Introducing Ontological Realism for Unsupervised Detection and Annotation of Operationally Significant Activity in Surveillance Videos

October 27-28, 2010
George Mason University, Fairfax, Virginia

W. Ceusters¹, J. Corso², Y. Fu², M. Petropoulos², V. Krovi³

¹ Ontology Research Group, NYS Center of Excellence in Bioinformatics & Life Sciences, 701 Ellicott Street, Buffalo NY
² School of Computer Science and Engineering, University at Buffalo
³ Department of Mechanical and Aerospace Engineering, University at Buffalo
The ISTARE Team

**Jason Corso**
Computer Science and Engineering
Computer Vision
Statistical Learning
Ontological Reasoning

**Raymond Fu**
Computer Science and Engineering
Computer Vision
Manifold Learning

**Werner Ceusters**
Psychiatry
Ontology Research Group
Ontology
Referent Tracking

**Michalis Petropoulos**
Computer Science and Engineering
Databases Storage
Query Languages

**Venkat Krovi**
Mechanical and Aerospace Engineering
Robotics
Biomechanics
Articulated Motion

**ISTARE**
Intelligent Spatiotemporal Activity Reasoning Engine
DARPA’s Mind’s Eye Program (1)

- **Purpose**: develop software for a smart camera, which is mountable on, f.i., man-portable UGVs and which exhibits capabilities necessary to perform surveillance in operational missions.

- **Capabilities requested**:
  - recognize the primitive actions that take place between objects in the visual input, with a particular emphasis on actions that are relevant in typical operational scenarios (e.g., vehicle APPROACHES checkpoint; person EXITS building).
DARPA’s Mind’s Eye Program (2)

Capabilities requested (continued):

- learning and cross-scene application of invariant spatio-temporal patterns,
- issuing alerts to activities of interest,
- performing interpolation to fill in likely explanations for gaps in the perceptual experience,
- explaining its reasoning by displaying relevant video segments for what has been observed, and by generating visualizations for what is hypothesized.
## Actions of interest

<table>
<thead>
<tr>
<th>approach</th>
<th>carry</th>
<th>dig</th>
<th>fall</th>
<th>give</th>
<th>hit</th>
<th>lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>arrive</td>
<td>catch</td>
<td>drop</td>
<td>flee</td>
<td>go</td>
<td>hold</td>
<td>move</td>
</tr>
<tr>
<td>attach</td>
<td>chase</td>
<td>enter</td>
<td>fly</td>
<td>hand</td>
<td>kick</td>
<td>open</td>
</tr>
<tr>
<td>bounce</td>
<td>close</td>
<td>exchange</td>
<td>follow</td>
<td>haul</td>
<td>jump</td>
<td>pass</td>
</tr>
<tr>
<td>bury</td>
<td>collide</td>
<td>exit</td>
<td>get</td>
<td>have</td>
<td>leave</td>
<td>pick up</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>push</th>
<th>run</th>
<th>touch</th>
<th>pass</th>
<th>receive</th>
<th>take</th>
</tr>
</thead>
<tbody>
<tr>
<td>put down</td>
<td>snatch</td>
<td>turn</td>
<td>pick up</td>
<td>replace</td>
<td>throw</td>
</tr>
<tr>
<td>raise</td>
<td>stop</td>
<td>walk</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Run - Take
Drop - Lift
Filling gaps
System overview

New York State Center of Excellence in Bioinformatics & Life Sciences

Integrate SOA computer vision

Visual Processing
- Low Level Processing
- Object and Part Detection
- Object and Part Tracking

Camera

Stored Video Examples

Operationally realistic scenes

Research focus on enabling technologies

Visual Intelligence
- Visual Event Learning
- Spatiotemporal Activity Models
- Visual Inspection & Envisionment
- Grounding
- Activity Ontology
- Symbolic Reasoning
- Declarative Knowledge

