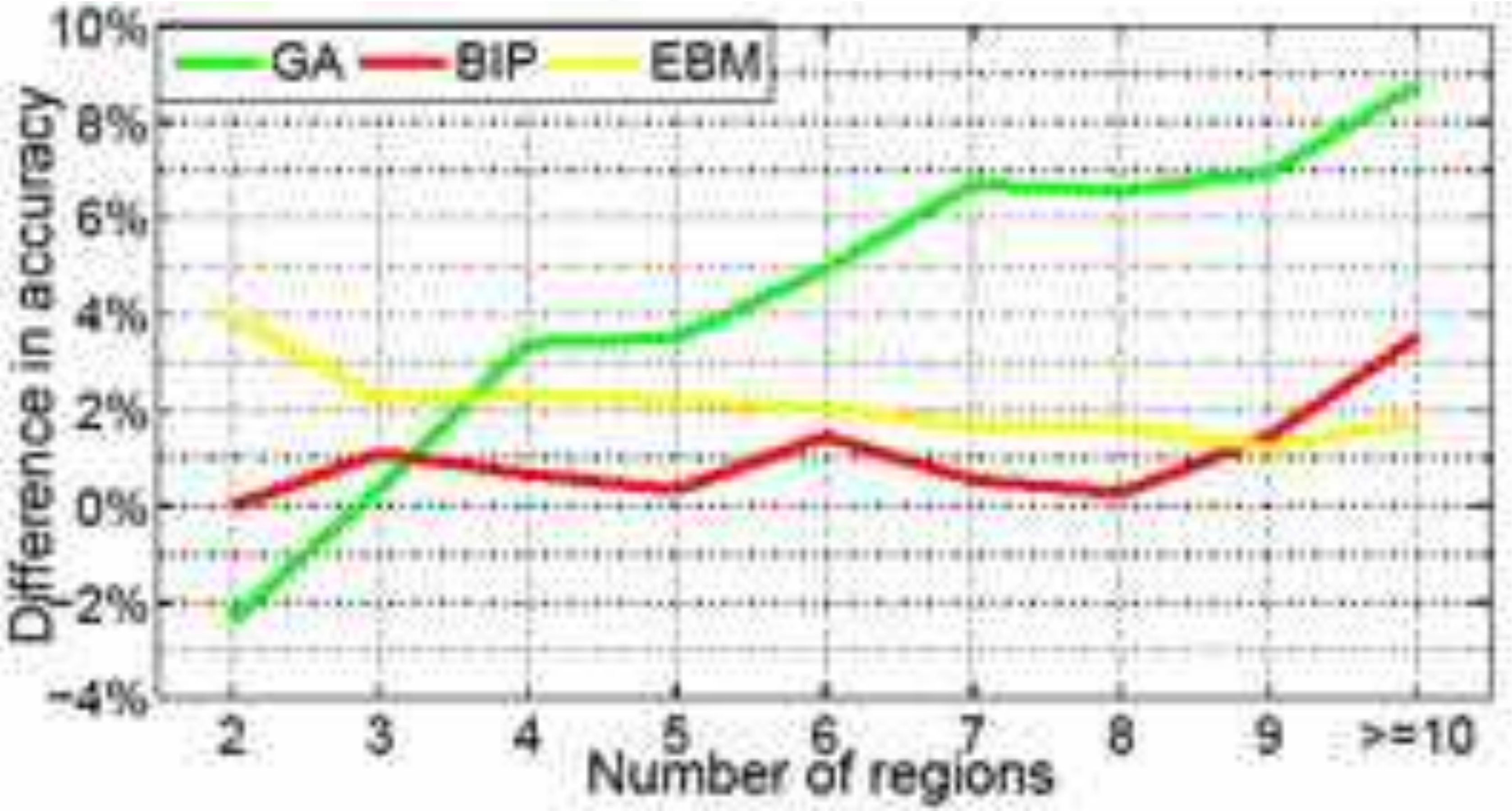
Indicative concept-level results (cont'd)

D4 dataset

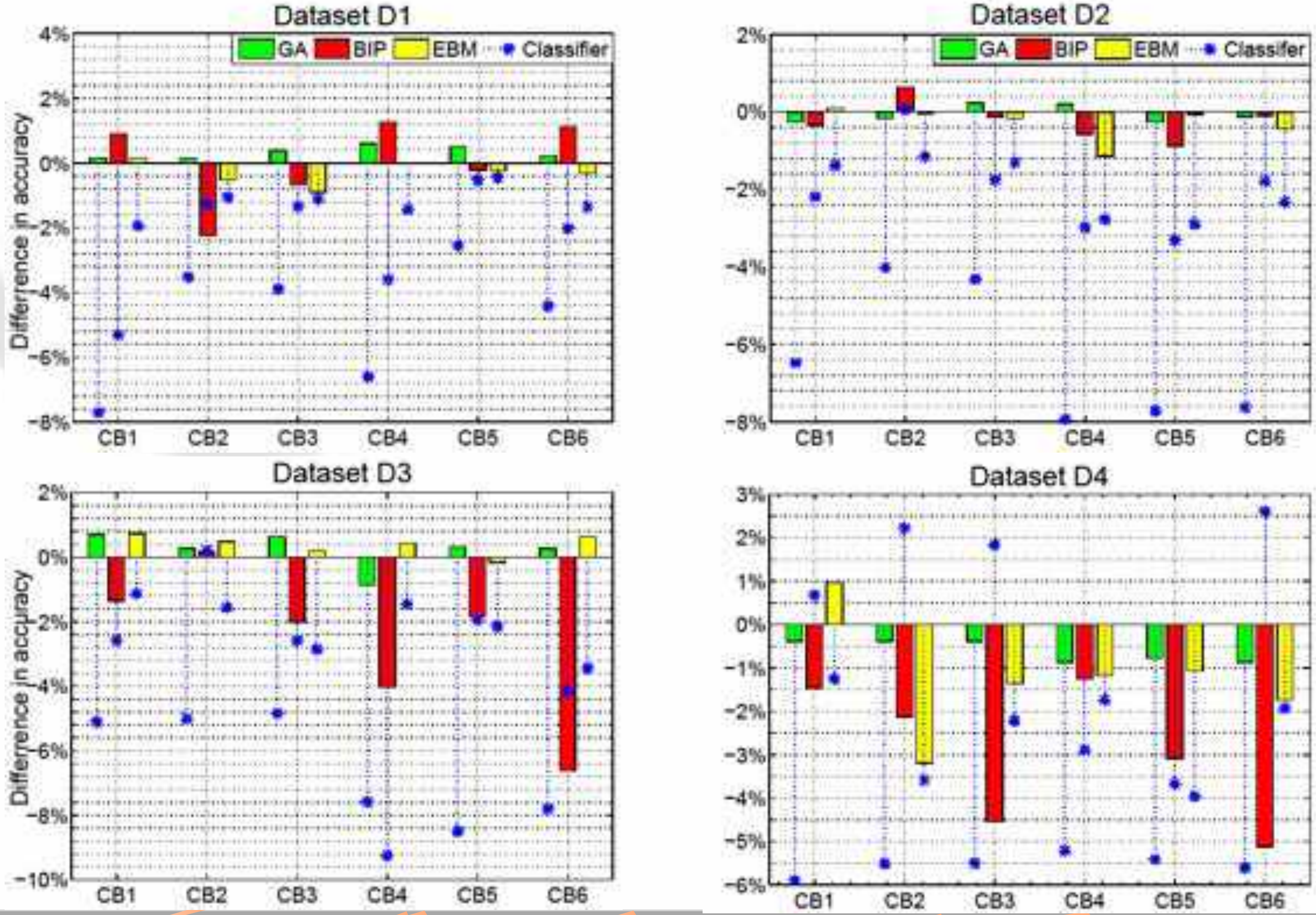


Supported concepts: c1: Building, c2: Grass, c3: Tree, c4: Cow, c5: Sheep, c6: Sky, c7: Aeroplane, c8: Water, c9: Face, c10: Car, c11: Bicycle, c12: Flower, c13: Sign, c14: Bird, c15: Book, c16: Chair, c17: Road, c18: Cat, c19: Dog, c20: Body, c21: Boat

Effect of the number of image regions



Effect of the amount of data used for context acquisition



Conclusions on the use of spatial context

- Spatial context is efficient in improving the initial (i.e. based solely on visual features) region-concept association results
- The highest performance is achieved when complex spatial constraints are acquired and their weight against the visual and co-occurrence information is efficiently adjusted
- The improvement over the initial classification results tends to decrease when the number of supported concepts increases
- For a given dataset, the highest initial classification performance leads also to the highest spatial context performance
- Fuzzy spatial constraints are less likely to result in performance decreases when the amount of training data is reduced, compared to binary constraints

Comparison among techniques

Factors considered	Spatial context techniques		
	GA	BIP	EBM
Concepts favored	Concepts with more well-defined spatial context and concepts with low initial classification rate	Concepts with less well-defined spatial context	Concepts with more well-defined spatial context
Number of image regions	Continuous increase in performance improvement, when the number of regions increases ($N \geq 4$)	Significant performance improvement only when the number of regions is significantly high ($N \geq 10$)	Highest performance when very few image regions are present ($N = 2$)
Reduction in amount of training data	Small changes in performance (changes $< 1\%$)	Significant performance reduction in datasets with many concepts (up to –6,62%)	Performance reduction in datasets with many concepts (up to –3,19%)

Fuzzy DL reasoning

Reasoning in multimedia content analysis

- Imprecision
  - Uncertainty (degrees of probability)/Vagueness (degrees of truth), Incompleteness (missing input, background knowledge)
- Formal approaches
  - Fuzzy/probabilistic/possibilistic logic (DLs, rules), abductive reasoning, inductive reasoning…
- Statistical approaches
  - Bayesian inference, HMMs…
- Used for:
  - Fusion / Integration
  - Consistency checking
  - Higher-level abstraction results

Why Reasoning in MM Annotation?

