# Using Ontologies in a Cognitive-Grounded System: Automatic Action Recognition in Video Surveillance

Alessandro Oltramari
Department of Psychology
Carnegie Mellon University
Pittsburgh, Pennsylvania 15217
Email: aoltrama@andrew.cmu.edu

Christian Lebiere
Department of Psychology
Carnegie Mellon University
Pittsburgh, Pennsylvania 15217
Email: cl@cmu.edu

***Abstract*—This article presents an integrated cognitive system for automatic video surveillance: in particular, we focus on the task of classifying the actions occurring in a scene. For this purpose, we developed a semantic infrastructure on top of a hybrid computational ontology of actions. The article outlines the core features of this infrastructure, illustrating how the processing mechanisms of the cognitive system benefit from knowledge capabilities in fulfilling the recognition goal. Ultimately, the paper shows that ontologies can enhance a cognitive architecture's functionalities, allowing for high-level performance in complex task execution.**

## I. Introduction

The automatic detection of anomalous and threatening behaviour has recently emerged as a new area of interest in video surveillance: the aim of this technology is to disambiguate the context of a scene, discriminate between different types of human actions, eventually predicting their outcomes. In order to achieve this level of complexity, state-of-the-art computer vision algorithms [1] need to be complemented with higher-level tools of analysis involving, in particular, knowledge representation and reasoning (often under conditions of uncertainty). The goal is to approximate human visual intelligence in making effective and consistent detections: humans evolved by learning to adapt and properly react to environmental stimuli, becoming extremely skilled in filtering and generalizing over perceptual data, taking decisions and acting on the basis of acquired information and background knowledge.

In this paper we first discuss the core features of human 'visual intelligence' and then describe how we can simulate and approximate this comprehensive faculty by means of an integrated framework that augments ACT-R cognitive architecture (see figure 1) with background knowledge expressed by suitable ontological resources (see section III-B2). ACT-R is a modular framework whose components include perceptual, motor and memory modules, synchronized by a procedural module through limited capacity buffers (refer to [2] for more details). ACT-R has accounted for a broad range of cognitive activities at a high level of fidelity, reproducing aspects of human data such as learning, errors, latencies, eye movements and patterns of brain activity. Although it is not our purpose in this paper to present the details of the architecture, two specific mechanisms need to be mentioned here to sketch how the system works: i) *partial matching* - the probability that two different knowledge units (or *declarative chunks*) can be associated on the basis of an adequate measure of similarity (this is what happens when we consider, for instance, that a bag is more likely to resemble to a basket than to a wheel); ii) *spreading of activation* - when the same knowledge unit is part of multiple contexts, it contributes to distributionally activate all of them (like a chemical catalyst may participate in multiple chemical transformations). Section 7 will show in more details how these two mechanisms are exploited by the cognitive system to disambiguate action signals: henceforth, we will refer to this system as the *Cognitive Engine*. As much as humans understand their surroundings coupling perception with knowledge, the *Cognitive Engine* can mimic this capability by leveraging scene-parsing and disambiguation with suitable ontology patterns and models of actions, aiming at identifying relevant actions and spotting the most anomalous ones.

In the next sections we present the different aspects of the *Cognitive Engine*, discussing the general framework alongside specific examples.

## II. The Conceptual Features of Visual Intelligence

The territory of 'visual intelligence' needs to be explored with an interdisciplinary *eye*, encompassing cognitive psychology, linguistics and semantics: only under these conditions can we aim at unfolding the variety of operations that visual intelligence is responsible for, the main characteristics of the emergining representations and, most importantly in the present context, at reproducing them in an artificial agent.

As claimed in [3],"events are understood as action-object couplets" (p. 456) and "segmenting [events as couplets] reduces the amount of information into manageable chunks" (p. 457), where the segment boundaries coincide with achievements and accomplishments of goals (p.460). Segmentation is a key-feature when the task of disambiguating complex scenarios is

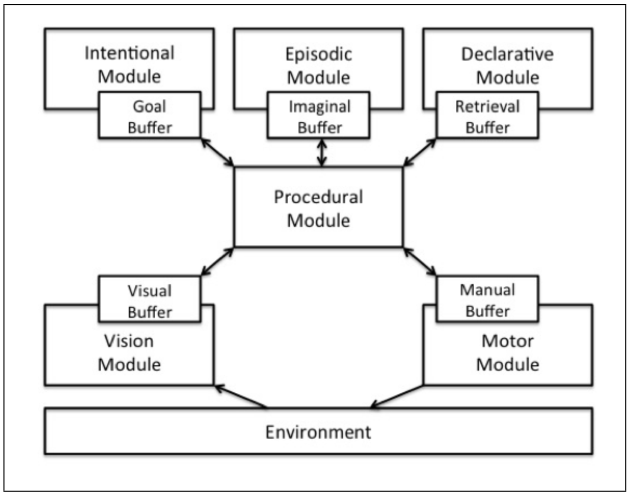

Fig. 1. ACT-R modular structure elaborates information from the environment at different levels.

considered: recognition doesn't correspond to the process of making an inventory of all the actions occurring in a scene: a selection process is performed by means of suitable 'cognitive schemas' (or *gestalts*, e.g. up/down, figure/ground, force, etc.), which carve visual presentations according to principles of mental organization and optimize the perceptual effort" [4]. Besides cognitive schemas, conceptual primitives have also been studied: in particular, [5] applied Hayes' naïve physics theory [6] to build an event logic. Within the adopted common sense definitions, we can mention i) *substantiality* (objects generally cannot pass through one another); ii) *continuity* (objects that diachronically appear in two locations must have moved along the connecting path); iii) *ground plane* (ground acts as universal support for objects).