Visual Interaction
- Dialog Processing
- Referent Tracking

Human-In-The-Loop

Mixed-Factor Graphical Models

Articulated Activities

Anomaly Detection

Visual Inspection & Envisionment

Parts-Pursuit

Analysis-By-Synthesis

Visual Intelligence

ISTARE
Intelligent Spatio-Temporal Activity Reasoning Engine

Query
**Representation strategy**

- Hierarchical and generative representation.
- Allows for reuse of parts at the low- and mid-levels.
- Each object is a rooted graph with a mixture of directed and undirected edges, which govern the nature of the connections.
- At the low-level, atoms subtend spatiotemporal lines, planes, and regions over which we learn non-linear manifolds.
Representation of activities

- Activities are detected as changes in conglomerates of these object graphs, which have connections at both the mid- (part) and the high-level (object).
Learning with human in the loop

- For learning the structure at the mid-to-high levels.
  - Proceed with an active semi-supervised learning schema to iteratively refine a distance metric.
  - At mid-level: clustering is the game.
    - Iteratively seek grouping constraints.
    - Interactively link clusters to **ontology**.
    - Working with both spectral and hierarchical clustering.
  - At high-level: get feedback from the user based on our **ontological** grounding: ask Questions, such as “is the person in this clip giving the object to the other person?”
ISTARE Ontology

• Roles:
  – **Learning**: it will help guide a learning algorithm to remain in plausible configurations.
  – **Inference**: it will support reasoning of plausible explanations of objects and activities in existing and missing parts of the signal.

• Components:
  – L1 $\leftrightarrow$ L1:
    – How humans interact with objects and other humans in various scenarios.
    – How motions of object-parts contribute to full object motion.
  – L1 $\leftrightarrow$ L3:
    – How manifolds in the video correspond to entities videotaped.
  – L1 $\leftrightarrow$ L2 $\leftrightarrow$ L3:
    – How analysts interpret videos and corresponding reality.
L1 – L3
The basis: Ontological Realism

Basic axioms of realism:
1. There is an external reality which is ‘objectively’ the way it is;
2. That reality is accessible to us;
3. We build in our brains cognitive representations of reality;
4. We communicate with others about what is there, and what we believe there is there.

Three levels of reality, two sorts of representations

L1

R

L2

beliefs

L3

symbolizations

‘about’
Main distinctions in BFO

- Some continuant universal
- Some occurred universal
- Some continuant particular
- Some occurred particular
Sorts of relations

Unconstrained reasoning

\(\text{UtoU}: \text{isa, partOf, …}\)

\(\text{PtoU}: \text{instanceOf, lacks, denotes…}\)

\(\text{PtoP}: \text{partOf, denotes, subclassOf,…}\)

OWL-DL reasoning
Region Connection Calculus (RCC8)

8 possible relations between regions at a time

- DC
- EC
- PO
- TPP
- NTPP
- TPPI
- NTPPI

RCC8 reasoning

- $\text{rel}_1(x,y,t) \land \text{rel}_2(y,z,t) \Rightarrow \text{rel}_3(x,z,t) ?$
- e.g. $\text{DC}(x,y,t) \land \text{DC}(y,z,t)$
- maintained in tables

RCC8: conceptual neighborhood

If $\text{rel}_1$ at $t_1$, what possible relations at $t_2$?

- DC
- EC
- PO
- TPP
- NTPP
- TPPI
- NTPPI

### Basic ‘Motion Classes’

<table>
<thead>
<tr>
<th></th>
<th>DC</th>
<th>EC</th>
<th>PO</th>
<th>TPP</th>
<th>NTPP</th>
<th>EQ</th>
<th>TPPI</th>
<th>NTPPI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Starts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC</td>
<td><strong>Split</strong></td>
<td><strong>Peripheral</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PO</td>
<td><strong>Leave or Reach</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTPP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTPPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ends</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC</td>
<td><strong>External</strong></td>
<td><strong>Hit</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PO</td>
<td><strong>Leaf or Reach</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTPP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTPPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reasoning with motion classes

- \( mc_1(x,y,t) \land mc_2(y,z,t) \implies mc_3(x,z,t) \)?
- e.g. \( \text{leave}(x,y,t) \land \text{leave}(y,z,t) \)