- **Problem Definition**
  - Machine learning provides now generic methodologies for supporting more than 100 concepts
    - captures conveniently complex associations between perceptual features and semantics
    - successful application examples, yet highly variable general performance
  - **Semantics goes beyond perceptual manifestations**
    - *possibly contradictory (Mountain, Sand and Indoor)*
    - *possibly overlapping / complementary (Beach and Sea)*
    - *of restricted abstraction w.r.t. semantic expressiveness (Person inside Sea vs Swimmer)*
  - Learning-based extracted annotations need to be ***semantically interpreted*** into a ***consistent,*** meaningful final description

Semantics goes beyond perceptual manifestations

$(image : Countryside\_buildings) \geq 0.65$
$(image : Roadside) \geq 0.57$
$(image : Rockyside) \geq 0.44$
$(image : Forest) \geq 0.45$
$(image : Seaside) \geq 0.47$
$(image : \exists contains.Sand) \geq 0.66$
$(image : \exists contains.Sky) \geq 0.95$
$(image : \exists contains.Person) \geq 0.62$
$(image : \exists contains.Foliage) \geq 0.70$

$(image : Rockyside) \geq 0.42$
$(image : Countryside\_buildings) \geq 0.52$
$(image : Seaside) \geq 0.51$
$(image : Forest) \geq 0.52$
$(image : Roadside) \geq 0.71$
$(image : \exists contains.Sky) \geq 0.98$
$(image : \exists contains.Sea) \geq 0.73$
$(image : \exists contains.Person) \geq 0.60$
$(image : \exists contains.Sand) \geq 0.75$

Our Approach: Fuzzy DLs based Reasoning in Multimedia Annotation

- **Goal:** enhance the robustness and completeness of learning-based extracted annotations
  - annotations at object and scene level
  - different implementations
- **How:** semantics utilisation
  - to integrate initial annotations
  - to detect and resolve inconsistencies
  - to enrich by means of logical entailment
- **Methodology:** fuzzy DLs reasoning framework
  - Crisp TBox to conceptualise the domain semantics
  - Fuzzy assertions to capture the imprecision of initial annotations

DLs in brief

- Family of knowledge representation languages characterised by **formal** semantics and **sound & complete** inference algorithms

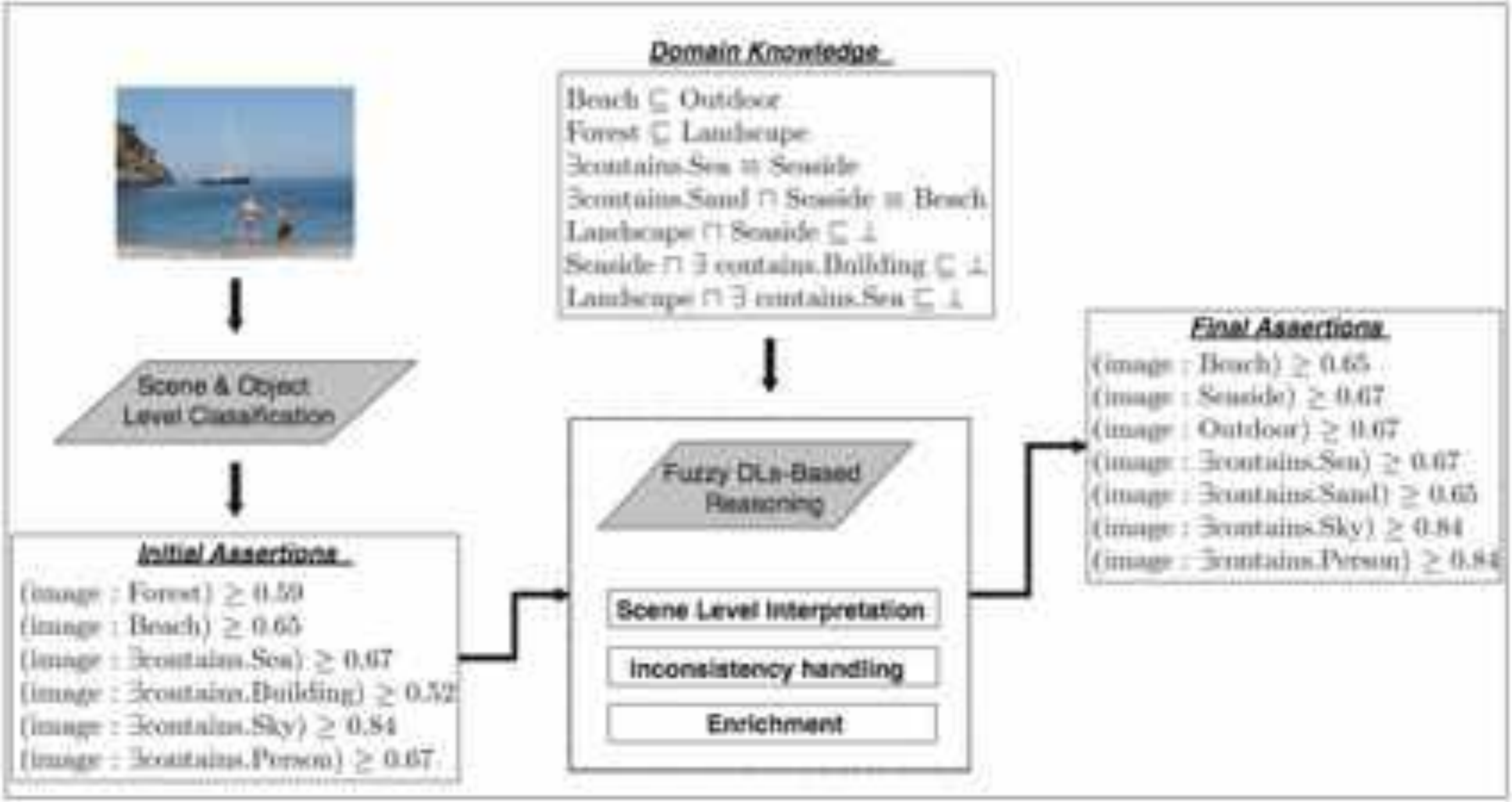
- **Terminological Box** (TBox): vocabulary (concepts & roles) and interrelations describing the application domain
  - equivalence
  - subsumption
  - complex descriptions inductively build with constructors

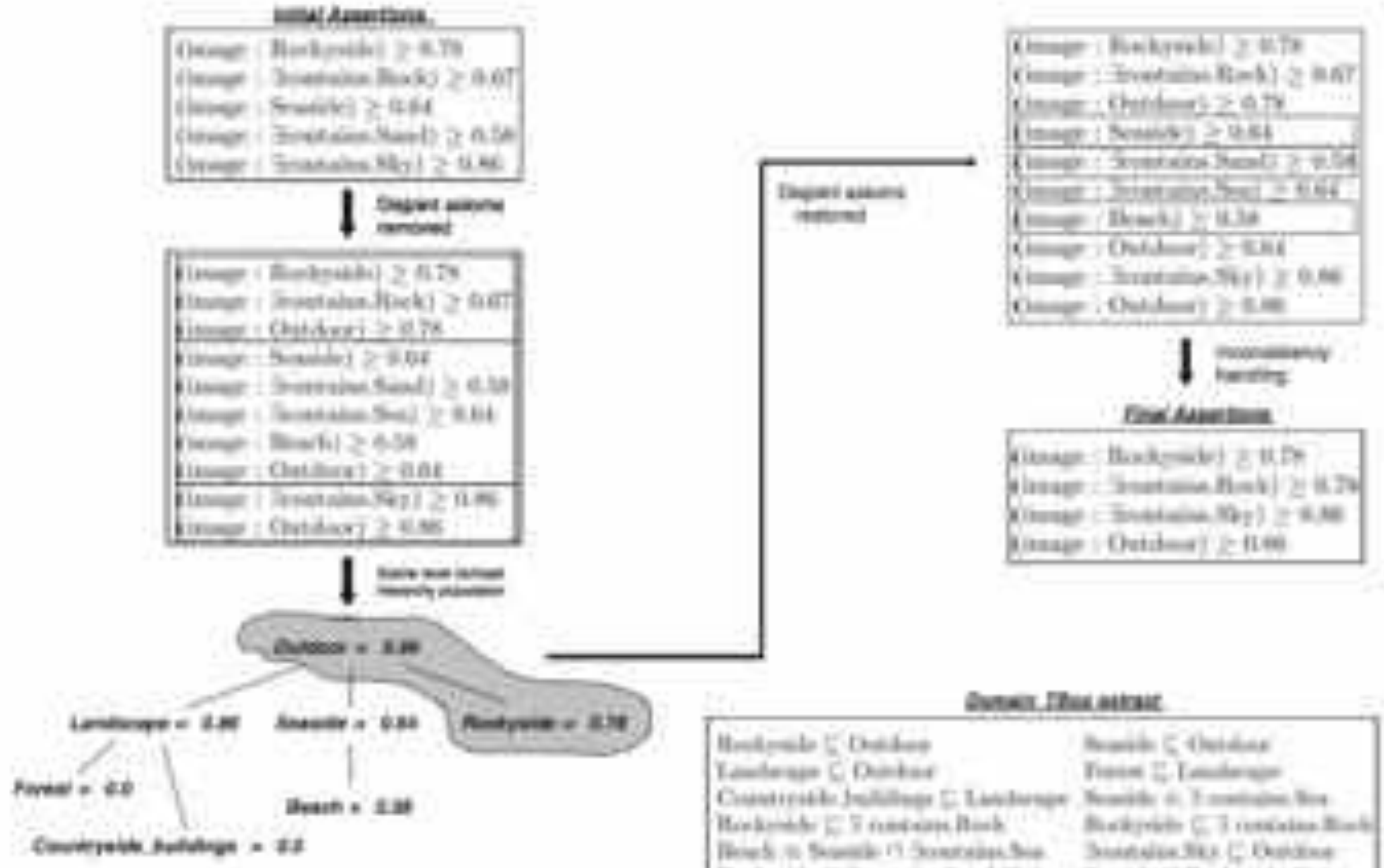
$$\mathrm{Mother} \equiv \mathrm{Woman} \sqcap \exists\ \mathrm{hasChild.Person}$$
$$\mathrm{Tree} \sqsubseteq \exists\ \mathrm{hasPart.Leaf} \sqcap \exists\ \mathrm{hasPart.Trunk}$$