As far as action-object pairs are central to characterize the 'ontology of events', verb-noun 'frames' are also relevant at the linguistic level[1]; in particular, identifying roles played by objects in a scene is necessary to disambiguate action verbs and highlight the underlying goals. In this respect, studies of event categorization revealed that events are always *packaged*, that is distinctly equipped with suitable semantic roles [8]: for example, the events which are exemplified by motion verbs like walk, run, fly, jump, crawl, etc. are generally accompanied with information about source, path, direction and destination/goal, as in the proposition "John ran out of the house (*source*), walking south (*direction*) along the river (*path*), to reach Emily's house (*destination/goal*)"; conversely, verbs of possession such as have, hold, carry, get, etc. require different kind of semantic information, as in the proposition "John (*owner*) carries Emily's bag (*possession*)". Note that it is not always the case that all possible semantic roles are filled by linguistic phrases: in particular, *path* and *direction* are not necessarily specified when motion is considered, while *source* and *destination/goal* are (we do not focus here on agent and patient which are the core semantic roles).

As this overview suggests, there is an intimate connection between linguistics, cognition and ontology both at the level of scene parsing (mechanism-level) and representation (content-level). In particular, in order to build a visual intelligent system for action recognition, three basic functionalities are required:

- **Ontology pattern matching** - comparing events on the basis of the similarity between their respective pattern components: e.g., *a person's burying an object* and *a person's digging a hole* are similar because they both include some basic body movements as well as the act of removing the soil;
- **Conceptual packaging** - eliciting the conceptual structure of actions in a scene through the identification of the roles played by the detected objects and trajectories: e.g. if you watch *McCutchen hitting an homerun*, the Pittsburgh Pirates' player number 22 is the 'agent', the ball is the patient, the baseball bat is the 'instrument', toward the tribune is the 'direction', etc.).
- **Causal selectivity**: attentional mechanisms drive the visual system in picking the causal aspects of a scene, i.e. selecting the most distinctive actions and discarding collateral or accidental events (e.g., in the above mentioned *homerun* scenario, focusing on the movements of the first baseman is likely to be superfluous).

In the next section we describe how the *Cognitive Engine* realizes the first two functionalites by means of combining the architectural features of ACT-R with ontological knowledge, while **Causal selectivity** will be addressed in future work.

## III. Building the Cognitive Engine

### A. The Context

The *Cognitive Engine* represents the core module of the Extended Activity Reasoning system (EAR) in the CMU-Minds Eye architecture (see figure 2). Mind's Eye is the name of the DARPA program[2] for building AI systems that can filter surveillance footage to support human (remote) operators, and automatically alert them whenever something suspicious is recognized (such as someone leaving a package in a parking lot and running away – see also [9]). In this framework, visual intelligent systems play the role of filtering computer vision data, suitably coupling relevant signals with background knowledge and – when feasible – searching for a 'script' that ties together all the most salient actions in a scene. This comprehensive capability requires intensive information processing at interconnected levels: basic optical features (low-level), object detection (mid-level) and event classification (high-level). EAR has been conceived to deal with the last one: in particular the *Cognitive Engine* receives outputs from the Immediate Activity Recognition module (IAR), which collects the results of different pre-processing algorithms and adopts learning–based methods to output action probability distributions [10].

[1] We refer here to the very broad notion of 'frame' introduced by Minsky: "frames are data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party" [7].

[2] http://www.darpa.mil/Our_Work/I2O/Programs/Minds_Eye.aspx

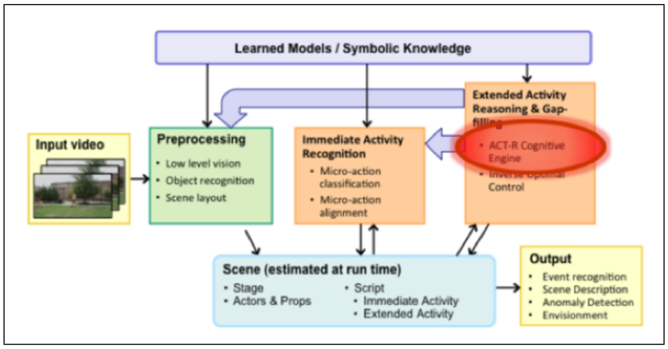

Fig. 2. CMU Mind's Eye architecture

Specific parsing functions are included in EAR to convert the IAR output into sequences of quasi-propositional descriptions of atomic events to be fed to the *Cognitive Engine*. For example, the sample video strip in figure 3 can be converted into *(a)*:

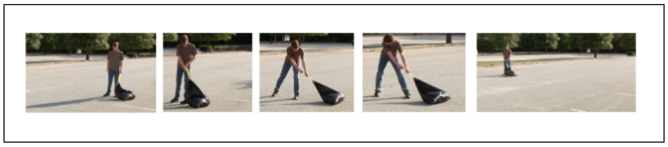

Fig. 3. Significative moments of a composite action

*(a) Person1 Holds Bag2 + Person1 Bends Over + Person1 Drags Bag2 + Person1 Stops.*

These sequences reflect the most likely atomic events (so called 'micro-actions', 'micro-states' and'micro-poses') occurring in the environment, detected and thresholded by machine vision algorithms. The addition symbol exemplifies temporal succession while numbers stand for entity unique identifiers. For the sake of readability, we omit here the temporal information about start and end frames of the single atomic-events, as well as spatial coordinates of the positions of objects. Leveraging the semantic properties of sequences like *(a)*, the *Cognitive Engine* aims at generalizing over action components and distill the most likely 'unifying story': for instance, figure 3 depicts a person hauling an object to the top left side of the scene. Ontology patterns [11] of action play a key-role in the process of sequence disambiguation: in this regard, III-B reviews some of the core patterns we adopted in the recognition mechanisms of the *Cognitive Engine* and outlines the basic classes and properties of the ontology of actions used for high-level reasoning. The benefits of using ontologies for event recognition in the context of the Mind's Eye program have been also discussed in [12], although our two approaches differ both in the theoretical underpinnings (as the next sections will show, we propose a hybridization of linguistic and ontological distinctions rather than embracing ontological realism) and in the general system design (in [12] the authors outline a framework in which ontological knowledge is directly plugged into visual algorithms, while in our proposal ACT-R is exploited as an intermediate module to bridge the vision and the knowledge levels, stressing the role of cognitive mechanisms in action understanding).

## *B. The Knowledge Infrastructure*

*1) Ontology patterns of actions:* In recent years,'Ontology Design Patterns' (or just 'ontology patterns') have become an important resource in the areas of Conceptual Modeling and Ontology Engineering: the rationale is to identify some minimal conceptual structures to be used as the *building blocks* for designing ontologies [13]. Ontology patterns are small models of entities and their basic properties: the notion originates in [14], where the author argues that a good (architectural) design can be achieved by means of a set of rules that are packaged in the form of patterns, such as 'windows place', or 'entrance room'. Design patterns are then assumed as archetypal solutions to design problems in a certain context. Ontology patterns are built and formalized on the basis of a preliminary requirement analysis, which can be driven either by applications tasks or by specific problems in the domain of interest. In our context, ontology patterns enable the classification of actions by means of pinpointing the basic semantic roles and constituent atomic events of relevant actions. In these regards, table I shows the composition of the core ontology patterns used in the *Cognitive Engine*: e.g. an instance of the action-type 'pick-up' depends on the occurrence of at least four basic components (C1-C4), namely 'bend-over', 'lower-arm', 'stand-up' (necessary body-movements) and 'holding' (referring to the interaction between a person and an object); moreover, those action-verbs require specific conceptual roles to be exemplified, respectively, *protagonist* for the first and the third component, *agent* for the second and the fourth (which includes also 'patient' as object-role). But what did inspire our modeling choices? How could we identify those roles and atomic events? Which rules/principles allowed us to assemble them in that very fashion? In order to answer to these questions, in the next section we introduce HOMINE, the Hybrid Ontology for the Mind's Eye project.

*2) Ontology of actions:* Ontologies play the role of 'semantic specifications of declarative knowledge' in the framework of cognitive architectures [15]. As [16], [17], [18], [19] demonstrate, most research efforts have focused on designing methods for mapping large knowledge bases to the ACT-R declarative module. Here we commit on taking a different approach: instead of tying to a single monolithic large knowledge base, we built a hybrid resource that combines different semantic modules, allowing for high scalability and interoperability. Our proposal consists in suitably linking distinctive lexical databases, i.e. WordNet [20] and FrameNet [21] with a computational ontology of actions, plugging the obtained semantic resource in the dynamic mechanisms of the ACT-

TABLE I

ONTOLOGY PATTERNS OF ACTIONS FOR THE COGNITIVE ENGINE

| Action | Role1 | Role2 | Role3 | Role4 | Object | C1 | C2 | C3 | C4 |
|---|---|---|---|---|---|---|---|---|---|
| Arrive | self-mover | theme | | | | walk | stop | | |
| Leave | self-mover | theme | | | | walk | exit | | |
| Give | agent | carrier | agent | | patient | holding | transport | drop | |
| Take | carrier | agent | agent | | patient | transport | drop | holding | |
| Exchange | agent | agent | agent | | patient | give | take | swap | |
| Carry | agent | carrier | agent | | patient | holding | transport | pull | |
| Pick-up | protagonist | agent | protagonist | agent | patient | bend-over | lower-arm | stand-up | holding |
| Put-down | agent | protagonist | agent | figure1 | patient | holding | bend-over | lower-arm | on |
| Bury | protagonist | agent | protagonist | agent | patient | bend-over | lower-arm | fill-with-tool | stand-up |
| Dig | protagonist | agent | agent | protagonist | patient | bend-over | lower-arm | dig-with-tool | stand-up |
| Haul | protagonist | agent | agent | agent | patient | bend-over | extend-arm | holding | drag |