Reasoning with motion classes

- $mc_1(x, y, t) \land mc_2(y, z, t) \rightarrow mc_3(x, z, t)$
- e.g. $leave(x, y, t) \land leave(y, z, t)$

Reasoning with motion classes

- $mc_1(x,y,t) \land mc_2(y,z,t) \rightarrow mc_3(x,z,t)$ ?
- e.g. leave(x,y,t) $\land$ leave(y,z,t)

- all possibilities also in tables

RCC8/MC14 and Ontological Realism

In ontological realism:
- regions don’t move
- material entities are located in regions
- while material entities move or shrink/expand:
  - they are located at each t in a different region
  - each such region is part of the region formed by all the regions visited, thus constituting a path
  - ...

An unambiguous mapping is possible
RCC8/MC14 and action verbs

‘approach’
RCC8/MC14 and action verbs

‘approach’

- Invariant:
  - shrink of the region between the entities involved in an approach
RCC8/MC14 and action verbs

<table>
<thead>
<tr>
<th>approach</th>
<th>carry</th>
<th>dig</th>
<th>fall</th>
<th>give</th>
<th>hit</th>
<th>lift</th>
<th>push</th>
<th>run</th>
<th>touch</th>
</tr>
</thead>
<tbody>
<tr>
<td>arrive</td>
<td>catch</td>
<td>drop</td>
<td>flee</td>
<td>go</td>
<td>hold</td>
<td>move</td>
<td>put down</td>
<td>snatch</td>
<td>turn</td>
</tr>
<tr>
<td>attach</td>
<td>chase</td>
<td>enter</td>
<td>fly</td>
<td>hand</td>
<td>kick</td>
<td>open</td>
<td>raise</td>
<td>stop</td>
<td>walk</td>
</tr>
<tr>
<td>bounce</td>
<td>close</td>
<td>exchange</td>
<td>follow</td>
<td>haul</td>
<td>jump</td>
<td>pass</td>
<td>receive</td>
<td>take</td>
<td></td>
</tr>
<tr>
<td>bury</td>
<td>collide</td>
<td>exit</td>
<td>get</td>
<td>have</td>
<td>leave</td>
<td>pick up</td>
<td>replace</td>
<td>throw</td>
<td></td>
</tr>
</tbody>
</table>

- all can be expressed in terms of mc14 (with the addition of direction and some other features)
- from mc to the verbs: requires additional information on the nature of the entities involved – to be encoded in the ontology
Action verbs and Ontological Realism

• Many caveats:
  – the way matters are \textit{expressed} in natural language does not correspond faithfully with the way matters \textit{are}

‘approach’

$x$ orbiting around $y$

$x$ taking distance from $y$?

→ $x$’s process didn’t change

→ ‘to approach’ is a verb, but it does not represent a process, rather implies a process.
Action verbs and Ontological Realism

• Approaching following a forced path
RCC8/MC14 & video as 2D+T representation of 3D+T

man entering building: the first-order view
RCC8/MC14 & video as 2D+T representation of 3D+T

man entering building: the video view
RCC8/MC14 & video as 2D+T representation of 3D+T

egg crashing on wall: the video view

• Requires additional mapping from the motion of manifolds in the video to the corresponding motion of the corresponding entities in reality
ISTARE Ontology design principles

- **Main objective:**
  - Being pragmatic, yet adhering to the principles of Ontological Realism.