- **Assertional ABox** (ABox): facts describing a specific state of the application domain
  - concept assertions
  - role assertions

$$\mathrm{Athlete(John)},\ \mathrm{Woman(Myriam)}$$
$$\mathrm{hasChild(Myriam,John)}$$

Three Reasoning Tasks




Initial Assertions
Final Assertions

Three Reasoning Tasks (cont'd)

- T1 - scene level interpretation: find plausible (logically admitted) interpretations

- T2 – consistency handling: track and resolve inconsistencies

- T3 – Enrichment: augment final interpretation making entailments explicit

Three Reasoning Tasks - I

- **T1 – Scene level interpretation**
  - involves both asserted and inferred assertions of scene level concepts
  - removes disjointness axioms from TBox to consider all related assertions (disjointness semantics maintained separately)
  - computes scene level concept hierarchy
  - starting from the leaf concepts maintains between conflicting assertions the one with highest degree
  - propagates degrees according to fuzzy subsumption semantics to the next level
  - repeats procedure, if current prevalent assertions contradict the previous level (i.e. have higher plausibility) remove and update accordingly the previous level
  - procedure ends when reaching the top level concepts

Scene level interpretation demonstration

Domain TBox

$Natural \equiv Outdoors \sqcup \neg ManMade$
$Mountainous \equiv Natural \sqcup \neg Coastal$
$Beach \equiv Coastal \sqcap \exists contains.Sand$
$\exists contains.Mountain \sqsubseteq Mountainous$
$\exists contains.Sea \sqsubseteq Coastal$
$\exists contains.Sand \sqcap Mountainous \sqsubseteq \bot$
$Outdoor \sqcap Indoor \sqsubseteq \bot$

Initial Assertions

$(image{:}Indoor) \geq 0.67$
$(image{:}\exists contains.Sea) \geq 0.73$
$(image{:}\exists contains.Sand) \geq 0.58$
$(image{:}\exists contains.Mountain) \geq 0.85$

Disjointness axioms removed

$(image{:}Indoor) \geq 0.67$
$(image{:}\exists contains.Sea) \geq 0.73$
$(image{:}\exists contains.Sand) \geq 0.58$
$(image{:}Coastal) \geq 0.73$
$(image{:}Beach) \geq 0.58$
$(image{:}Natural) \geq 0.73$
$(image{:}Outdoor) \geq 0.73$
$(image{:}\exists contains.Mountain) \geq 0.85$
$(image{:}Mountainous) \geq 0.85$
$(image{:}Natural) \geq 0.85$
$(image{:}Outdoor) \geq 0.85$

Scene level hierarchy

Outdoor (0.85) Indoor (0.67)
Natural (0.85) ManMade
Coastal (0.58) Mountainous (0.85)
Beach (0.58)

Three Reasoning Tasks - II

- **T2 – Consistency handling**
  - performs over the initial set of annotations
  - removes all assertions (asserted & inferred) pertaining to object level concepts disjoint to T1 interpretation
  - removes all assertions pertaining to scene level concepts disjoint to T1 interpretation
  - removal of assertions is performed w.r.t. to the type of inclusion axioms they appear in
  - in case of more than one consistent descriptions we chose the one that requires the removal of assertions with the lowest average degree

Consistency handling demonstration

Initial Assertions

(image:Indoor) $\geq$ 0.67
(image:$\exists$contains.Sea) $\geq$ 0.73
(image:$\exists$contains.Sand) $\geq$ 0.58
(image:$\exists$contains.Mountain) $\geq$ 0.85

Disjointness axioms removed

(image:Indoor) $\geq$ 0.67
(image:$\exists$contains.Sea) $\geq$ 0.73
(image:$\exists$contains.Sand) $\geq$ 0.58
(image:Coastal) $\geq$ 0.73
(image:Beach) $\geq$ 0.58
(image:Natural) $\geq$ 0.73
(image:Outdoor) $\geq$ 0.73
(image:$\exists$contains.Mountain) $\geq$ 0.85
(image:Mountainous) $\geq$ 0.85
(image:Natural) $\geq$ 0.85
(image:Outdoor) $\geq$ 0.85

Scene level hierarchy

Outdoor (0.85) Indoor (0.67)
Natural (0.85) ManMade
Coastal (0.58) Mountainous (0.85)
Beach (0.58)

T1 step

Domain TBox

Natural $\equiv$ Outdoors $\sqcup \neg$ ManMade
Mountainous $\equiv$ Natural $\sqcup \neg$ Coastal
Beach $\equiv$ Coastal $\sqcap \exists$contains.Sand
$\exists$contains.Mountain $\sqsubseteq$ Mountainous
$\exists$contains.Sea $\sqsubseteq$ Coastal
$\exists$contains.Sand $\sqcap$ Mountainous $\sqsubseteq \perp$
Outdoor $\sqcap$ Indoor $\sqsubseteq \perp$

directly disjoint

inferred disjoint

(image:Indoor) $\geq$ 0.67
(image:$\exists$contains.Sea) $\geq$ 0.73
(image:$\exists$contains.Sand) $\geq$ 0.58
(image:Coastal) $\geq$ 0.73
(image:Beach) $\geq$ 0.58
(image:$\exists$contains.Mountain) $\geq$ 0.85
(image:Mountainous) $\geq$ 0.85
(image:Natural) $\geq$ 0.85
(image:Outdoor) $\geq$ 0.85

Disjoint axioms restored

Inconsistency handling

(image:$\exists$contains.Mountain)$\geq$ 0.85

Three Reasoning Tasks - III

- **T3 – Enrichment**
  - performs on the set of assertions maintained after step T2
  - through typical DLs reasoning inferred assertions are obtained, leading to enriched descriptions

Enrichment demonstration

Initial Assertions

(image:Indoor) $\geq$ 0.67
(image:$\exists$contains.Sea) $\geq$ 0.73
(image:$\exists$contains.Sand) $\geq$ 0.58
(image:$\exists$contains.Mountain) $\geq$ 0.85

Disjointness axioms removed

(image:Indoor) $\geq$ 0.67
(image:$\exists$contains.Sea) $\geq$ 0.73
(image:$\exists$contains.Sand) $\geq$ 0.58
(image:Coastal) $\geq$ 0.73
(image:Beach) $\geq$ 0.58
(image:Natural) $\geq$ 0.73
(image:Outdoor) $\geq$ 0.73
(image:$\exists$contains.Mountain) $\geq$ 0.85
(image:Mountainous) $\geq$ 0.85
(image:Natural) $\geq$ 0.85
(image:Outdoor) $\geq$ 0.85

Scene level hierarchy

Outdoor (0.85) Indoor (0.67)
Natural (0.85) ManMade
Coastal (0.58) Mountainous (0.85)
Beach (0.58)

T1 step

Domain TBox

$\text{Natural} \equiv \text{Outdoors} \sqcup \neg \text{ManMade}$
$\text{Mountainous} \equiv \text{Natural} \sqcup \neg \text{Coastal}$
$\text{Beach} \equiv \text{Coastal} \sqcap \exists\text{contains.Sand}$
$\exists\text{contains.Mountain} \sqsubseteq \text{Mountainous}$
$\exists\text{contains.Sea} \sqsubseteq \text{Coastal}$
$\exists\text{contains.Sand} \sqcap \text{Mountainous} \sqsubseteq \bot$
$\text{Outdoor} \sqcap \text{Indoor} \sqsubseteq \bot$

directly disjoint

inferred disjoint

(image:Indoor) $\geq$ 0.67
(image:$\exists$contains.Sea) $\geq$ 0.73
(image:$\exists$contains.Sand) $\geq$ 0.58
(image:Coastal) $\geq$ 0.73
(image:Beach) $\geq$ 0.58
(image:$\exists$contains.Mountain) $\geq$ 0.85
(image:Mountainous) $\geq$ 0.85
(image:Natural) $\geq$ 0.85
(image:Outdoor) $\geq$ 0.85

Disjoint axioms restored

Inconsistency handling

(image:$\exists$contains.Mountain) $\geq$ 0.85

T2 step

(image:$\exists$contains.Mountain) $\geq$ 0.85
(image:Mountainous) $\geq$ 0.85
(image:Natural) $\geq$ 0.85
(image:Outdoor) $\geq$ 0.85

