Combining visual and knowledge processing for semantics extraction

Dr. Yiannis Kompatsiaris
Multimedia Knowledge Laboratory
CERTH - Informatics and Telematics Institute

http://mklab.iti.gr
Outline

• Introduction
• Content – Applications
• Problem Definition
• Context and Reasoning
• Combined Approaches
  • Visual + Context
  • Visual + fuzzy DL Reasoning
  • Visual + Rules
• Conclusions
Evolution of Content ...

- 3 billions photos in Flickr and 2 billions a year ago
- 3,024,780,142 photos @ 11:52, 12 Nov 2008
- 1-2 exabytes (millions of terabytes) of new information produced world-wide annually
- 80 billion of digital images are captured each year
- Over 1 billion images related to commercial transactions are available through the Internet
- This number is estimated to increase by ten times in the next two years.
- 4 000 new films are produced each year
- 33 000 television stations
- 100 billions of hours of audiovisual content
...results to be LOST in content
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”
Content - Applications

Content

Knowledge Extraction

Applications

Personal

Sports - News

Industrial

3D

News

Commercial

Semantic Desktop

Personalization

Mobile

Retrieval

News

Web 2.0

Commercial

Semantic Desktop

Personalization

Mobile

Retrieval
Addressing the **Semantic Gap**

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

```xml
<?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>
```

Dominant Color Descriptor of a *sky* region

This image contains a *sky* region and is a *holiday* image
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
Visual features based Classification

Segment’s hypothesis set

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

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
### Semantics goes beyond perceptual manifestations

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

*Discrepancy between intended and learned semantics*

*Snoek et al., “Adding Semantics to Detectors for Video Retrieval”, IEEE Multimedia, 2007*
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

KAA

scene classification

beach scene

person/face detection

other analysis methods
Knowledge Extraction
A common view

- Implicit Knowledge
  - Signal Level
- Explicit Knowledge – Logic - Semantics & Hybrid Level

- Single Modality Analysis
- Manual Annotation - Models
- Additional Analysis Information
- Knowledge Infrastructure (Multimedia Ontology)
- Semantic Analysis

Manual Annotation - Models
**Knowledge Extraction**

**A common view**

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

**Information**
- Classifiers fusion
  - Global vs. Local
  - Modalities fusion
- Context
  - "Ambulance"

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

**Manual**
- Multimedia content annotation tools
- Training (Statistical)
- Modeling

**Knowledge Infrastructure (Multim)**
- Domain
  - Multimedia content
  - Annotations
- Algorithms - Features
  - Context
Multimedia Content Annotation

M-Ontomat-Annotizer and VAT
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)
Visual information processing and context usage

Visual information processing

• Segmentation
  • Using extension of the Recursive Shortest Spanning Tree (RSST) algorithm
    • Fast
    • Produces accurate region boundaries

• Region-level features extraction
  • MPEG-7 descriptors
    • Scalable Color, Edge Histogram, Homogeneous Texture, Region Shape, Color Structure and Color Layout
    • More descriptors could be employed to improve performance

• Common data set for all classification algorithms
Region classification

- **Aim:** perform the mapping of the low-level numerical image data to high-level semantic concepts

- **Semantic concepts**
  - Denote real-world objects that can be present in the examined images
  - Their detection can facilitate image manipulation tasks based on semantic criteria
  - Predefined by expert
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
Spatial Context - Genetic algorithm

• Extensively used in a wide variety of optimization problems

• Proposed approach
  • Employed on top of the SVM-based classifier
  • Treats semantic image analysis as a global optimization problem
  • Takes into account spatial information
    • Fuzzy directional relations
    • Spatial-related contextual information, acquired by training
  • Fitness function

\[ f(V) = \lambda \times FS_{norm} + (1 - \lambda) \times SC_{norm} \]

Visual features similarity
Spatial relations consistency

• Parameter \( \lambda \) estimated according to a separate optimization procedure
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
Fuzzy Spatial Relations Extraction

- Compute a “reduced box” of the ground object’s MBR in terms of its compactness
- Form cone-shaped areas and assign the corresponding relations
- The percentage of the figure object included in each of the cone-shaped areas determines the degree to which the respective relation is satisfied

