Introducing Ontological Realism for Semi-Supervised Detection and Annotation of Operationally Significant Activity in Surveillance Videos

Werner Ceusters[1], Jason Corso[2], Yun Fu[2],
Michalis Petropoulos[2], Venkat Krovi[3]

[1] Ontology Research Group, NYS Center of Excellence in Bioinformatics & Life Sciences, 701 Ellicott Street, Buffalo NY
[2] Department of Computer Science and Engineering, University at Buffalo
[3] Department of Mechanical and Aerospace Engineering, University at Buffalo
{ceusters, jcorso, yunfu, mpetropo, vkrovi}@buffalo.edu

Abstract. As part of DARPA's Mind's Eye program, a video-analysis software platform able to detect operationally significant activity in videos is being developed. The goal is to describe such activity semi-automatically in terms of verb phrases mapped to a realism-based ontology that can be used to infer and even predict further activities that are not directly visible. We describe how Region Connection Calculus and its derivative, Motion Class Calculus, can be used together to link the spatiotemporal changes that pixel-aggregates undergo in video-displays to the corresponding changes of the objects in reality that were recorded and to linguistic descriptions thereof. We discuss how Ontological Realism can be used as a safeguard to drawing such correspondences naively.

Keywords: ontological realism, video analysis, activity detection

1 Introduction

Automatic video-understanding is a relatively new field for which the research agenda has been set only fairly recently. Cetin identified in 2005 two grand challenges for video-analysis: the first was to develop applications that allow a natural high-level interaction with multimedia databases; the second was finding adequate algorithms for detecting and interpreting humans and human behavior in videos containing also audio and text information [1]. Early successes have focused on particular sub-problems, such as face detection [2].

State of the art systems are capable of detecting instances of objects – sometimes referred to as 'the nouns' of the scene – among few hundreds of object classes [3] and contests such as the PASCAL Challenge annually pit the world's best object detection methods on novel datasets [4]. Now, however, a more elusive problem presents itself: finding the 'verbs' of the scene. As Biederman stated nearly 30 years ago: specifying not only the elements in an image but also the manner in which they are interacting

and relating to one another is integral to full image understanding [5]. However, recognizing human actions, especially with a view to understanding their underlying motivation, has proved to be an extremely challenging task. This is because (1) behavior is the result of a complex combination of coordinated actions, (2) motion and behavior are described linguistically at a wide variety of spatiotemporal scales, and most importantly (3) the unambiguous extraction of intent from motion alone can never be achieved due to the significant dependence upon contextual knowledge.

Solving these problems, specifically in the context of surveillance, is the objective of DARPA's *Mind's Eye* program which seeks to embed in a *smart camera* sufficient visual intelligence to detect, interpolate and even predict activities in an area of observation and, as a specific requirement, to describe these activities in terms of 'verbs' (Table 1) [6].

As successful proposers to this program, our answer is ISTARE: a platform which will suitably represent articulated motion in a three-layer hierarchical dynamical graphical model consisting of (1) a lowest level of representation formed by points, lines and regions in their spatiotemporal context, (2) a mid-level capturing the spatio-temporal coherence inherent in the appearance, structure and motions of the atoms in the lower level, and (3) generalizations of the reusable mid-level parts into full objects and activities at the high-level (Fig.1). Part of that platform is an ontology which grounds the models with proper semantics thereby driving both learning and inference. A human-in-the-loop is the bridge between models and symbolic representations in case of ambiguities. But rather than requiring laborious annotation in such case, the human simply needs to answer yes/no questions generated by our methods.

In this communication, we describe our strategy to make the ISTARE approach in general, and the computational structures resulting from automated video analysis and annotation within the ISTARE platform specifically, compatible with ongoing research in the field. Using Motion Classes (MC) as an example, we demonstrate how Ontological Realism is an important building block in this endeavor and how it is able to tie the various pieces – reality, spatiotemporal models and linguistic descriptions – together.