Final Assertions

Experimental Results

- Domain of outdoor images (~360 images)
- Use of fuzzyDL(*) as inference engine for core fuzzy DLs reasoning services
- Experiment I
  - three implementations for scene level classifiers, two implementations for object level
- Experiment II
  - one implementation for scene level classifiers
  - one implementation for object level classifiers

(*) http://faure.isti.cnr.it/~straccia/software/fuzzyDL/fuzzyDL.html

Outdoor images TBox extract

$\text{Countryside\_buildings} \sqsubseteq \exists \text{contains.Buildings} \sqcap \exists \text{contains.Foliage}$
$\text{Countryside\_buildings} \sqsubseteq \text{Landscape}$
$\exists \text{contains.Forest} \sqcup \exists \text{contains.Grass} \sqcup \exists \text{contains.Tree} \sqsubseteq \exists \text{contains.Foliage}$
$\text{Rockyside} \sqsubseteq \exists \text{contains.Cliff}$
$\text{Rockyside} \sqsubseteq \exists \text{contains.Mountainous}$
$\text{Roadside} \sqsubseteq \exists \text{contains.Road}$
$\text{Roadside} \sqsubseteq \text{Landscape}$
$\exists \text{contains.Sea} \equiv \text{Coastal}$
$\text{Coastal} \sqsubseteq \text{Natural}$
$\exists \text{contains.Forest} \sqsubseteq \text{Landscape}$
$\text{Beach} \equiv \text{Coastal} \sqcap \exists \text{contains.Sand}$
$\text{Beach} \sqsubseteq \text{Natural}$
$\text{Cityscape} \sqsubseteq \text{ManMade}$
$\exists \text{contains.Sky} \sqsubseteq \text{Outdoor}$
$\exists \text{contains.Trunk} \sqsubseteq \exists \text{contains.Tree}$
$\text{Mountainous} \sqcap \text{Coastal} \sqsubseteq \bot$
$\text{Natural} \sqcap \text{ManMade} \sqsubseteq \bot$

Experiment I – Scene level concepts

	Analysis			Reasoning		
Concept	Recall	Precision	F-M	Recall	Precision	F-M
Indoor	0.00	NaN	NaN	1.00	0.75	0.85
Outdoor	0.99	0.99	0.99	0.99	0.99	0.99
Natural	0.97	0.96	0.97	0.98	0.96	0.97
ManMade	0.18	0.40	0.25	0.18	0.40	0.25
Cityscape	0.18	0.40	0.25	0.18	0.40	0.25
Landscape	0.75	0.63	0.68	0.76	0.68	0.71
Mountainous	0.64	0.28	0.39	0.48	0.30	0.37
Coastal	0.00	NaN	NaN	0.86	0.49	0.63
Beach	0.89	0.30	0.45	0.90	0.31	0.47

Analysis extracted descriptions are 'semantically treated', i.e. detection of Beach is considered as positive detection of Outdoor also. Not much impact because of low semantic association between object level and scene level concepts.

Experiment I – Object level concepts

	Analysis			Reasoning		
Concept	Recall	Precision	F-M	Recall	Precision	F-M
Building	1.00	0.17	0.29	0.09	0.83	0.17
Grass	0.06	0.40	0.10	0.01	1.00	0.03
Foliage	0.99	0.70	**0.82**	0.90	0.80	**0.85**
Sky	0.93	0.87	0.89	0.93	0.87	0.89
Cliff	0.98	0.21	**0.35**	0.54	0.42	**0.47**
Tree	0.22	0.65	0.33	0.18	0.58	0.27
Trunk	0.38	0.65	0.48	0.38	0.65	0.48
Sand	0.49	0.37	**0.42**	0.92	0.41	**0.56**
Sea	0.72	0.46	**0.56**	0.88	0.49	**0.63**
Conifers	1.00	0.01	**0.02**	0.50	0.02	**0.03**
Mountain	0.14	0.01	**0.01**	0.43	0.04	**0.06**
Boat	0.10	0.40	**0.16**	0.10	0.50	**0.17**
Road	0.15	0.50	0.23	0.02	0.25	0.03
Ground	0.06	0.57	0.19	0.11	0.57	0.19
Person	0.49	0.54	0.52	0.49	0.54	0.52

Concepts semantically related to scene level concepts are affected the most, e.g. the Sand concept. In general, precision is improved due to the utilisation of disjoint semantics.

Experiment II – Scene level concepts

	Analysis			Reasoning		
Concept	**Recall**	**Precision**	**F-M**	**Recall**	**Precision**	**F-M**
Countryside_buildings	0.30	1.0	**0.46**	0.60	0.86	**0.71**
Rockyside	0.68	0.70	**0.69**	0.68	0.79	**0.74**
Roadside	0.68	0.69	**0.69**	0.68	0.72	**0.70**
Forest	0.75	0.63	**0.69**	0.74	0.68	**0.71**
Coastal	0.85	0.67	**0.75**	0.86	0.72	**0.78**
Outdoor	-	-	-	0.00	1.00	0.99
Indoor	-	-	-	NaN	NaN	NaN
Natural	-	-	-	0.97	1.00	0.98
ManMade	-	-	-	NaN	NaN	NaN
Cityscape	-	-	-	NaN	NaN	NaN
Mountainous	-	-	-	0.67	0.80	0.74
Beach	-	-	-	0.45	0.76	0.57

Higher impact as the analysis supported concepts are characterised are more strongly related to each other.

Experiment II – Object level concepts

	Analysis			Reasoning		
Concept	Recall	Precision	F-M	Recall	Precision	F-M
Building	0.54	0.69	**0.60**	0.62	0.86	**0.72**
Roof	0.33	0.54	**0.41**	0.33	0.75	**0.46**
Grass	0.49	0.42	0.45	0.30	0.52	0.38
Foliage	0.48	0.84	**0.61**	0.86	0.86	**0.86**
Dried-Plant	0.07	0.11	**0.08**	0.07	0.13	**0.10**
Ground	0.26	0.33	0.29	0.26	0.33	0.29
Person	0.75	0.51	0.61	0.75	0.51	0.61
Sky	0.95	0.93	0.94	0.95	0.93	0.94
Cliff	0.65	0.45	**0.53**	0.69	0.70	**0.69**
Tree	0.49	0.52	0.51	0.56	0.47	0.51
Trunk	0.26	0.28	0.27	0.26	0.28	0.27
Sand	0.02	0.10	**0.03**	0.57	0.45	**0.50**
Sea	0.69	0.60	**0.64**	0.85	0.69	**0.76**
Wave	0.25	0.5	0.33	0.25	0.5	0.33
Boat	0.41	0.71	0.52	0.33	0.66	0.44
Road	0.50	0.69	**0.58**	0.69	0.71	**0.70**

Again, higher impact as the analysis supported concepts bear stronger semantic relatedness.

Interesting to note the lower performance for Boat, which is due to analysis mistaken degrees estimation of the scene level concepts (Cityscape appears prevalent, which is disjoint with Boat).

Cliff detector has better performance than the corresponding Rockyside scene level one; replacing though Rockyside µ ∃contains.Cliff with ∃contains.Cliff µ Rockyside would be a customisation of domain knowledge, not generally applicable.