R cognitive architecture (see IV). Accordingly, HOMINE is built on the top-level of DOLCE-SPRAY [22], a simplified version of DOLCE [23]: we used DOLCE-SPRAY as a general model for aligning WordNet (WN) and FrameNet (FN) – following the line of research of [24]: figure 4 shows some selected nodes of DOLCE backbone taxonomy. The root of the hierarchy of DOLCE-SPRAY is ENTITY, which is defined as anything which is identifiable by humans as an object of experience or thought. The first distinction is among CONCRETE-ENTITY, i.e. objects located in definite spatial regions, and ABSTRACT-ENTITY, whose instances don't have spatial properties. In the line of [25], CONCRETE-ENTITY is further split in CONTINUANT and OCCURRENT, namely entities without inherent temporal parts (e.g. artifacts, animals, substances) and entities with inherent temporal parts (e.g. events, actions, states) respectively. The basic ontological distinctions are maintained: DOLCE's ENDURANT and PERDURANT match DOLCE-SPRAY's CONTINUANT and OCCURRENT. The main difference of DOLCE-SPRAY's top level with respect to DOLCE, is the merging of DOLCE's ABSTRACT and NON-PHYSICAL-ENDURANT categories into the DOLCE-SPRAY's category of ABSTRACT-ENTITY. Among abstract entities, DOLCE-SPRAY's top level distinguishes CHARACTERIZATION, defined as mapping of *n*-uples of individuals to truth values. Individuals belonging to CHARACTERIZATION can be regarded to as 'reified concepts', and the irreflexive, antisymmetric relation CHARACTERIZE associates them with the objects they denote. Whether CHARATERIZATION is formally a metaclass, and whether CHARACTERIZE bears the meaning of set membership is left opaque in this ontology.

HOMINE's linguistic-semantic layer is based on a partition of WN related to verbs of action, such as 'haul', 'pick-up', 'carry', 'arrive', 'bury' etc. WN is a semantic network whose nodes and arcs are, respectively, synsets ("sets of synonym terms") and semantic relations. Over the years, there has been an incremental growth of the lexicon (the latest version, WordNet 3.0, contains about 120K synsets), and substantial enhancements aimed at facilitating computational tractability. In order to find the targeted group of relevant synsets, we basically started from two pertinent top nodes[3], move #1 and move#2[4]. As one can easily notice, the former synset denotes a change of position accomplished by an agent or by an object (with a sufficient level of autonomy), while the latter is about causing someone or something to move (both literally and figuratively). After extracting the sub–hierarchy of synsets related to these generic verbs of action, we introduced a topmost category 'movement-generic', abstracting from the two senses of 'move' (refer to figure 5 for the resulting taxonomy of actions).

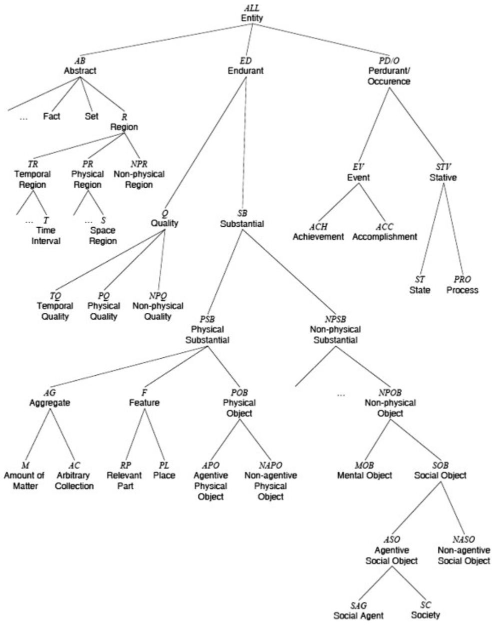

Fig. 4. An excerpt of DOLCE-SPRAY top level

FrameNet (FN) is the additional conceptual layer of HOMINE. Besides wordnet-like databases, a computational lexicon can be designed from a different perspective, for example focusing on frames, to be conceived as orthogonal

[3] AKA Unique Beginners (Fellbaum 1998).

[4] 01835496 move#1, travel#1, go#1, locomote#1 (change location; move, travel, or proceed) "How fast does your new car go?"; "The soldiers moved towards the city in an attempt to take it before night fell". 01850315 move#2, displace#4 (cause to move or shift into a new position or place, both in a concrete and in an abstract sense) "Move those boxes into the corner, please"; "The director moved more responsibilities onto his new assistant".