*Figure object:* the object whose relative position is to be estimated

*Ground object:* the object used as reference
Genetic algorithm demonstration

Initial image

Segmentation Mask
<table>
<thead>
<tr>
<th>Region</th>
<th>Sea</th>
<th>Sky</th>
<th>Sand</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region 1</td>
<td>0.05</td>
<td>0.03</td>
<td>0.07</td>
<td>1.00</td>
</tr>
<tr>
<td>Region 2</td>
<td>0.28</td>
<td>0.42</td>
<td>0.30</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 3</td>
<td>0.54</td>
<td>0.74</td>
<td>0.32</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 4</td>
<td>0.79</td>
<td>0.54</td>
<td>0.43</td>
<td>0.08</td>
</tr>
<tr>
<td>Region 5</td>
<td>0.00</td>
<td>0.80</td>
<td>0.03</td>
<td>0.09</td>
</tr>
<tr>
<td>Region</td>
<td>Sea</td>
<td>Sky</td>
<td>Sand</td>
<td>Person</td>
</tr>
<tr>
<td>---------</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>Region 1</td>
<td>0.05</td>
<td>0.03</td>
<td>0.07</td>
<td>1.00</td>
</tr>
<tr>
<td>Region 2</td>
<td>0.28</td>
<td>0.42</td>
<td>0.30</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 3</td>
<td>0.54</td>
<td>0.74</td>
<td>0.32</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 4</td>
<td>0.79</td>
<td>0.54</td>
<td>0.43</td>
<td>0.08</td>
</tr>
<tr>
<td>Region 5</td>
<td>0.00</td>
<td>0.80</td>
<td>0.03</td>
<td>0.09</td>
</tr>
</tbody>
</table>
### Spatial Relations

<table>
<thead>
<tr>
<th>Region</th>
<th>Sea</th>
<th>Sky</th>
<th>Sand</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region 1</td>
<td>0.05</td>
<td>0.03</td>
<td>0.07</td>
<td>1.00</td>
</tr>
<tr>
<td>Region 2</td>
<td>0.28</td>
<td>0.42</td>
<td>0.30</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 3</td>
<td>0.54</td>
<td>0.74</td>
<td>0.32</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 4</td>
<td>0.79</td>
<td>0.54</td>
<td>0.43</td>
<td>0.08</td>
</tr>
<tr>
<td>Region 5</td>
<td>0.00</td>
<td>0.80</td>
<td>0.03</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Sky cannot be "Left" of Sea.
Sky cannot be “Left” of Sea

<table>
<thead>
<tr>
<th>Region</th>
<th>Sea</th>
<th>Sky</th>
<th>Sand</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region 1</td>
<td>0.05</td>
<td>0.03</td>
<td>0.07</td>
<td>1.00</td>
</tr>
<tr>
<td>Region 2</td>
<td>0.28</td>
<td>0.42</td>
<td>0.30</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 3</td>
<td>0.54</td>
<td>0.74</td>
<td>0.32</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 4</td>
<td>0.79</td>
<td>0.54</td>
<td>0.43</td>
<td>0.08</td>
</tr>
<tr>
<td>Region 5</td>
<td>0.00</td>
<td>0.80</td>
<td>0.03</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Spatial Relations

Confidence Values

Multimedia Knowledge Laboratory
Informatics and Telematics Institute
Sky cannot be "Below" Sea

<table>
<thead>
<tr>
<th>Region</th>
<th>Sea</th>
<th>Sky</th>
<th>Sand</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region 1</td>
<td>0.05</td>
<td>0.03</td>
<td>0.07</td>
<td>1.00</td>
</tr>
<tr>
<td>Region 2</td>
<td>0.28</td>
<td>0.42</td>
<td>0.30</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 3</td>
<td>0.54</td>
<td>0.74</td>
<td>0.32</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 4</td>
<td>0.79</td>
<td>0.54</td>
<td>0.43</td>
<td>0.08</td>
</tr>
<tr>
<td>Region 5</td>
<td>0.00</td>
<td>0.80</td>
<td>0.03</td>
<td>0.09</td>
</tr>
<tr>
<td>Region</td>
<td>Sea</td>
<td>Sky</td>
<td>Sand</td>
<td>Person</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>Region 1</td>
<td>0.05</td>
<td>0.03</td>
<td>0.07</td>
<td>1.00</td>
</tr>
<tr>
<td>Region 2</td>
<td>0.28</td>
<td>0.42</td>
<td>0.30</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 3</td>
<td>0.54</td>
<td>0.74</td>
<td>0.32</td>
<td>0.00</td>
</tr>
<tr>
<td>Region 4</td>
<td>0.79</td>
<td>0.54</td>
<td>0.43</td>
<td>0.08</td>
</tr>
<tr>
<td>Region 5</td>
<td>0.00</td>
<td>0.80</td>
<td>0.03</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Sky cannot be “Below” Sea
ML Algorithms Evaluation framework