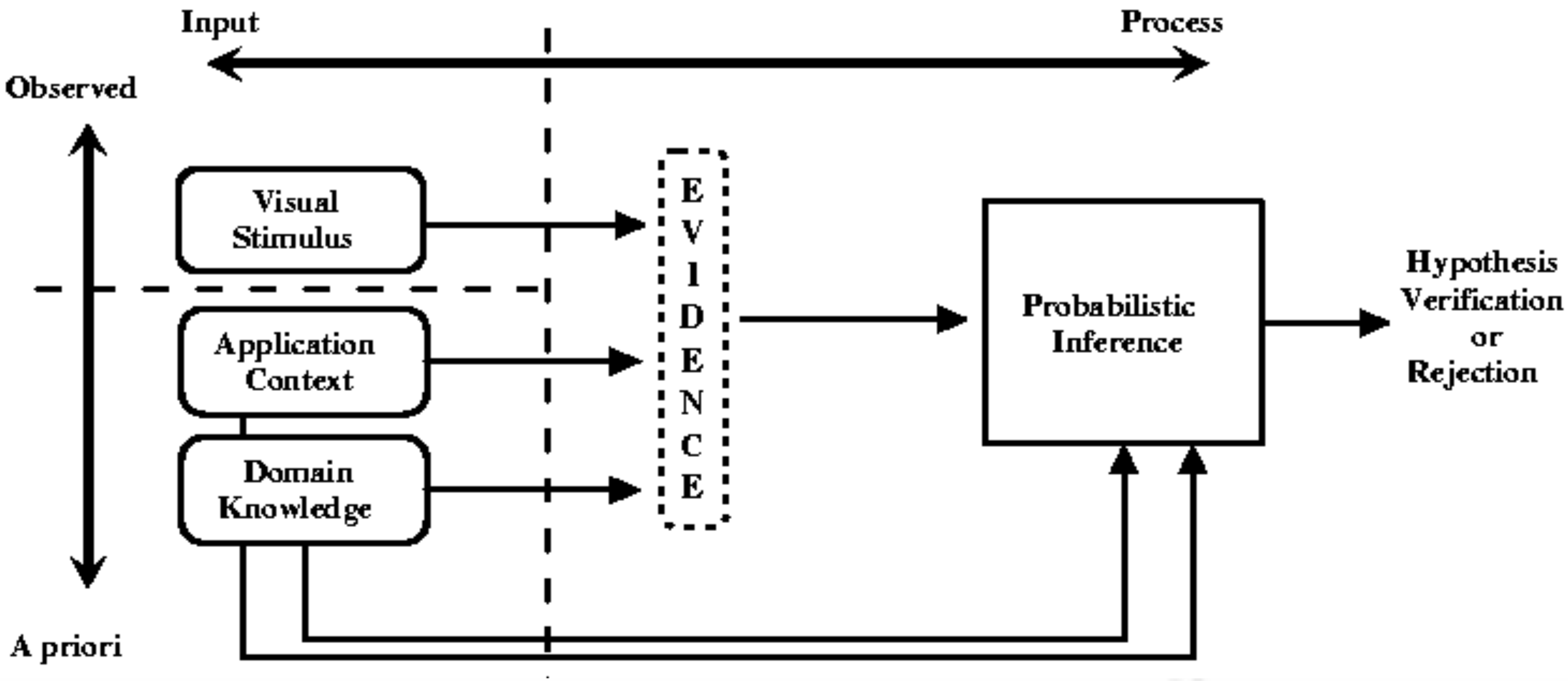
Conclusions

- The proposed fuzzy DLs reasoning enables
  - formal handling of the imprecision inherent in the input classifications
  - utilisation of domain semantics
  - consistent interpretations / image descriptions
- The use of explicit semantics and logic-based reasoning is crucial in multimedia semantics extraction
  - yet not the only necessary component
- Largely miscalculated classification degrees can mislead the interpretation
  - combined usage of additional (probabilistic) knowledge could be a possible solution, along with logic-based reasoning to account for possible worlds/interpretations

Probabilistic inference

Our approach

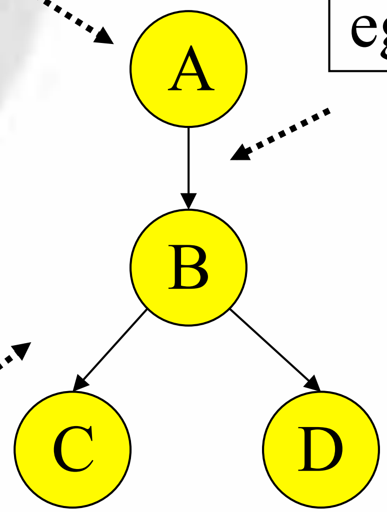
Goal: Combine explicit (provided by humans) and implicit (extracted from training data) knowledge for enhancing image analysis



BN: A Directed Acyclic Graph

Each node in the graph is a random variable

A node X is a parent of another node Y if there is an arrow from node X to node Y eg. A is a parent of B



Informally, an arrow from node X to node Y means X has a direct influence on Y

A Set of Tables for Each Node

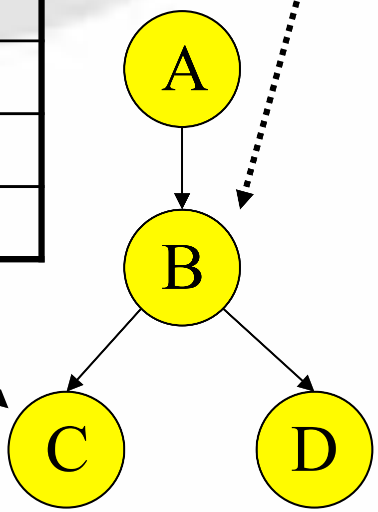
A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

Each node X_i has a conditional probability distribution $P(X_i \mid Parents(X_i))$ that quantifies the effect of the parents on the node

The parameters are the probabilities in these conditional probability tables (CPTs)

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1



B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

Bayesian Networks

Two important properties:

1. Encodes the conditional independence relationships between the variables in the graph structure
2. Is a compact representation of the joint probability distribution over the variables

Inference

- Using a Bayesian network to compute probabilities is called inference
- In general, inference involves queries of the form:

P(X | E)

E = The evidence variable(s)

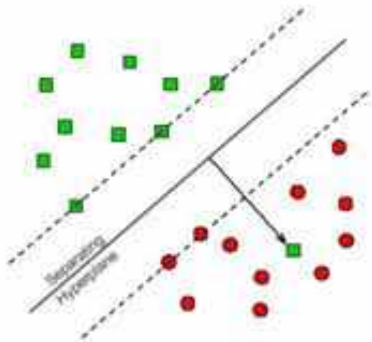
X = The query variable(s)

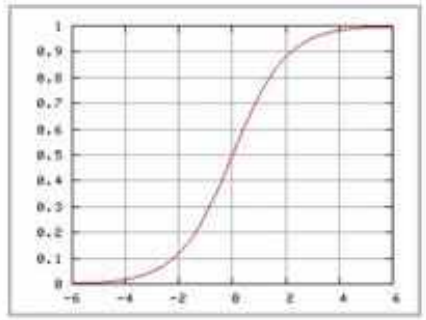
Designing, training and performing inference on the BN

BN	Our approach
Network Structure • nodes • arcs	Ontology • concepts • ontology relations
Network Parameters • Conditional Probability Tables • Prior Probabilities	Annotated data (context) • frequency of co-occurrence between concepts • frequency of concepts' appearance
Evidence • Probabilities for the evidence variables	Visual stimulus • Probabilistic output of the SVM-based classifiers

Framework Components

- Visual stimulus
  - Segmentation using a Recursive Shortest Spanning Tree algorithm [Adamek & O'Connor, 2005]
  - MPEG-7 visual descriptors (region and global level)
  - SVM-based concept classifiers producing probability estimates by fitting the decision values by a sigmoid function





- Domain knowledge
  - Ontology (OWL-DL)
- Application context
  - Quantifies the effect/causality between concepts using co-occurrence information
- Probabilistic inference
  - Bayesian Networks and message passing belief propagation

Image Analysis - Running Example

SVM-based Concept Classifiers (global level)

SVM-based Concept Classifiers (region level)

	Global Classifiers
Tennis	45,97
Roadside	54,21
Rockyside	47,07
Seaside	**56,31**
Forest	52,48
C.Buildings	50,00

	Local Classifiers			
	Region1	Region2	Region3	Region4
Board	45,51	49,51	50,62	45,84
Court	49,72	52,53	51,83	54,00
Person	52,56	50,16	51,54	52,93
Graffiti	49,07	53,01	51,40	51,67
Court line	50,13	49,25	51,85	51,93
Racket	48,95	50,99	50,28	48,34
Sky	47,05	**60,69**	**74,54**	**56,03**
Dried plant	47,14	49,63	48,09	47,97
Grass	**56,36**	53,29	47,66	50,31
Road	52,28	48,47	49,45	54,97
Vegetation	48,17	47,18	46,85	53,74
Ground	50,58	48,56	50,70	49,51
Road line	48,55	48,93	49,89	50,30
Car	47,87	49,51	47,60	47,92
Trunk	48,82	47,68	49,00	48,41
Rock	48,13	47,40	47,79	48,00
Glacier	50,00	50,00	50,00	50,00
Sea	48,73	51,32	47,87	44,08
Sand	49,62	47,86	49,66	47,41
Wave	52,85	46,54	47,98	48,29
Boat	49,78	49,74	47,63	48,77
Tree	50,47	47,61	48,41	48,25
Stone	49,58	49,50	49,39	49,55
Building	44,62	47,71	46,14	43,97
Roof	48,60	52,30	49,47	49,90

	Belief Evolution					
%	Tennis	Roadside	Rockyside	Seaside	Forest	C.Buildings
Prior	15,4	17,6	15,7	17,3	**18,5**	15,4
Global	17,4	20,2	14,2	**21,2**	20,1	18,8
Evidence Region1	13,4	**21,7**	14,2	21,2	20,9	20,6
Evidence Region2	13,4	**23,5**	15,2	22,9	21,5	22,6
Evidence Region3	13,4	**27,2**	17,7	26,3	22,6	26,8
Evidence Region4	13,4	**27,2**	17,6	26,9	22,8	27,4