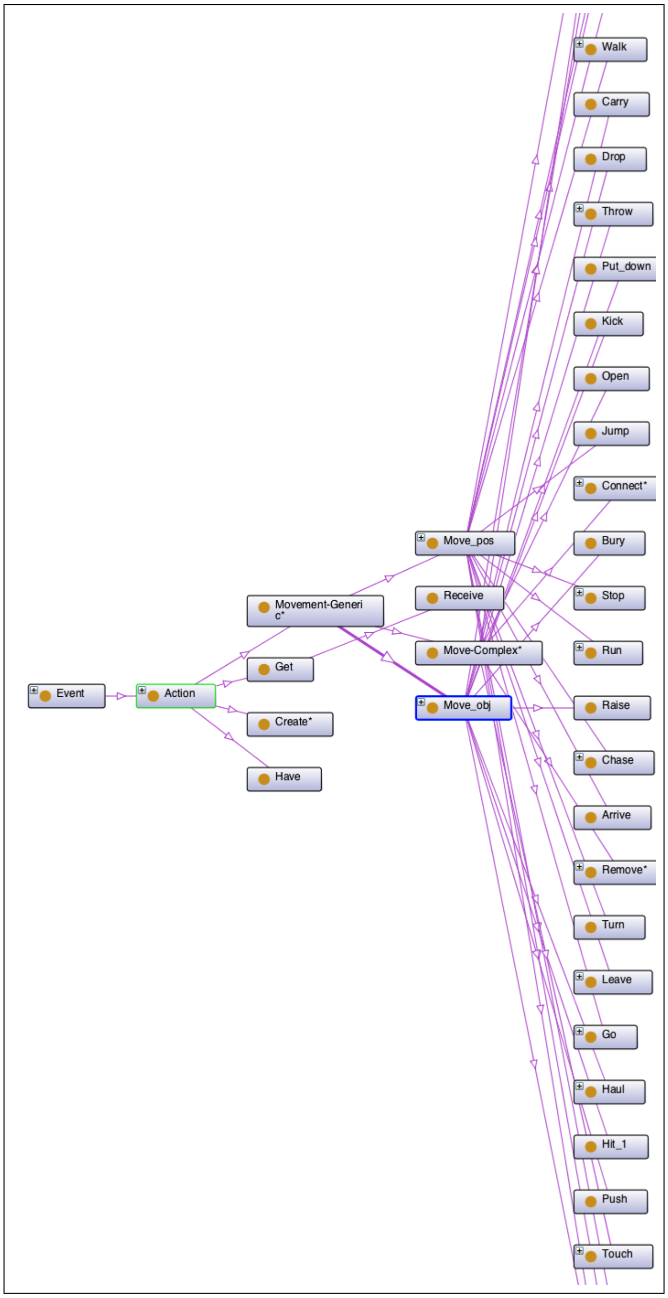

Fig. 5. An excerpt of HOMINE backbone taxonomy

to domains. Inspired by frame semantics [26], FN aims at documenting "the range of semantic and syntactic combinatory possibilities (valences) of each word in each of its senses" through corpus-based annotation. Different frames are evoked by the same word depending on different contexts of use: the notion of 'evocation' helps in capturing the multi-dimensional character of knowledge structures underlying verbal forms. For instance, if you consider the *bringing* frame, namely an abstraction of a state of affairs where sentient agents (e.g., persons) or generic carriers (e.g. ships) bring something somewhere along a given path, you will find several 'lexical units' (LUs) evoking different roles (or frame elements - FEs): i.e., the noun 'truck' instantiates the 'carrier' role. In principle, the same Lexical Unit (LU) may evoke distinct frames, thus dealing with different roles: 'truck', for example, can be also associated to the vehicle frame ('the vehicles that human beings use for the purpose of transportation'). FN contains about 12K LUs for 1K frames annotated in 150000 sentences. WN and FN are based on distinct models, but one can benefit from the other in terms of coverage and type of information conveyed. Accordingly, we have analyzed the evocation-links between the action verbs we have extracted from WN and the related FN frames: those links can be generated through 'FN Data search', an on–line navigation interface used to access and query FN[5]. Using a specific algorithm [27], WordNet synsets can be associated with FrameNet frames, ranking the results by assigning weights to the discovered connections [28]. The core mechanism can be resumed by the following procedure: first of all the user has to choose a term and look for the correspondent sense in WordNet; once the correct synset is selected, the tool searches for the corresponding lexical units (LUs) and frames of FrameNet. Afterwards, all candidate frames are weighted according to three important factors: the similarity between the target word (the LU having some correspondence to the term typed at the beginning) and the wordnet relative (which can be the term itself - if any - and/or its synonyms, hypernyms and antonyms); a variable boost factor that rewards words that correspond to LU as opposed to those that match only the frame name; the spreading factor, namely the number of frames evoked by that word:

$$\boxed{\frac{similarity(wordnet\_relative,target\_word)*BoostFactor}{spreading\_factor(wordnet\_relative)}}$$

If DOLCE-SPRAY provides the axiomatic basis for the formal characterization of HOMINE[6], and WN and FN computational lexicons populate the ontology with linguistic knowledge, SCONE is the selected framework of implementation[7]. SCONE is an open–source knowledge-base system intended for use as a component in many different software applications: it provides a LISP-based framework to represent and reason over symbolic common–sense knowledge. Unlike most diffuse KB systems, SCONE is not based on OWL (Ontology Web Language[8]) or Description Logics in general [30]: its inference engine adopts marker–passing algorithms [31] (originally designed for massive parallel computing) to perform fast queries at the price of losing logical completeness and decidability. In particular, SCONE represents knowledge as a *semantic network* whose nodes are locally weighted (*marked*) and associated to arcs (*wires*[9]) in order to optimize basic reasoning tasks (e.g. class membership, transitivity, inheritance

[5] https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=luIndex

[6] For instance, DOLCE adapts Allen's temporal axioms [29], which are considered as state of the art in temporal representation and reasoning.