Table 1. Verbs of interest for activity detection in the Mind's Eye video-analysis program

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

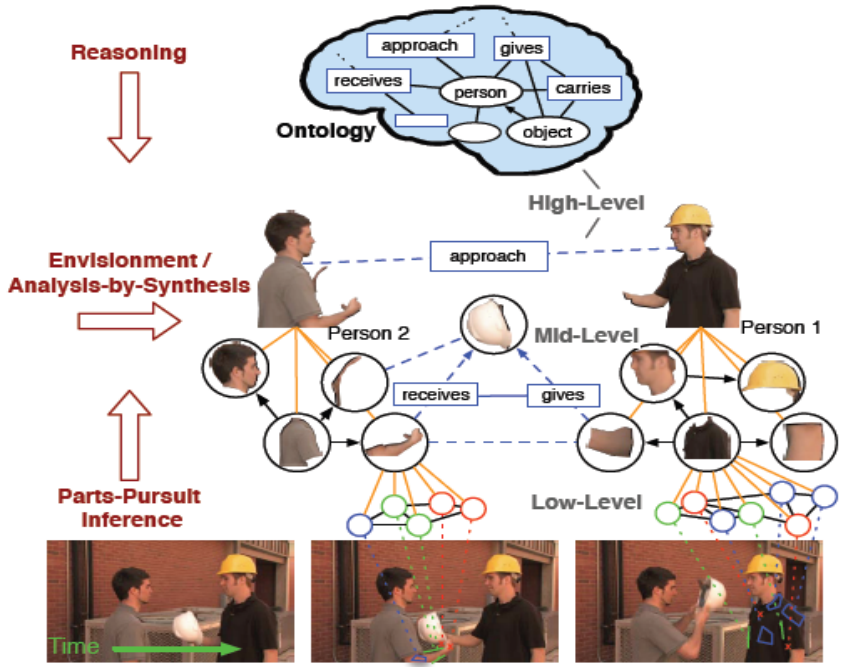

Fig. 1: Interaction of spatio-temporal models at three levels of granularity within the ISTARE platform.

2 Motion Classes

Several formalisms have been introduced to represent and reason with actions. The basic elements of situation calculus [7-8] are: (1) *actions* that can be performed in the world, (2) *fluents* that describe the state of the world, each fluent thus being the representation of some property, and (3) *situations*, a situation being '*a complete state of the universe at an instant of time*' [9], a position which is also maintained in fluent calculus [10]. Event calculus does without situations, and uses only actions and fluents, whereby the latter are functions – rather than predicates as is the case in situation calculus – which can be used in predicates such as *HoldsAt* to state at what time which fluents hold [11]. These approaches, unfortunately, don't take ontological commitment very serious or are based on representational artifacts which do not follow the principles of ontological realism [12].

The Motion Classes (MC) paradigm [13] which builds further on Region Connection Calculus (RCC) [14] to describe motions do not suffer from this. RCC describes how two regions are spatially located in relation to each other, thereby recognizing eight relations (Fig.2). Five of them are ontologically distinct: disconnected (DC), externally connected (EC), partially overlapping (PO), tangential proper part (TPP) and non-tangential proper part (NTPP). Three others are there for

notational purposes: equality (EQ – if we write 'EQ(x,y)', then there is in fact only one region denoted by two distinct symbols 'x' and 'y'), and TPPI and NTPPI as the inverses of TPP and NTPP.

The Motion Classes paradigm exploits what is called the 'conceptual neighborhood' of the RCC-relations which for each relation is defined as the set of possible relations that may hold at some point in time when another relation held earlier (Fig.2). A motion class is the set of transitions from one RCC configuration to another one that can be achieved by the same basic sort of motion (Fig.3). As an example, any transition from PO, TPP NTPP, EQ, TPPI and NTPPI to DC, EC or PO can be achieved through a LEAVE motion, i.e. a motion which separates the two regions from each other. Although there are 64 (8^2) distinct types of transitions which thus theoretically could be caused by 64 distinct types of motions, closer inspection reveals that there are only nine distinct motion types (Fig.3). Five more distinct classes can be defined through pair-wise combination of the nine basic motions: HIT-SPLIT, PERIPHERAL-REACH, PERIPHERAL-LEAVE, REACH-LEAVE, and LEAVE-REACH. The 76 (9^2-5) other combinations do not lead to a distinct sort of motion; HIT followed by REACH, for instance, was already REACH from the very beginning.

In the same way as RCC calculus uses tables to list the possible configurations for region pair (x,z) when the RCC-relations for the pairs (x,y) and (y,z) are known, so provides MC tables for what motion classes are possible for the pair (x,z) when the motion classes for (x,y) and (y,z) are known [13]. MC, in addition to being a representational framework for motion, can also be used as the semantic underpinning for motion verbs. Almost all verbs from Table 1 can be analyzed in terms of a motion class: 'leave' and 'give' involve LEAVE, 'hit' and 'collide' involve HIT, 'bounce' involves HIT-SPLIT, 'approach' involves REACH, and so forth. The feasibility of this approach has already been determined although some further representational frameworks for spatiotemporal reference and direction are required [15]. But, as we will discuss in section 4, an adequate ontological analysis as applied in related contributions to geographic information science [16], is required to determine precisely what sort of involvement is the case.