Ontology-to-BN – Network Structure *(Explicit Knowledge)*

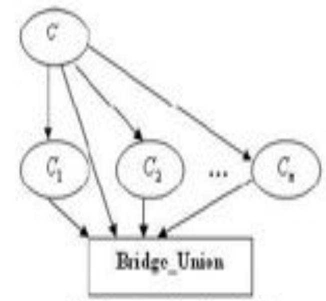
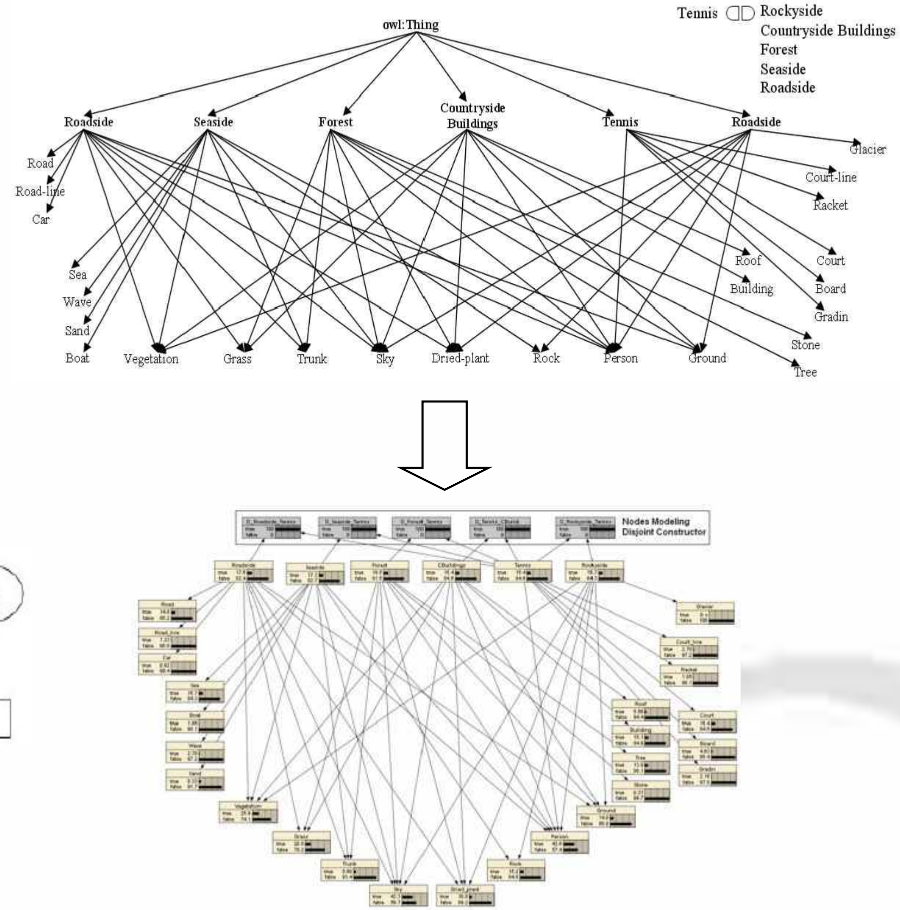


Fig.3. – "owl:unionOf"



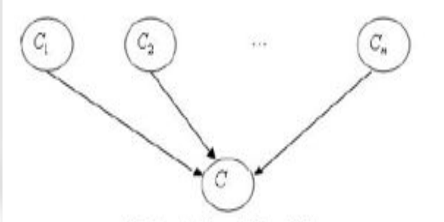


Fig.1. – "rdfs:subClassOf"

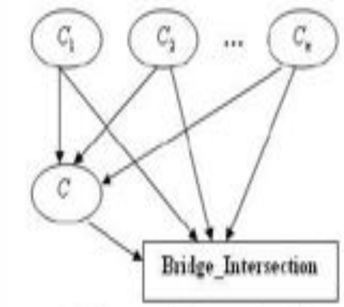
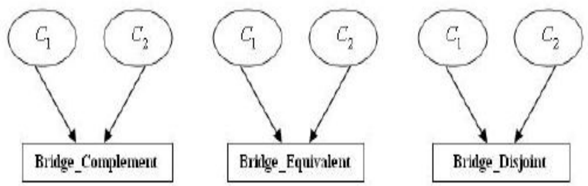


Fig.2. – "owl:intersectionOf"



Z. Ding, Y. Peng, and R. Pan, "A bayesian approach to uncertainty modeling in owl ontology," in Proc. of the Int. Conf. on Advances in Intelligent Systems - Theory and Applications, Nov. 2004

Ontology-to-BN - Parameter Learning *(Implicit Knowledge)*

Concept labels

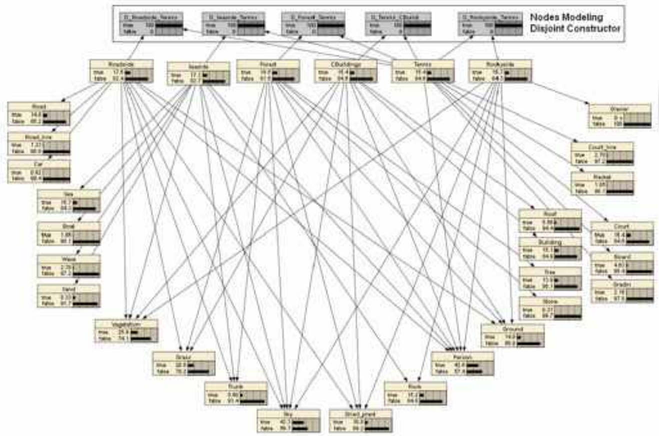
// ~->[CASE-1]->~

IDnum	Building		Roof	Grass	Vegetation		Dried_plant		Ground	Person	Sky	Rock	Glacier	Tree	Trunk	Stone	Sand	Sea
0	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
9	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
12	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
13	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
14	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
15	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
16	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
17	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
18	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
19	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
20	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

- Concept labels were obtained by manual annotation
- Concepts co-occurrence was used to estimate causality relations

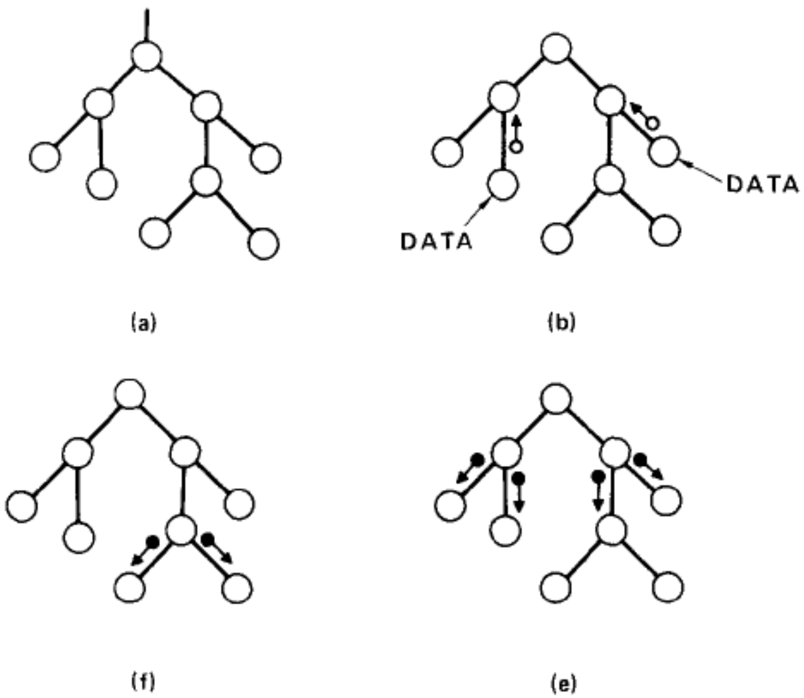
- Expectation Maximization was applied on concept labels
- CPTs were estimated between the connected nodes

BN with estimated CPTs



Belief Propagation with Message passing

- Pearl's message passing algorithm for belief propagation
    - Top-down and bottom-up message passing between parent and child nodes
- Junction tree variation for coping with complexity issues



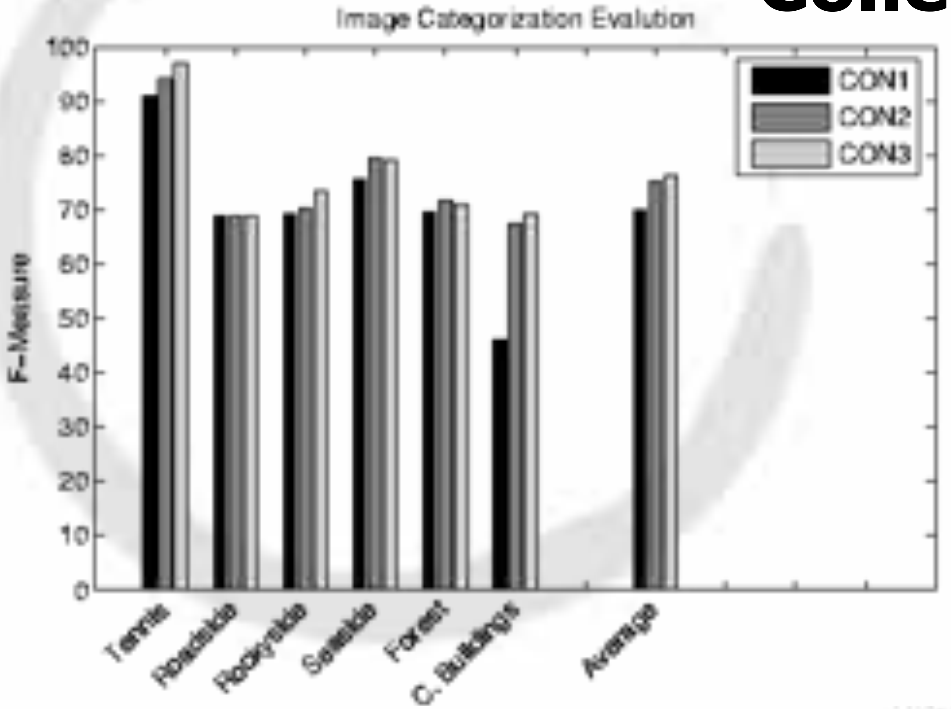
- *Pearl, "Fusion, propagation, and structuring in belief networks," Artif. Intell., vol. 29, no. 3, pp. 241–288, 1986.*

- *F. V. Jensen and F. Jensen, "Optimal junction trees," in Proc. of the 10th Conf. on Uncertainty in Artif. Intel., C. M. Kaufmann, Ed., San Mateo, 1994*