- **Each representational unit (RU) denotes either**
  - a *universal*:
    - Human Being, Motion;
  - a *fiat class*, i.e, a portion of reality demarcated by human fiat, and relevant for the goals of ISTARE:
    - Canonically-limbed human being,
    - High five.
A partly worked out example
Human anatomy (L1)

- c1 member-of Canonically-Limbed Human Being at t, then:
  - c1 has-part c2 at t
  - c2 instance-of Left Arm at t
  - c1 has-part c3 at t
  - c3 instance-of Right Arm at t
  - c1 has-part c4 at t
  - c4 instance-of Left Lower Limb at t
  - c1 has-part c5 at t
  - c5 instance-of Right Lower Limb at t
  - c1 has-part c6 at t
  - c6 instance-of Head at t
  - c1 has-part c7 at t
  - c7 instance-of Torso at t
  - c2 adjacent-to c7 at t
  - c3 adjacent-to c7 at t
  - c4 adjacent-to c7 at t
  - c5 adjacent-to c7 at t
  - c6 adjacent-to c7 at t
  - …
Human physiology (L1)

- $c_1$ member-of Canonically-Limbed Human Being at $t$, then:
  - $sdc_1$ inheres-in $c_1$ at $t$
  - $sdc_1$ instance-of Disposition-to-Walk at $t$
  - $sdc_2$ inheres-in $c_1$ at $t$
  - $sdc_2$ instance-of Disposition-to-Run at $t$
  - ...

<table>
<thead>
<tr>
<th>approach</th>
<th>carry</th>
<th>dig</th>
<th>fall</th>
<th>give</th>
<th>hit</th>
<th>lift</th>
<th>push</th>
<th>run</th>
<th>touch</th>
</tr>
</thead>
<tbody>
<tr>
<td>arrive</td>
<td>catch</td>
<td>drop</td>
<td>flee</td>
<td>go</td>
<td>hold</td>
<td>move</td>
<td>put down</td>
<td>snatch</td>
<td>turn</td>
</tr>
<tr>
<td>attach</td>
<td>chase</td>
<td>enter</td>
<td>fly</td>
<td>hand</td>
<td>kick</td>
<td>open</td>
<td>raise</td>
<td>stop</td>
<td>walk</td>
</tr>
<tr>
<td>bounce</td>
<td>close</td>
<td>exchange</td>
<td>follow</td>
<td>haul</td>
<td>jump</td>
<td>pass</td>
<td>receive</td>
<td>take</td>
<td></td>
</tr>
<tr>
<td>bury</td>
<td>collide</td>
<td>exit</td>
<td>get</td>
<td>have</td>
<td>leave</td>
<td>pick up</td>
<td>replace</td>
<td>throw</td>
<td></td>
</tr>
</tbody>
</table>

impossible under certain circumstances
Human physiology (L1)

- o1 member-of Canonical-Human-Walking, then:
  - o1 realization-of sdc1
  - sdc1 instance-of Disposition-to-Walk at t
  - sdc1 inheres-in c1 at t
  - c1 instance-of Canonically-Limbed Human Being at t
  - o1 has-agent c1 at t
  - o1 has-part o2
  - o2 instance-of Walking Leg Motion
  - o2 has-agent c2 at t
  - c2 part-of c1 at t
  - c2 instance-of Left Lower Limb at t
  -  
  - o3 instance-of Walking Leg Motion
  - o3 has-agent c3 at t
  - c3 part-of c1 at t
  - c3 instance-of Right Lower Limb at t
  - c1 located-in r1 at t0
  - t0 earlier t
  - c1 located-in r2 at t1
  - t earlier t1
  - ...

But: elliptical work-out, walking in circle, …
Human anatomy applied to video frames

- Detection and grounding of manifolds:
  - m1, denotes c6 at t?
  - m2, denotes c7 at t?
  - m3, denotes c2 at t?
  - m4, denotes c4 at t?
  - m5, denotes c5 at t?
Elements of ontology-based reasoning

• Projection of RCC and MCC in L3 to portions of reality in L1:
  – EC → adjacent-to
  – shrink → shrinking
    → moving away from camera
  – hit → approach in front or behind object
  – hit < shrink → ‘shrinking’ object passed behind
  – …

• Human in the loop
Work in progress

• Finished:
  – Feasibility assessment by team
  – Core ontological principles

• Started:
  – Population of ontology
  – Documentation of invariants

• Next steps:
  – annotation of ‘verbs’ in terms of RUs from the ontology
  – development of reasoning engine
  – integration with manifold detection and tracking