[7] http://www.cs.cmu.edu/~sef/scone/

[8] http://www.w3.org/TR/owl-features/

[9] In general, a *wire* can be conceived as a binary relation whose domain and range are referred to, respectively, as A-node and B-node.

of properties, etc. ). The philosophy that inspired SCONE is straightforward: from vision to speech, humans exploit the brain's massive parallelism to fulfill all recognition tasks; if we want to build an AI system which is able to deal with the large amount of knowledge required in common-sense reasoning, we need to rely on a mechanism which is fast and effective enough to simulate parallel search. Accordingly, SCONE implementation of marker–passing algorithms aims at simulating a pseudo-parallel search by assigning specific marker bits to each knowledge unit. For example, if we want to query a KB to get all the parts of cars, SCONE would assign a marker M1 to the A-node CAR and search for all the statements in the knowledge base where M1 is the A-wire (domain) of the relation PART-OF , returning all the classes in the range of the relation (also called 'B-nodes'). SCONE would finally assign the marker bit M2 to all B-nodes, also retrieving all the inherited subclasses[10]. The modularization and implementation of HOMINE with SCONE allows for an effective formal representation and inferencing of core ontological properties of events, such as: i) participation of actors and objects in actions; ii) temporal features based on the notions of 'instant' and 'interval'; iii) common-sense spatial information.

The *Cognitive Engine* is the result of augmenting ACT-R with HOMINE: in general we refer to ACT-R including the SCONE extra-module as ACT-RK, meaning 'ACT-R with improved Knowledge capabilities' (the reader can easily notice the evolution from the original ACT-R architecture – figure 1 – to the knowledge-enabled one – figure 6). We engineered a SCONE-MODULE as a bridging component between the cognitive architecture and the knowledge resource: this integration allows for dynamic queries to be automatically submitted to HOMINE by ACT-RK whenever the visual information is incomplete, corrupted or when reasoning with common-sense knowlege is needed to generalize over actor and actions in a scene. In this way, the *Cognitive Engine* is able to overcome situations with missing input: ACT-R mechanisms of partial matching and spreading activation [2] can fill the gap(s) left by the missing atomic events and retrieve the best–matching ontology pattern. In the last section of the paper we describe how *Cognitive Engine* performs action-recognition task for the example orginally sketched in figure 3.

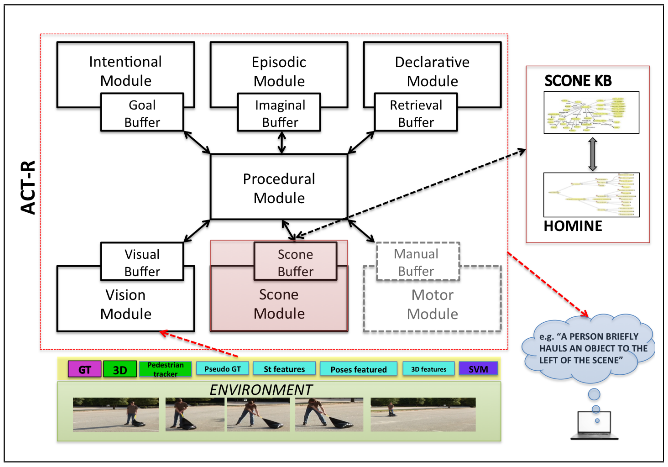

Fig. 6. The *Cognitive Engine*

## IV. Using the *Cognitive Engine* for action recognition: an example

In the context of the Mind's Eye program, a visual intelligent systems is considered to be successful if it is able to process a video-dataset of actions[11] and output the probability distribution (per video) of a pre-defined list of verbs, including 'walk', 'run', 'carry', 'pick-up', 'haul', 'follow', 'chase', etc[12]. Performance is measured in terms of consistency with human responses to stimuli (*Ground-Truth*): subjects have to acknowledge the presence/absence of every verb in each video. In order to meet these requirements, we devised the *Cognitive Engine* to work in a human-like fashion (see section II), trying to disambiguate the scene in terms of the most reliable conceptual structures. Because of space limitations, we can't provide here the details of a large-scale evaluation: nevertheless, we can discuss the example depicted earlier in the paper (figure 3) in light of the core mechanisms of the *Cognitive Engine*. Considering figure 7, the *Cognitive Engine* parses the atomic events extracted by IAR, namely 'hold' (micro-state) and 'bend-over', 'drag', 'stop' (micro-actions), associating frames and roles to visual input from the videos. This specific information is retrieved from the FrameNet module of HOMINE: frames and frame roles are assembled in suitable knowledge units and encoded in the declarative memory of ACT-RK. As with human annotators performing semantic role labeling [32], the *Cognitive Engine* associates verbs denoting atomic events to corresponding frames. When related mechanisms are activated, the *Cognitive Engine* retrieves the roles played by the entities in the scene, for each atomic event: for example, 'hold' evokes the *manipulation* frame, whose core role *agent* can be be associated to 'person1' (as showed in light-green box of the figure). In order to prompt a choice within the available ontology patterns of action (see table I), sub-symbolic computations for *spreading activation* are executed [2]. Spreading of activation from the contents of frames and roles triggers the evocation of related ontology patterns. As mentioned in the introduction, *partial matching* based on similarity measures and *spreading of activation* based on compositionality are the main mechanisms used by *Cognitive Engine*: in particular, we constrained semantic similarity within verbs to the 'gloss-vector' measure computed over WordNet synsets [33]. Base-level activations of verbs actions have been derived by frequency analysis of the American National Corpus: in particular, this choice reflects the fact that the more frequent is a verb, the more is likely to be activated

[10] Far from willing to deepen a topic that is out of scope to treat in this manuscript, we refer the reader to [31] for details concerning marker–passing algorithms.