Experimental Study

- Image analysis Tasks
  - **Image categorization:** Select the category concept that best describes an image as a whole.
  - **Localized region labeling**: assign labels to pre-segmented image regions, with one of the available regional concepts
  - **Weak annotation of video shot key-frames:** associate multiple concepts to an image, but not with specific image regions
- Test-beds
  - **Personal Collection**, 648 images, 6 global concepts, 25 local concepts
  - **MSRC**, 591 images, 21 local concepts
  - **TRECVID 2005**, 61600 images, 374 global concepts

Image Categorization (Personal Collection)



- CON1: Baseline configuration that is based solely on visual stimulus
- CON2: Concept hierarchy information is used but no semantic constraints are taken into account
- CON3: Semantic constraints are also taken into account

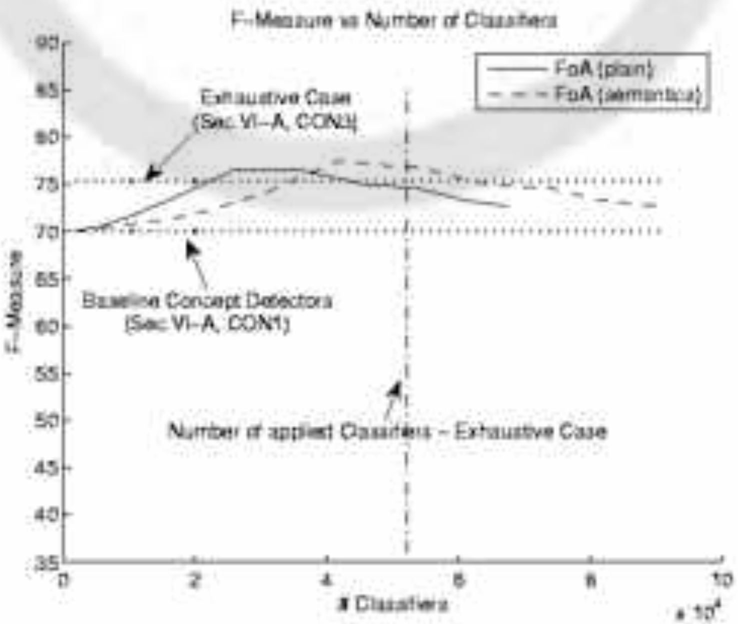
TABLE III
CONFUSION MATRIX FOR IMAGE CATEGORIZATION - CON2 LOWER OF THE CELLS - CON3 UPPER OF THE CELLS

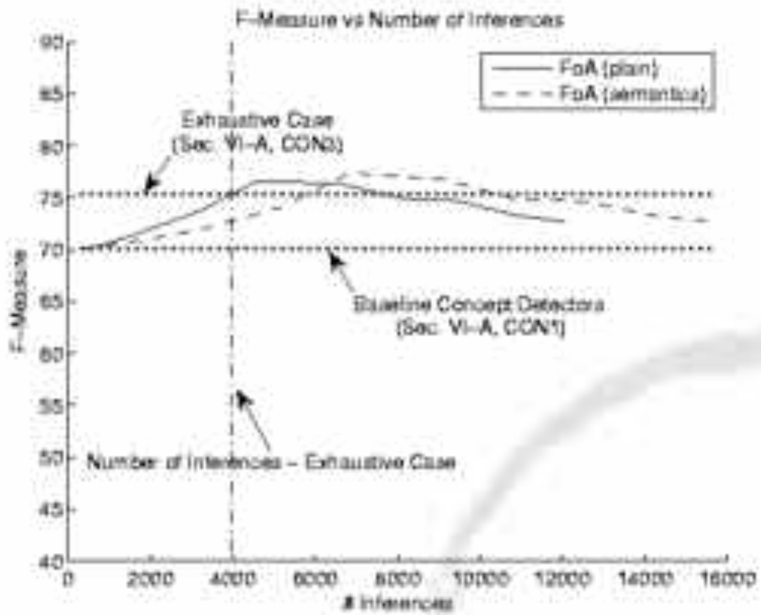
%	Tennis	Roadside	Rockyside	Seaside	Forest	C. Buildings
Tennis	94.00 / 88.00	0.00 / 0.00	2.00 / 0.00	4.00 / 2.00	0.00 / 0.00	0.00 / 0.00
Roadside	3.50 / 1.75	73.68 / 73.68	0.00 / 0.00	8.77 / 8.77	11.28 / 10.53	3.26 / 3.26
Rockyside	0.00 / 3.88	3.92 / 3.92	70.78 / 64.71	3.88 / 3.88	19.61 / 19.61	0.00 / 0.00
Seaside	0.00 / 0.00	3.56 / 5.36	3.57 / 3.57	91.07 / 83.93	0.00 / 2.00	0.00 / 0.00
Forest	0.00 / 0.00	10.00 / 10.00	8.33 / 8.33	50.00 / 10.00	71.67 / 71.67	0.00 / 0.00
C. Buildings	0.00 / 2.00	24.00 / 24.00	4.00 / 6.00	12.00 / 12.00	2.00 / 2.00	58.00 / 54.00

Image Categorization using a Focus of Attention (FoA) mechanism

-Start from the most likely hypothesis and attempt to verify it by looking for evidence that would have normally been present if the hypothesis was true.

- Exploit the mutual information (learned from training data) between the hypothesis and evidence concepts.