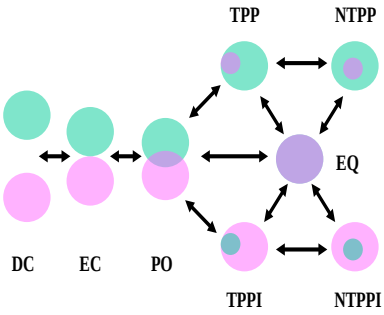

Fig.2: Relationships and transitions involving spatial regions in Region Connection Calculus.

		Ends							
		DC	EC	PO	TPP	NTPP	EQ	TPPI	NTPPI
Starts	DC	Ext	Hit	Reach					
	EC	Split	Periph.						
	PO	Leave		Leave / Reach					
	TPP			Leave	Internal	Expand			
	NTPP								
	EQ						Internal		
	TPPI				Shrink			Internal	
	NTPPI								

Fig.3: 9 basic motion classes representing the simplest type of change that two regions possibly underwent relative to their start and end configuration as expressed in RCC8.

3. Ontological Realism

Ontological realism is a paradigm which rests on a view of the world as consisting of entities of at least two sorts, called 'particulars' and 'universals' respectively [12, 17]. *Particulars*, according to this doctrine, are concrete individual entities that exist in space and time and that exist only once, examples being the persons and helmets depicted in Fig.1. Persons and helmets are *continuants*: whenever they exist, they exist in total. Also the motion in which the white helmet participated ('*participate'* is here a technical term which expresses in the least informative or committing way the relationship between a continuant and the change it undergoes [18]) while being given to the person depicted on the left is a particular. Particulars such as motions are *occurrents*, i.e. entities that at every point in time that they exist, exist only partially.

The word '*depicted'* in the sentence above is not used arbitrarily, because the three bottom images in Fig.1 are themselves three distinct particulars and so are the parts of these pictures which depict the persons. But whereas the persons themselves are not about anything, the corresponding pictures are about these persons. Therefore, the persons are so-called *L1-entities* (first-order entities) while the pictures are *L3-entities*, i.e. communicable representations [19]. It is this communicability that distinguishes L3-entities from cognitive representations (*L2-entities*) such as beliefs, for example the belief sustained by an intelligence analyst that the person on the left in each of these images is the same person, or the belief that this person is John Smith. The analyst can of course express his belief in an annotation to the pictures, that annotation then being an L3-entity and thus clearly distinct from the belief itself: the belief is in the analyst's head, the annotation is in the report.

Universals, in contrast to particulars, are repeatable. This means that they that can exist in indefinitely many instances – thus the persons depicted in Fig.1 were instances of the universal *HUMAN BEING* at the time the pictures (each of the latter being instances of the universal *PICTURE*) were taken – and they are the sorts of things that can be represented by means of general terms used in the formulation of theories, for instance that pictures shot by good cameras contain regions of which the colors correspond with the colors exhibited by the entities in reality to which those regions correspond.

Ontological realism is embodied in two artifacts which roughly correspond with the universal/particular distinction. *Basic Formal Ontology* (BFO) [20] represents the universals which are practically necessary for successful domain and application ontology construction and ensures (1) that there is an unbroken path from every type in the ontology to the ontology's root, and (2) that definitions for all terms in the ontology can be coherently formulated. *Referent Tracking* (RT) provides a set of templates to express formally *portions of reality* (PORs), i.e. how particulars relate to each other, what universals represented in BFO (or ontologies developed there from such as for instance UCORE-SL [21]) they instantiate, and what terms from other terminological systems are used to further describe them [22].

4. Video, spatiotemporal semantics and ontological realism

How do videos of PORs and descriptions about these PORs on the basis of what is depicted in a video, relate to reality under the view of ontological realism?