• **Aim of framework**
  - In depth investigation of the behavior and the resulting performance of the developed algorithms
  - Extensive experimentations under varying experimental conditions

• **Concepts defined**
  - Sand, Sea, Vegetation, Person, Sky, Rock, Tree, Grass, Ground, Trunk, Wave, Boat, Dried-Plant, Building and Pavement
  - Listed in descending order with respect to their frequency of appearance in the assembled image set
Evaluation framework (cont’d)

• Experiments performed
  • Examine the generalization ability and the algorithms' behavior when variable amount of data is available for training purposes
    • Percentage of total images used for training
      • 10%
      • 30%
      • 50%
  • Investigate the classification performance of each method under varying problem complexity
    • Number of supported concepts considered
      • 5
      • 10
      • 15
    • Concepts selected based on their estimated frequency of appearance
Experimental results

- Domain of experimentation: Vacation images
- Assembled image set
  - 500 images
  - Obtained from the Flickr online photo management and sharing application
  - Depicts cityscape, seaside, mountain and landscape locations
  - Each image was segmented and manually annotated
Indicative region hypothesis set

Region’s hypothesis set (SVM-based classifier)

<table>
<thead>
<tr>
<th>Object</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sand</td>
<td>0.44</td>
</tr>
<tr>
<td>Vegetation</td>
<td>0.42</td>
</tr>
<tr>
<td>Sky</td>
<td>0.19</td>
</tr>
<tr>
<td>Tree</td>
<td>0.07</td>
</tr>
<tr>
<td>Ground</td>
<td>0.15</td>
</tr>
<tr>
<td>Wave</td>
<td>0.12</td>
</tr>
<tr>
<td>Dried-plant</td>
<td>0.12</td>
</tr>
<tr>
<td>Building</td>
<td>0.89</td>
</tr>
<tr>
<td>Pavement</td>
<td>0.29</td>
</tr>
<tr>
<td>Sea</td>
<td>0.13</td>
</tr>
<tr>
<td>Person</td>
<td>0.34</td>
</tr>
<tr>
<td>Rock</td>
<td>0.22</td>
</tr>
<tr>
<td>Grass</td>
<td>0.21</td>
</tr>
<tr>
<td>Trunk</td>
<td>0.31</td>
</tr>
<tr>
<td>Boat</td>
<td>0.11</td>
</tr>
</tbody>
</table>
Indicative region-concept association results

Input image

Region-concept association

GA

PSO

SVM
Concept detection accuracy

10% of images used for training

30% of images used for training

50% of images used for training
Beach domain region-classification results

- Supported concepts: *Sea, Sky, Sand, Person*
- Extracted MPEG-7 descriptors: *Scalable Color, Homogeneous Texture, Region Shape, Edge Histogram*
- 40 training images
- 400 testing images

<table>
<thead>
<tr>
<th>Object</th>
<th>SVM</th>
<th>GA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Sky</td>
<td>58.73%</td>
<td>92.50%</td>
</tr>
<tr>
<td>Sea</td>
<td>89.47%</td>
<td>53.13%</td>
</tr>
<tr>
<td>Sand</td>
<td>76.79%</td>
<td>97.73%</td>
</tr>
<tr>
<td>Person</td>
<td>72.34%</td>
<td>82.92%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>75.95%</td>
<td></td>
</tr>
</tbody>
</table>
Outdoor images domain region-classification results

- Supported concepts: Sky, Water, Ground, Building, Vegetation, Rock
- Extracted MPEG-7 descriptors: Scalable Color, Homogeneous Texture, Region Shape, Edge Histogram
- 200 training images
- 400 testing images

![Graph showing classification results]

<table>
<thead>
<tr>
<th>Technique</th>
<th>Total Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>71.39%</td>
</tr>
<tr>
<td>GA (Triangular)</td>
<td>72.22%</td>
</tr>
<tr>
<td>GA (Euclidean)</td>
<td>75.83%</td>
</tr>
</tbody>
</table>
Outdoor images domain region-classification results (con’d)
Experimental results (cont’d)

• Observations and conclusions
  • The combined use of an optimization method (GA, PSO) with a more traditional classifier (SVM) leads to increased classification accuracy
  • The use of an increased number of images for training purposes is beneficial
    • Highlights the need for the availability of large annotated media sets for appropriately training any classification method
  • The PSO classification scheme is shown to be particularly suitable in the presence of an extremely limited training set
  • The GA classifier is advantageous when more samples are available for training purposes
  • The performance of the employed methods degrades gracefully when the number of concepts increases
Reasoning in multimedia content analysis