[11] http://www.visint.org/datasets.html.

[12] This list has been provided in advance by DARPA.

by a recognition system. Additionally, strengths of associations are set (or learned) by the architecture to reflect the number of patterns to which each atomic event is associated, the so-called 'fan effect' controlling information retrieval in many real-world domains [34].

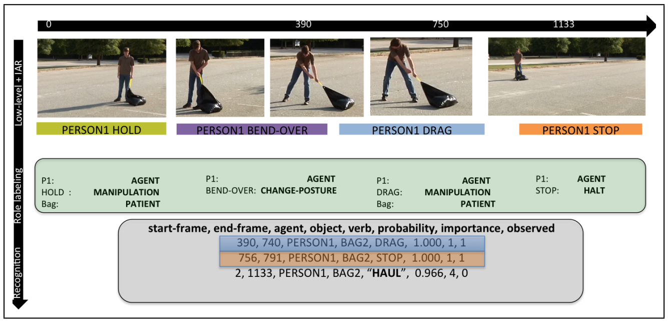
0
390
750
1133
Low-level + IAR
Role labeling
Recognition
PERSON1 HOLD
PERSON1 BEND-OVER
PERSON1 DRAG
PERSON1 STOP
P1: AGENT
HOLD : MANIPULATION
Bag: PATIENT
P1: AGENT
BEND-OVER: CHANGE-POSTURE
P1: AGENT
DRAG: MANIPULATION
Bag: PATIENT
P1: AGENT
STOP: HALT
start-frame, end-frame, agent, object, verb, probability, importance, observed
390, 740, PERSON1, BAG2, DRAG, 1.000, 1, 1
756, 791, PERSON1, BAG2, STOP, 1.000, 1, 1
2, 1133, PERSON1, BAG2, "HAUL", 0.966, 4, 0

Fig. 7. A Diagram of the Recognition Task performed by the *Cognitive Engine*. The horizontal black arrow represents the sequence time framing while the vertical one represents the interconnected levels of information processing. The light-green box displays the results of semantic disambiguation of the scene elements, while the gray box contains the schema of the output, where importance reflects the number of components ina detected pattern (1-4) and *observed* is a boolean parameter whose value is 1 when a verb matches an IAR detection and 0 when the verbs is an actual result of EAR processing.

The core sub-symbolic computations performed by the *Cognitive Engine* through ACT-RK can be expressed by the equation in figure 8:

$$\boxed{A_i = \ln \sum_j t_j^{-d} + \sum_k W_k S_{ki} + \sum_l MP_l Sim_{li} + N(0,\sigma)}$$

Fig. 8. Equation for Bayesian Activation Pattern Matching

- **1st term**: the more recently and frequently a chunk $i$ has been retrieved, the higher its activation and the chances of being retrieved. In our context $i$ can be conceived as a pattern of action (e.g., the pattern of HAUL), where $t_j$ is the time elapsed since the $j^{th}$ reference to chunk $i$ and $d$ represents the memory decay rate.
- **2nd term**: the contextual activation of a chunk $i$ is set by the attentional weight $S_{\mathrm{ki}}$ given the element $k$, the element $i$ and the strength of association between an element $k$ and the $i$. In our context, $k$ can be interpreted as the value BEND-OVER of the pattern HAUL in figure 7.
- **3rd term**: under partial matching, ACT-RK can retrieve the chunk $l$ that matches the retrieval constraints $i$ to the greatest degree, computing the similarity $Sim_{li}$ between $l$ and $i$ and the mismatch score MP (a negative score that is assigned to discriminate the 'distance' between two terms). In our context, for example, the value PULL could have been retrieved, instead of DRAG. This mechanism is particularly useful when verbs are continuosly changing - as in the case of a complex visual input stream.
- **4th term**: randomness in the retrieval process by adding Gaussian noise.

Last but not least, the *Cognitive Engine* can output the results of extra-reasoning functions by means of suitable queries submitted to HOMINE via the SCONE-MODULE. In the example in figure 7, object classifiers and tracking algorithms could not detect that 'person1' is dragging 'bag2' by pulling a rope: this failure in the visual algorithms is motivated by the fact that the rope is a very tiny and morphologically unstable artifact, hence difficult to be spotted by state-of-the-art machine vision. Nevertheless, HOMINE contains an axiom stating that:
"For every $x$,$y$,$e$,$z$ such that P($x$) is a person, GB($y$) is a Bag and DRAG($e$,$x$,$y$,$T$) is an event $e$ of type DRAG (whose participants are $x$ and $y$) occurring in the closed interval of time $T$, there is at least a $z$ which is a proper part of $y$ and that participates to $e$"[13].
Moreover, suppose that in a continuation of the video, the same person drops the bag, gets in a car and leaves the scene. The visual algorithms would have serious difficulties in tracking the person while driving the car, since the person would become partially occluded, assume an irregular shape and would be no more properly lightened. Again, the *Cognitive Engine* could overcome these problems in the visual system by using SCONE to call HOMINE and automatically perform the following schematized inferences:

- Cars move;
- Every car needs exactly one driver to move[14];
- Drivers are persons;
- A driver is located inside a car;
- If a car moves then the person driving the car also moves in the same direction.