Digital images taken from PORs contain pixels most of which combine into curves and regions which each have their own shape and texture, all of these entities being continuants. In the ideal case, regions in the image depict (roughly) the characteristics of the surface of the material entities visible to the camera which are all continuants too. Digital video files of PORs are continuants which when processed by display technology lead to the generation of occurrents of which the curves and regions, as well as their shapes and textures, are the only participants. These occurrents are the coming into existence, disappearance, or change in location, shape, size and/or texture of curves and regions. In the ideal case, with an immobile camera and without zooming having been applied, the occurrents visible on the screen correspond to occurrents in which the material entities that are depicted participate. But whereas the on-screen (L3) entities are instances of a very restricted number of universals, there are many more universals of which the corresponding L1-entities are instances. Furthermore, although each particular on-screen entity corresponds (roughly) to exactly one L1-entity, distinct on-screen entities in distinct images or videos may correspond to distinct L1-entities despite being of exactly similar shape, size and texture. Video-analysis can under this view thus be seen as an effort to identify (1) the on-screen regions and their changes which correspond to L1-particulars, (2) the universals instantiated by these L1-particulars, and (3) the identity of these particulars, either absolute (e.g. establishing that John Smith is depicted in the video) or relative (e.g. the person leaving the building is the same as the one that earlier entered the building).

Video-annotation under the Mind's Eye program requires the use of certain descriptive verbs (Table 1) which brings in additional complexity involving L2-entities. Not only is there a wide variability in the way motion classes are linguistically expressed [15], it has also been shown that the cognitive salience of topological relations is not equal for all topologically defined ending relations [23]. Various pitfalls need thus to be avoided. as demonstrated by a verb such as 'to approach'. One pitfall is leaving it open whether a descriptive verb is used to describe an on-screen entity or a first-order entity: although one on-screen entity might indeed be described as approaching another one, it might be such that the corresponding entities in reality are moving away from each other, the on-screen approach being the result of the reduction from 3D to 2D when reality is viewed at through the lens of a camera. Another pitfall is that some motion verbs behave grammatically as action verbs when used in the description of a scene, while in reality the process that is described as such, although ontologically being an instance of motion, is not at all an instance of action: the canoe floating down the river towards the lake is indeed approaching the lake, but without any action going on. Yet another pitfall is that two entities might be described as being involved in an approach although the shortest distance between them increases: think of two cars driving towards each other on a curved road around some mountain. It might be tempting to say that in this case there are two motions going on, one of approaching and one of moving away, but that is of course ontological nonsense. And as a last example, but for sure not the last pitfall,

many motion verbs do not *denote* motions at all, but rather certain configurations of entities in which some sort of motion is involved. 'To approach' is in this case. Imagine a satellite orbiting around Earth for years and that at some point in time a second satellite is launched in an orbit which is such that during some part of their motions the two satellites can be said to approach each other while during other parts they can be said to move away from each other. It seems obvious that the process in which the first satellite participated for all these years does suddenly not become a different process because of some event that does not have any effect on its motion. Yet, the descriptions are valid at the respective times.

Ongoing efforts

Automatically extracting from a video regions that correspond to concrete objects and parts of objects, and then identifying what these objects exactly are, is a challenging problem. Although progress toward object boundary segmentation at the low-level continues to be made, all sufficiently successful approaches are either limited to specific object classes or have not been applied to videos. To overcome these challenges, ISTARE works currently with a hierarchy of pixel aggregates which is induced directly from the pixel data and hence, does not impose an artificial structure. At the low-level, direct pixel intensities are used to decide whether or not to join into an aggregate, and thus do suffer from the above noted limitations. But, as the algorithm moves up the hierarchies, more informative features, such as texture and shape, are used to describe the aggregates and hence assist in deciding which ones should be joined. The aggregation occurs directly on spatiotemporal pixel cubes, defined over short segments of the video (e.g., 2 seconds) [24]. The goal is now to improve the recognition at this level by using information provided by an ontology developed along the lines just sketched.

Acknowledgments. The work described was funded in part by DARPA's Mind's Eye Program and the Army Research Lab under grant W911NF-10-2-0062.

References

1. Cetin, E., *Interim report on progress with respect to partial solutions, gaps in know-how and intermediate challenges of the NoE MUSCLE*. 2005.
2. Viola, P. and M.J. Jones, *Robust real-time face detection*. International Journal on Computer Vision, 2004. **57**(23): p. 137-154.
3. Fei-Fei, L., R. Fergus, and P. Perona., *Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories*, in *IEEE Conference on Computer Vision and Pattern Recognition Workshop on Generative-Model Based Vision*. 2004, IEEE.
4. Everingham, M., et al., *The Visual Object Classes (VOC) Challenge*. International Journal of Computer Vision, 2010. **88**(2): p. 303-338.
5. Biederman, I., *On the Semantics of a Glance at a Scene*, in *Perceptual Organization*, M. Kubovy and K.R. Pomerantz, Editors. 1981, Lawrence Erlbaum Publisher. p. 213-263.