- Support for imprecision - uncertainty
- Formal approaches
  - fuzzy DLs, probabilistic DLs, possibilistix DLs, fuzzy logic programming, abductive reasoning, inductive reasoning...
- Statistical approaches
  - Bayesian inference, HMMs...
- Used for:
  - Fusion / Integration
  - Consistency checking
  - Higher-level 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 versatile 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** final description
Semantics goes beyond perceptual manifestations

- Conifers detector semantics pertain to mountainous scenes
- Landscape detector pertains to field-like scenes
- Sand detector semantics pertains to beach scenes

- Sea and Sand detectors entail Beach scene
- Beach scenes entails Natural and Outdoor scenes
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 uncertainty of initial annotations
General Architecture

- Image
  - Segment Classification
  - Face Detection and Recognition
  - Person Detection
  - Scene Classification

- Textual annotation
  - Natural Language Processing

- KAA&TC ABoxes
- FDR ABox
- PD ABox
- ScC ABox
- NLP ABox

- Multimedia reasoning
- Final ABox
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

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

Examined Image

Machine Learning descriptions extraction

Scene level descriptions

Initial Assertions

Fuzzy DLs based reasoning framework

Scene level interpretation

Consistency handling

Enrichment

Final Semantic Annotation

Multimedia Knowledge Laboratory
Informatics and Telematics Institute
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**

<table>
<thead>
<tr>
<th>Natural</th>
<th>Outdoors ⊔ ¬ ManMade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mountainous</td>
<td>Natural ⊔ ¬ Coastal</td>
</tr>
<tr>
<td>Beach</td>
<td>Coastal ⊓ ∃contains.Sand</td>
</tr>
<tr>
<td>∃contains.Mountain ⊆ Mountainous</td>
<td></td>
</tr>
<tr>
<td>∃contains.Sea ⊆ Coastal</td>
<td></td>
</tr>
<tr>
<td>∃contains.Sand ⊓ Mountainous ⊆ ⊥</td>
<td></td>
</tr>
<tr>
<td>Outdoor ⊓ Indoor ⊆ ⊥</td>
<td></td>
</tr>
</tbody>
</table>

**Initial Assertions**

| (image:Indoor) ≥ 0.67 |
| (image:∃contains.Sea) ≥ 0.73 |
| (image:∃contains.Sand) ≥ 0.58 |
| (image:∃contains.Mountain) ≥ 0.85 |

Disjointness axioms removed

| (image:Indoor) ≥ 0.67 |
| (image:∃contains.Sea) ≥ 0.73 |
| (image:∃contains.Sand) ≥ 0.58 |
| (image:Coastal) ≥ 0.73 |
| (image:Beach) ≥ 0.58 |
| (image:Natural) ≥ 0.73 |
| (image:Outdoor) ≥ 0.73 |

Scene level hierarchy

<table>
<thead>
<tr>
<th>Outdoor (0.85)</th>
<th>Indoor (0.67)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural (0.85)</td>
<td>ManMade</td>
</tr>
<tr>
<td>Coastal (0.58)</td>
<td>Mountainous (0.85)</td>
</tr>
<tr>
<td>Beach (0.58)</td>
<td></td>
</tr>
</tbody>
</table>
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

T1 step

Initial Assertions

- (image:Indoor) ≥ 0.67
- (image:contains.Sea) ≥ 0.73
- (image:contains.Sand) ≥ 0.58
- (image:contains.Mountain) ≥ 0.85

Disjointness axioms removed

Disjoint axioms restored

Domain TBox

- Natural ⊑ Outdoors ⊔ ¬ ManMade
- Mountainous ⊑ Natural ⊔ ¬ Coastal
- Beach ⊑ Coastal ⊔ ¬ contains.Sand
- contains.Mountain ⊑ Mountainous
- contains.Sea ⊑ Coastal
- contains.Sand ⊑ Mountainous ⊑ ⊥
- Outdoor ⊑ Indoor ⊑ ⊥

Scene level hierarchy

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

Disjoint inferred disjoint
directly disjoint

Inconsistency handling

- (image:Outdoor) ≥ 0.85
- (image:contains.Mountain) ≥ 0.85
- (image:contains.Sea) ≥ 0.73
- (image:contains.Sand) ≥ 0.58
- (image:Coastal) ≥ 0.73
- (image:Beach) ≥ 0.58
- (image:Natural) ≥ 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) ≥ 0.67
- (image:contains.Sea) ≥ 0.73
- (image:contains.Sand) ≥ 0.58
- (image:contains.Mountain) ≥ 0.85