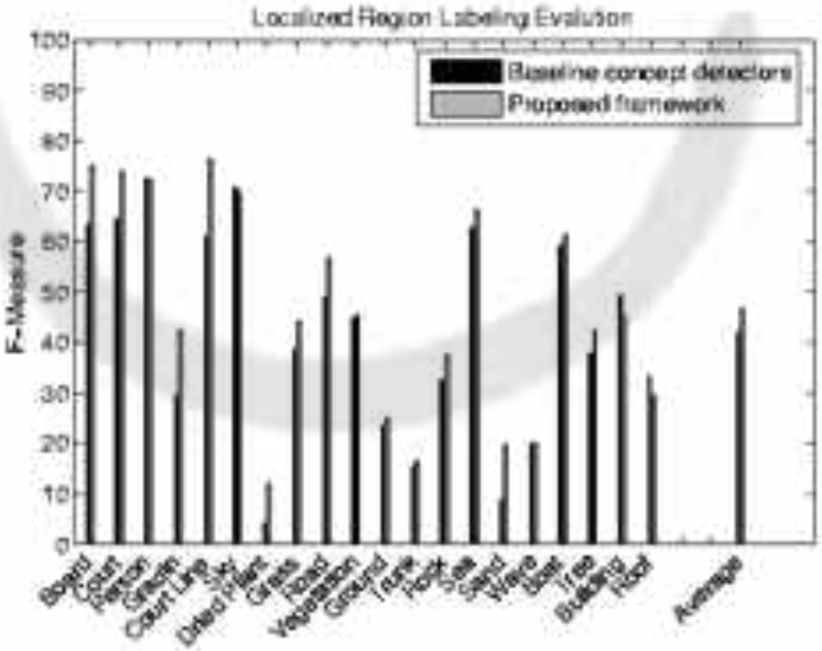




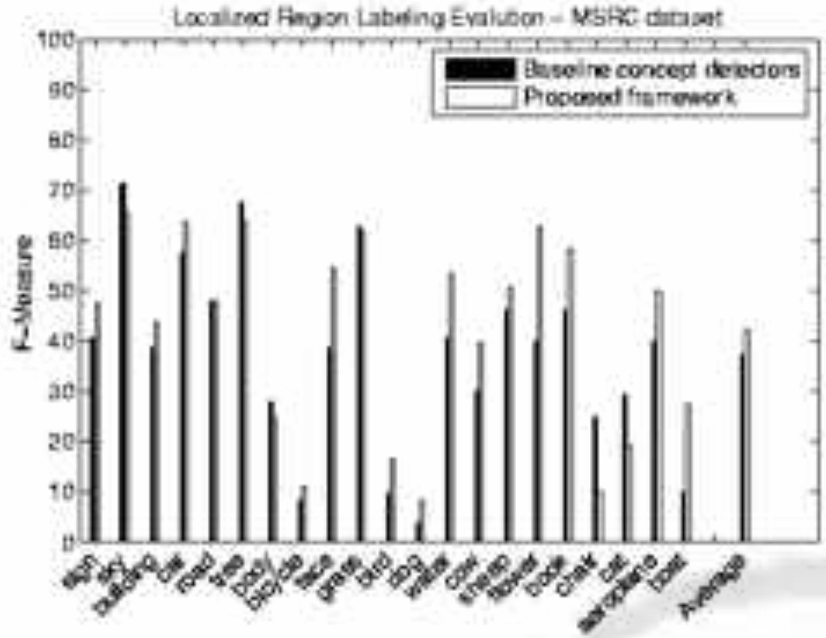
3172 (sec) gain in time

Localized region labeling

Personal Collection Dataset

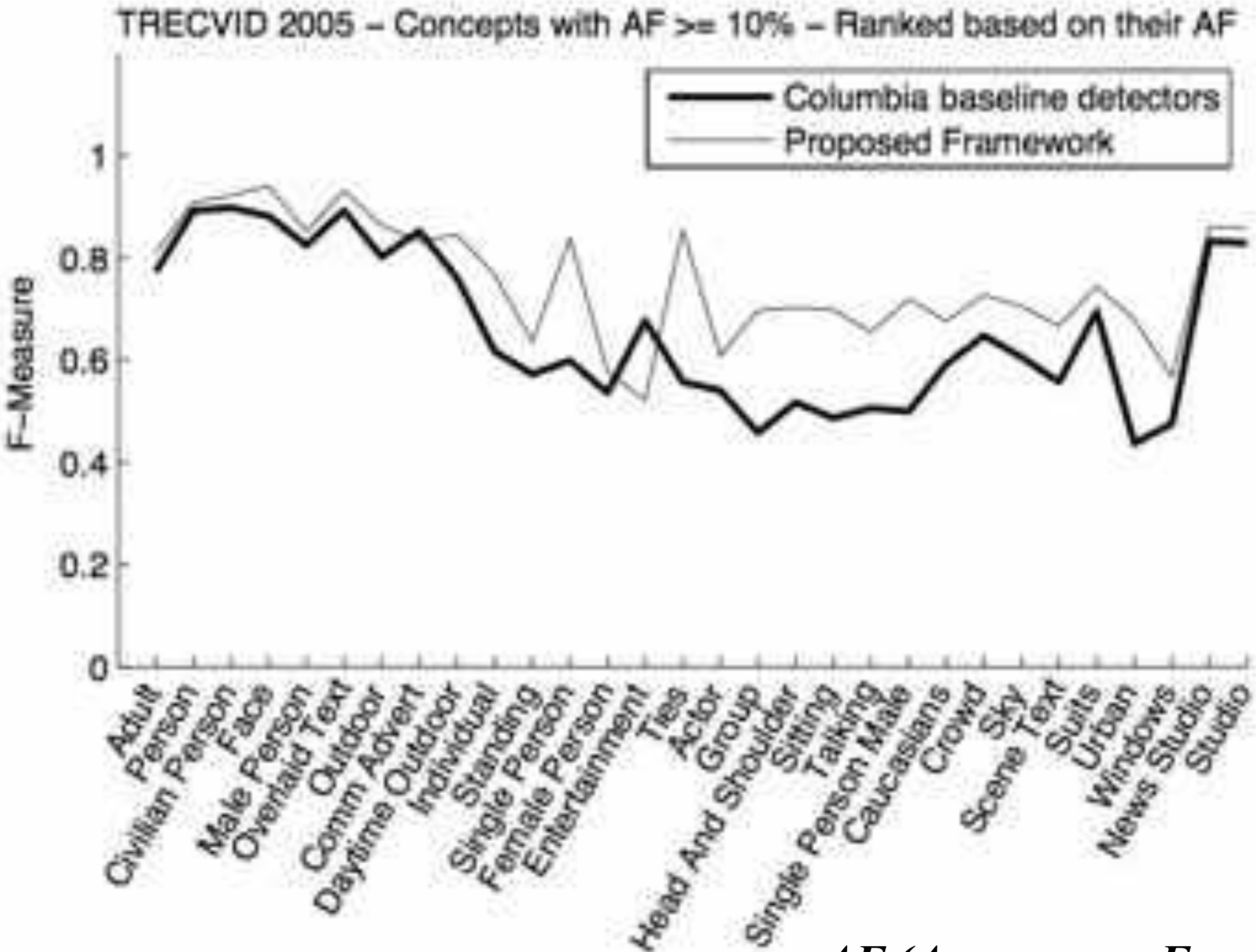


MSRC Dataset



When there is a conflict between the prediction of global and local classifiers we make two different hypothesis, we evaluate them into the BN and eventually select the one with highest probability.

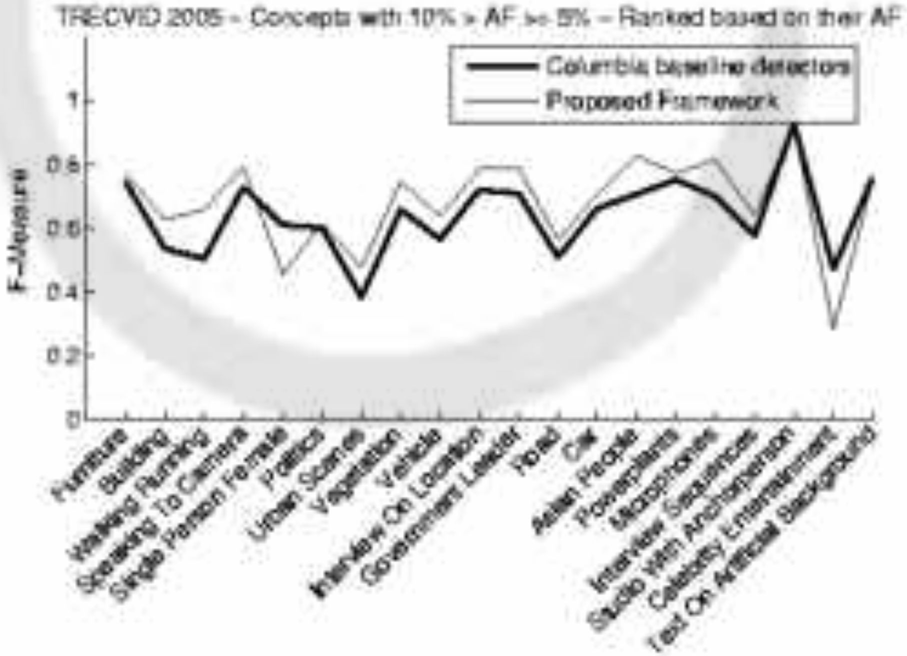
Weak annotation of video shot key-frames (1/2)



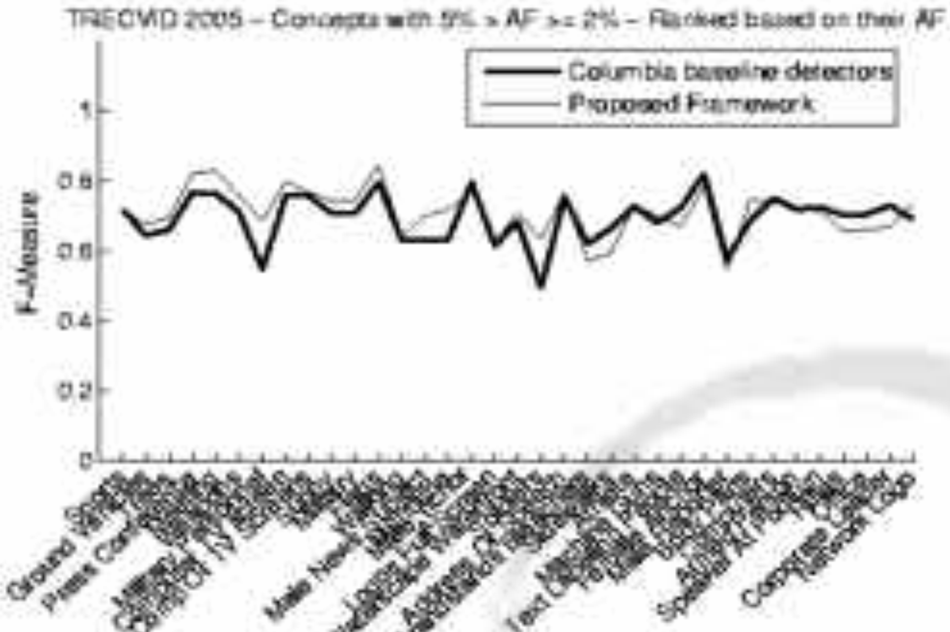
AF (Appearance Frequency) > 10%

Weak annotation of video shot key-frames (2/2)

10% > AF (Appearance Frequency) > 5%



5% > AF (Appearance Frequency) > 2%



Comparing with existing methods

COMPARING WITH EXISTING METHODS IN OBJECT RECOGNITION

	Buildings	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat
Textonboost [18]	**62**	**98**	86	58	50	83	60	53	**74**	63	**75**	63	35	**19**	**92**	15	86	**54**	19	**62**	7
PLSA-MRF/P [17]	52	87	68	**73**	**84**	94	**88**	**73**	70	68	74	**89**	33	19	78	**34**	**89**	46	**49**	54	**31**
Prop. Framework	32	55	**87**	40	73	**96**	57	56	50	**76**	8	64	**38**	12	46	5	51	12	8	29	18

- None of the three systems manages to outperform the others for a significant portion of the 21 classes
- Error rates are often quite different on individual classes showing that while there are some classes that can be modeled very efficiently using the visual features and the model proposed by one method, there are other classes that are best modeled using a different set of visual features and model
- Our work focus on using context and knowledge for improving the performance of a set of baseline concept classifiers, not to discover the optimal feature space

[17] J. J. Verbeek and B. Triggs, "Region classification with markov field aspect models," in CVPR, 2007

[18] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in ECCV (1), 2006, pp. 1–15.

Conclusions

- Combining explicit & implicit knowledge is beneficiary for enhancing image analysis …
  - Hierarchy information and causality relations between domain concepts was found to be useful in most of the cases
- The value of semantic information depends largely on the special characteristics of the domain …
  - Semantic constraints were only able to help image interpretation, when the imposed rules could be directly reflected into the visual space and when the domain is rich enough to impose meaningful interconnections
- A large amount of training data is required for approximating the prior and conditional probabilities using frequency information …
  - No improvement was delivered for the rarely appearing concepts of TRECVID dataset

Common (Open) Issues

- Evaluation
- Annotated content
- Ontologies
- Fusion in analysis
- Uncertainty in reasoning
- Large-Scale
- Generic vs. Specific approaches
- Multiple domains support

Conclusions

- Semantic analysis of multimedia is already providing results
- Fundamental and applied research in
  - Logic-based + signal approaches
  - Implicit + explicit (knowledge) approaches
  - Logic + statistical/learning based
- Different applications and requirements
- Ongoing research in all areas

Thank you!

http://mklab.iti.gr