6. Defense Advanced Research Projects Agency. *Mind's Eye Broad Agency Announcement*. 2010 [cited 2010 August 10]; Available from: http://www.darpa.mil/tcto/solicitations/BAA-10-53.html.
7. McCarthy, J., *Situations, actions and causal laws*. 1963, Stanford University Artificial Intelligence Laboratory: Stanford, CA.
8. Reiter, R., *The frame problem in the situation calculus: a simple solution (sometimes) and a completeness result for goal regression*, in *Artificial intelligence and mathematical theory of computation: papers in honour of John McCarthy*, V. Lifshitz, Editor. 1991, Academic Press Professional, Inc: San Diego, CA, USA. p. 359-380.
9. McCarthy, J. and P.J. Hayes, *Some philosophical problems from the standpoint of artificial intelligence.* Machine Intelligence, 1969. **4**: p. 463-502.
10. Thielscher, M., *Introduction to the Fluent Calculus.* Electronic Transactions on Artificial Intelligence, 1998. **2**(3-4): p. 179-192.
11. Kowalski, R., *Database updates in the event calculus.* Journal of Logic Programming, 1992. **12**(1-2): p. 121-46.
12. Smith, B. and W. Ceusters, *Ontological Realism as a Methodology for Coordinated Evolution of Scientific Ontologies.* Journal of Applied Ontology, 2010 (in press).
13. Ibrahim, Z. and A. Tawfik, *An Abstract Theory and Ontology of Motion Based on the Regions Connection Calculus*, in *Abstraction, Reformulation, and Approximation*, I. Miguel and W. Ruml, Editors. 2007, Springer Berlin / Heidelberg. p. 230-242.
14. Randell, D.A., Z. Cui, and A.G. Cohn, *A spatial logic based on regions and connection*, in *Proceedings of the Third International Conference on the Principles of Knowledge Representation and Reasoning*, B. Nebel, W. Swartout, and C. Rich, Editors. 1992, Morgan Kaufmann: Los Altos, CA. p. 165-176.
15. Pustejovsky, J. and J.L. Moszkowicz, *Integrating Motion Predicate Classes with Spatial and Temporal Annotations*, in *Coling 2008: Companion volume: Posters*. 2008, Coling 2008 Organizing Committee: Manchester. p. 95--98.
16. Worboys, M.F. and K. Hornsby, *From objects to events: GEM, the geospatial event model*, in *GIScience : Proceedings of the Third International Conference on GIScience*, M.J. Egenhofer, C. Freksa, and H. Miller, Editors. 2004, Springer Verlag: Berlin. p. 327-44.
17. Smith, B. and W. Ceusters, *Towards Industrial-Strength Philosophy; How Analytical Ontology Can Help Medical Informatics.* Interdisciplinary Science Reviews, 2003. **28**(2): p. 106-111.
18. Smith, B., et al., *Relations in biomedical ontologies.* Genome Biology, 2005. **6**(5): p. R46.
19. Smith, B., et al., *Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain*, in *KR-MED 2006, Biomedical Ontology in Action.* 2006: Baltimore MD, USA
20. Grenon, P. and B. Smith, *SNAP and SPAN: Towards dynamic spatial ontology.* Spatial Cognition and Computation, 2004. **4**(1): p. 69-103.
21. Smith, B., L. Vizenor, and J. Schoening, *Universal Core Semantic Layer*, in *OIC-2009: Ontologies for the Intelligence Community*, P. Costa, K. Laskey, and L. Obrst, Editors. 2009, CEUR: Fairfax, VA.
22. Ceusters, W. and S. Manzoor, *How to track Absolutely Everyting?*, in *Ontologies and Semantic Technologies for the Intelligence Community. Frontiers in Artificial Intelligence and Applications.*, L. Obrst, T. Janssen, and W. Ceusters, Editors. 2010, IOS Press: Amsterdam. p. 13-36.
23. Klippel, A. and R. Li, *The endpoint hypothesis: A topological-cognitive assessment of geographic scale movement patterns.* Conference on Spatial Information Theory (COSIT 2009), 2009: p. 177-194.
24. Corso, J.J., et al., *Efficient multilevel brain tumor segmentation with integrated bayesian model classification.* IEEE Transactions on Medical Imaging, 2008. **27**(5): p. 629-640.