Disjointness axioms removed

- (image:Indoor) ≥ 0.67
- (image:contains.Sea) ≥ 0.73
- (image:contains.Sand) ≥ 0.58
- (image:Coastal) ≥ 0.73
- (image:Beach) ≥ 0.58
- (image:Natural) ≥ 0.73
- (image:Outdoor) ≥ 0.73
- (image:contains.Mountain) ≥ 0.85
- (image:Mountainous) ≥ 0.85
- (image:Natural) ≥ 0.85
- (image:Outdoor) ≥ 0.85

Scene level hierarchy

**Domain TBox**

- Natural ⊑ Outdoors ⊔ ¬ ManMade
- Mountainous ⊑ Natural ⊔ ¬ Coastal
- Beach ⊑ Coastal ⊔ ∃contains.Sand
- ∃contains.Mountain ⊑ Mountainous
- ∃contains.Sea ⊑ Coastal
- ∃contains.Sand ⊔ Mountainous ⊑ ⊥
- Outdoor ⊔ Indoor ⊑ ⊥

Disjoint axioms restored

- (image:Indoor) ≥ 0.67
- (image:contains.Sea) ≥ 0.73
- (image:contains.Sand) ≥ 0.58
- (image:Coastal) ≥ 0.73
- (image:Beach) ≥ 0.58
- (image:Natural) ≥ 0.85
- (image:Outdoor) ≥ 0.85

Inconsistency handling

**Final Assertions**

- (image:contains.Mountain) ≥ 0.85
- (image:Mountainous) ≥ 0.85
- (image:Natural) ≥ 0.85
- (image:Outdoor) ≥ 0.85
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 two implementations for object level
• Experiment II
  • one implementation for scene level
  • one implementation for object level

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

Countryside_buildings ⊑ ∃contains.Buildings □ ∃contains.Foliage
Countryside_buildings ⊑ Landscape
∃contains.Forest □ ∃contains.Grass □ ∃contains.Tree ⊑ ∃contains.Foliage
RockySide ⊑ ∃contains.Cliff
RockySide ⊑ ∃contains.Mountainous
Roadside ⊑ ∃contains.Road
Roadside ⊑ Landscape
∃contains.Sea ≡ Coastal
Coastal ⊑ Natural
∃contains.Forest ⊑ Landscape
Beach ≡ Coastal □ ∃contains.Sand
Beach ⊑ Natural
Cityscape ⊑ ManMade
∃contains.Sky ⊑ Outdoor
∃contains.Trunk ⊑ ∃contains.Tree
Mountainous □ Coastal ⊑ ⊥
Natural □ ManMade ⊑ ⊥
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

<table>
<thead>
<tr>
<th>Concept</th>
<th>Analysis</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>Building</td>
<td>1.00</td>
<td>0.17</td>
</tr>
<tr>
<td>Grass</td>
<td>0.06</td>
<td>0.40</td>
</tr>
<tr>
<td>Foliage</td>
<td>0.99</td>
<td>0.70</td>
</tr>
<tr>
<td>Sky</td>
<td>0.93</td>
<td>0.87</td>
</tr>
<tr>
<td>Cliff</td>
<td>0.98</td>
<td>0.21</td>
</tr>
<tr>
<td>Tree</td>
<td>0.22</td>
<td>0.65</td>
</tr>
<tr>
<td>Trunk</td>
<td>0.38</td>
<td>0.65</td>
</tr>
<tr>
<td>Sand</td>
<td>0.49</td>
<td>0.37</td>
</tr>
<tr>
<td>Sea</td>
<td>0.72</td>
<td>0.46</td>
</tr>
<tr>
<td>Conifers</td>
<td>1.00</td>
<td>0.01</td>
</tr>
<tr>
<td>Mountain</td>
<td>0.14</td>
<td>0.01</td>
</tr>
<tr>
<td>Boat</td>
<td>0.10</td>
<td>0.40</td>
</tr>
<tr>
<td>Road</td>
<td>0.15</td>
<td>0.50</td>
</tr>
<tr>
<td>Ground</td>
<td>0.06</td>
<td>0.57</td>
</tr>
<tr>
<td>Person</td>
<td>0.49</td>
<td>0.54</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Concept</th>
<th>Analysis</th>
<th>Reasoning</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
<td>F-M</td>
</tr>
<tr>
<td>Countryside_buildings</td>
<td>0.30</td>
<td>1.0</td>
<td>0.46</td>
</tr>
<tr>
<td>Rockyssey</td>
<td>0.68</td>
<td>0.70</td>
<td>0.69</td>
</tr>
<tr>
<td>Roadside</td>
<td>0.68</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>Forest</td>
<td>0.75</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>Coastal</td>
<td>0.85</td>
<td>0.67</td>
<td>0.75</td>
</tr>
<tr>
<td>Outdoor</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Indoor</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Natural</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ManMade</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cityscape</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mountainous</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Beach</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Higher impact as the analysis supported concepts are characterised are more strongly related to each other.
**Experiment II – Object level concepts**

<table>
<thead>
<tr>
<th>Concept</th>
<th>Analysis</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td><strong>Building</strong></td>
<td>0.54</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>Roof</strong></td>
<td>0.33</td>
<td>0.54</td>
</tr>
<tr>
<td><strong>Grass</strong></td>
<td>0.49</td>
<td>0.42</td>
</tr>
<tr>
<td><strong>Foliage</strong></td>
<td>0.48</td>
<td>0.84</td>
</tr>
<tr>
<td><strong>Dried-Plant</strong></td>
<td>0.07</td>
<td>0.11</td>
</tr>
<tr>
<td><strong>Ground</strong></td>
<td>0.26</td>
<td>0.33</td>
</tr>
<tr>
<td><strong>Person</strong></td>
<td>0.75</td>
<td>0.51</td>
</tr>
<tr>
<td><strong>Sky</strong></td>
<td>0.95</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>Cliff</strong></td>
<td>0.65</td>
<td>0.45</td>
</tr>
<tr>
<td><strong>Tree</strong></td>
<td>0.49</td>
<td>0.52</td>
</tr>
<tr>
<td><strong>Trunk</strong></td>
<td>0.26</td>
<td>0.28</td>
</tr>
<tr>
<td><strong>Sand</strong></td>
<td>0.02</td>
<td>0.10</td>
</tr>
<tr>
<td><strong>Sea</strong></td>
<td>0.69</td>
<td>0.60</td>
</tr>
<tr>
<td><strong>Wave</strong></td>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td><strong>Boat</strong></td>
<td>0.41</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>Road</strong></td>
<td>0.50</td>
<td>0.69</td>
</tr>
</tbody>
</table>

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.

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).
Conclusions

• The proposed Fuzzy DLs reasoning enables
  • formal handling of annotations uncertainty semantics
  • utilisation of domain semantics
  • consistent interpretations / descriptions

• The use of explicit semantics is crucial in multimedia semantics extractions; yet not the only necessary component

• Largely miscalculated degrees can mislead the interpretation
  • combined usage of probabilistic knowledge could be an a possible solution
Use of sequential methods and knowledge for spatio-temporal localization of subevents

• Motivation/Contributions
• Spatial Localization
• Temporal Localization
• Knowledge Representation & Reasoning
  • SpatioTemporal Activity Ontology
  • Domain Events Ontology and Temporal Rules
• Experiments
Motivation

• Event detection from video has attracted much attention in many applications:
  • Surveillance
  • E-home
  • Entertainment (recommendations, ads)
• Cumbersome when done manually
• Difficult to automate in a non-adhoc manner due to the vast range of content
• Need to combine principled low-level methods with high-level knowledge structures
Contributions

• Automated detection of events in space and time using statistical processing of input data.
  • Spatial processing localizes active pixels
  • Temporal processing finds when their activity changes i.e. new events/activities

• The processing of only active pixels
  • Reduces false alarms
  • Increases computational efficiency
Example surveillance video
Spatial Processing System

Video → Inter-frame → Activity area extraction (kurtosis)
Temporal Processing System

Activity area

Temporal processing of data in activity area

Detection of times of change: beginning and end of events, via Sequential Likelihood Ratio Testing (SLRT)

Activity areas for sub-events
Spatial domain: Activity Area

- Activity Area = the pixels where activity occurs during the video
Activity Area Extraction

• Illumination changes between successive frames can be caused by motion or measurement noise:
  \[ H_0 : v_k(r) = z_k(r) \]
  \[ H_1 : v_k(r) = u_k(r) + z_k(r) \]
• Noise is modeled as Gaussian
• Gaussian random variables have zero kurtosis:
  \[ Kurt(z) = E[z^4] - 3E^2[z^2] \]
• We estimate the kurtosis of each pixel’s illumination changes over several frames: when they are very high, we consider that the pixel is active.
Temporal Sub-event Localization

• Currently changes in video are limited to detecting shot changes
• Often, more than one activity occurs during a shot
• We consider that different activities are characterized by different motion
• We detect time instants where the activity changes, to find the beginning and end of new activity.
Sequential Likelihood Ratio Testing (SLRT)

- Current activity corresponds to $H_0$, new activity to $H_1$ (after a change):
  
  \[
  H_0 : \nu_0 \sim P_0 \\
  H_1 : \nu_1 \sim P_1
  \]

- We approximate $P_0$ using an initial set of frames and $P_1$ using the current and a windowed set of past frames.
SLRT

• At each time instant $k$ (frame $k$), we estimate the “current” log-likelihood ratio:

$$T(k) = \sum_{i=1}^{k} \frac{P_1(k)}{P_0(k)}$$

• We approximate $P_0$ using an initial set of frames and $P_1$ using the current and a windowed set of past frames.
The SLRT is compared to the Wald threshold:

\[ \eta = \log \left( \frac{1 - \beta}{\alpha} \right) \]

Where \( \beta \) is the probability of miss and \( \alpha \) the probability of false alarm.

When the SLRT is above \( \eta \), a change has occurred in the activity, i.e. a new activity began.
Knowledge Representation & Reasoning

• Make involved semantics formal & explicit
  • knowledge sharing & reuse
  • extensible & modular approach

• Two dimensions to structure knowledge
  • temporal vs atemporal (e.g. “car overtake”, “car is type of vehicle”)
  • generic to spatiotemporal analysis vs specific to application domain (e.g. “trajectory attributes”, “people fighting presupposes people meeting”)
Representation Issues

• Ontology languages (e.g. OWL DL) are ideal for representing atemporal semantics but lack support for temporal semantics

• Variants of temporal logic are ideal for supporting temporal associations but are incompatible with the Semantic Web (SW)

• Temporal DLs are both SW compatible and support temporal semantics, but there are not temporal DLs reasoning engines implemented
Proposed Knowledge Infrastructure

- Hybrid representation combines ontologies and rules
  - use of ontologies for atemporal semantics
  - use of rules for temporal semantics
  - RacerPro(*) is used as the reasoning engine

- SpatioTemporal Activity Ontology (atemporal dimension) and Rules (temporal dimensions)
  - serve as common vocabulary for interchanging analysis results among different implementations
  - serve as generic, qualitative representation on top of which domain specific semantics can be inferred

- Domain Specific Event Rules
  - capture the temporal dimension of domain semantics
  - independent of video analysis implementation, which can be used as a black box
  - support for additional events affects only the extension of domain rules

(*) http://www.racer-systems.com/
SpatioTemporal Activity Ontology

• Main classes: ActivityArea, StaticArea, and Trajectory
  • A trajectory of an activity area is characterised by its start and ending location, and by its direction
  • Intermediate locations can be represented using the hasLocation property
  • Motion attributes include magnitude and variance

\[
\begin{align*}
\text{ActivityArea} & \sqsubseteq \text{StaticArea} \sqsubseteq \bot \\
\text{ActivityArea} & \sqsubseteq (\exists \text{has.Trajectory}) \sqcap (\exists \text{has.Motion}) \\
\text{Trajectory} & \sqsubseteq (\exists \text{hasStartLocation.Location}) \\
& \sqcap (\exists \text{hasEndLocation.Location}) \\
\text{Trajectory} & \sqsubseteq (\exists \text{has.Direction}) \\
\text{Motion} & \sqsubseteq (\exists \text{has.Magnitude}) \\
\text{Motion} & \sqsubseteq (\exists \text{has.Variance})
\end{align*}
\]
SpatioTemporal Activity Rules

- Nine generic activity events

<table>
<thead>
<tr>
<th>Event</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>same-direction</td>
<td>activity areas whose trajectories undergo similar direction</td>
</tr>
<tr>
<td>different-direction</td>
<td>activity areas whose trajectories undergo different directions</td>
</tr>
<tr>
<td>variance-increase</td>
<td>activity area whose motion is undergoing variance increase</td>
</tr>
<tr>
<td>magnitude-increase</td>
<td>activity area whose motion activity is increased</td>
</tr>
<tr>
<td>magnitude-reduction</td>
<td>activity area whose motion activity is reduced</td>
</tr>
<tr>
<td>move-towards</td>
<td>an activity area approaching a still one,</td>
</tr>
<tr>
<td></td>
<td>i.e. reducing the in-between distance</td>
</tr>
<tr>
<td>move-away</td>
<td>an activity area drawing away from a still one,</td>
</tr>
<tr>
<td></td>
<td>i.e. increasing the in-between distance</td>
</tr>
<tr>
<td>move-meet</td>
<td>activity areas approaching each other</td>
</tr>
<tr>
<td>move-split</td>
<td>activity areas drawing away from each other</td>
</tr>
</tbody>
</table>
"move-towards" rule definition

(def rule (move-towards ?obj1 ?obj2) ?t1 ?t2)
((?obj1 ActivityArea) ?t0 ?t_n)
((?obj2 StaticArea) ?t0 ?t_n)
((?traj1 Trajectory) ?t0 ?t_n)
((?obj1 ?traj1 has) ?t1 ?t2)
((?l1 Location) ?t0 ?t_n)
((?l2 Location) ?t0 ?t_n)
((?traj1 ?l1 hasStartLocation) ?t0 ?t_n)
((?obj2 ?l2 hasStartLocation) ?t0 ?t_n)
((different - location ?l1 ?l2) ?t1 ?t3)
((in - motion ?obj1) ?t3 ?t4)
((?l11 Location) ?t0 ?t_n)
((?l22 Location) ?t0 ?t_n)
((?traj1 ?l11 hasEndLocation) ?t0 ?t_n)
((?obj2 ?l22 hasEndLocation) ?t0 ?t_n)
((nearby - location ?l11 ?l22) ?t4 ?t2))
Domain Specific Rules

- Two domains have been investigated
  - people surveillance (7 events)
  - car traffic (3 events)
- People surveillance domain events rules

<table>
<thead>
<tr>
<th>Event</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>people-meet</td>
<td>people moving towards each other</td>
</tr>
<tr>
<td>people-split</td>
<td>people moving away from each other</td>
</tr>
<tr>
<td>people-walking-together</td>
<td>starting from nearby positions</td>
</tr>
<tr>
<td>people-fight</td>
<td>people meet and continue moving in similar direction</td>
</tr>
<tr>
<td>people-fight-chase</td>
<td>people meet and increased activity follows</td>
</tr>
<tr>
<td>people-fight-one-down</td>
<td>people fight followed by motion at the same direction</td>
</tr>
<tr>
<td>people-fight-run-away</td>
<td>people fight followed by activity reduction of one person</td>
</tr>
<tr>
<td>people-fight-run-away</td>
<td>people fight followed by activity increase of one person</td>
</tr>
</tbody>
</table>
“people meet” and “people fight” rules definition

\[
\text{(define – event – rule ((people – meet } ?obj_1 \ ?obj_2) \ ?t_1 \ ?t_2)} \\
\text{((?obj_1 Person) \ ?t_0 \ ?t_n)} \\
\text{((?obj_2 Person) \ ?t_0 \ ?t_n)} \\
\text{(move – meet \ ?obj_1 \ ?obj_2) \ ?t_1 \ ?t_2))}
\]

\[
\text{(define – event – rule ((people – fight } ?obj_1 \ ?obj_2) \ ?t_1 \ ?t_2)} \\
\text{(people – meet \ ?obj_1 \ ?obj_2) \ ?t_1 \ ?t_3)} \\
\text{((?m_1 Motion) \ ?t_0 \ ?t_n)} \\
\text{((?obj_1 \ ?m_1 has) \ ?t_0 \ ?t_n)} \\
\text{(variance – increase \ ?m_1) \ ?t_3 \ ?t_2))}
\]

Temporal aggregation of simpler events models the semantics of the more complex
“cars-approach” rule definition

(define - event - rule ((cars - approach ?obj1 ?obj2) ?t1 ?t2)
  ((?obj1 Car) ?t0 ?tn)
  ((?obj2 Car) ?t0 ?tn)
  ((move - meet ?obj1 ?obj2) ?t1 ?t2))

- The concepts and events included in the SpatioTemporal Ontology and Rules allow the straightforward extension to a different application domain.
Experiment – Taxi Sequence
Experiment – Meet Sequence
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
• Different applications and requirements
• Ongoing research in all areas
• Future direction: analysis+reasoning for social (Web 2.0) applications
Thank you!

CERTH-ITI / Multimedia Knowledge Laboratory

http://mklab.iti.gr

Dr. Alexia Briassouli
Dr. Stamatia Dasiopoulou
Georgios Th. Papadopoulos