Thanks to the inferential mechanisms embedded in its knowedge infrastructure, the *Cognitive Engine* is not bound to visual input as an exclusive source of information: in a human-like fashion, the *Cognitive Engine* has the capability of coupling visual signals with background knowledge, performing high-level reasoning and disambiguating the original input perceived from the environment. In this respect, the *Cognitive Engine* can be seen as exemplifying a general perspective on artificial intelligence, where data-driven learning mechanisms are integrated in a knowledge–centered reasoning framework.

## V. Conclusion

In this paper we presented the knowledge infrastructure of a high-level artificial visual intelligent system, the *Cognitive Engine*. In particular we described how the conceptual specifications of basic action types can be driven by an hybrid semantic resource, i.e. HOMINE and its derived ontology patterns: for each considered action verb, the *Cognitive Engine* can identify typical FrameNet roles and corresponding lexical fillers (WordNet synsets), logically constraining them

[13] Note that here we are paraphrasing an axiom that exploits Davidsonian event semantics [35] and basic principles of formal mereology (see [25] and [36]). Also, this axiom is valid if every bag has a rope: this is generally true when considering garbage bags like the one depicted in figure7, but exceptions would need to be addressed in a more comprehensive scenario.

[14] With some exceptions, especially in California, around Mountain View!

to a computational ontology of actions encoded in ACTR-K through the SCONE Knowledge-Base system. Future work will be devoted to improve the *Cognitive Engine* and address *causal selectivity* (see II) using (1) reasoning and statistical inferences to derive and predict goals of agents and (2) mechanisms of abduction to focus on the most salient information from complex visual streams. We also plan to extend the system functionalities in order to support a wider range of action verbs and run tests on a large video dataset.

## Acknowledgments

This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0061. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

## References

[1] D. A. Forsyth and J. Ponce, *Computer Vision, A Modern Approach*. Prentice Hall, 2004.

[2] J. Anderson and C. Lebiere, *The Atomic Components of Thought*. Erlbaum, 1998.

[3] B. Tversky, J. Zachs, and B. Martin, "The structure of experience," in *Understanding events: From Perception to Action*, T. Shipley and T. Zacks, Eds., 2008, pp. 436–464.

[4] L. Albertazzi, L. Van Tonder, and D. Vishwanath, Eds., *Perception Beyond Inference. The Information Content of Visual Processes*. The MIT Press, 2010.

[5] J. M. Siskind, "Grounding language in perception," *Artificial Intelligence Review*, vol. 8, pp. 371–391, 1995.

[6] P. J. Hayes, "The second naïve physics manifesto," in *Formal Theories of the Common Sense World*, J. Hobbes and R. Moore, Eds. Ablex Publishing Corporation, 1985.

[7] M. Minsky, "A framework for representing knowledge," in *Mind Design*, P. Winston, Ed. MIT Press, 1997, pp. 111–142.

[8] A. Majid, J. Boster, and M. Bowerman, "The cross-linguistic categorization of everyday events: a study of cutting and breaking," *Cognition*, vol. 109, pp. 235–250, 2008.

[9] P. W. Singer, *Wired for War*. The Penguin Press, 2009.

[10] P. Maitikanen, R. Sukthankar, and M. Hebert, "Feature seeding for action recognition," in *Proceedings of International Conference on Computer Vision*, 2011.

[11] M. Poveda, M. C. Suarez-Figueroa, and A. Gomez-Perez, "Ontology analysis based on ontology design patterns," in *WOP 2009 Workshop on Ontology Patterns at the 8th International Semantic Web Conference (ISWC 2009). Proceedings of the WOP 2009.*, W. . . W. on Ontology Patterns at the 8th International Semantic Web Conference (ISWC 2009), Ed. WOP 2009 Workshop on Ontology Patterns at the 8th International Semantic Web Conference (ISWC 2009), 2009. [Online]. Available: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-516/pap05.pdf

[12] W. Ceusters, J. Corso, Y. Fu, M. Petropoulos, and V. Krovi, "Introducing ontological realism for semi-supervised detection and annotation of operationally significant activity in surveillance videos," in *the 5th International Conference on Semantic Technologies for Intelligence, Defense,and Security (STIDS 2010)*, 2010.

[13] A. Gangemi and V. Presutti, "Ontology design patterns," in *Handbook on Ontologies*, ser. 2nd Edition, S. Staab and R. Studer, Eds. Springer, 2009.

[14] C. Alexander, *The Timeless Way of Building*. Oxford Press, 1979.

[15] A. Oltramari and C. Lebiere, "Mechanism meet content: Integrating cognitive architectures and ontologies," in *Proceedings of AAAI 2011 Fall Symposium of "Advances in Cognitive Systems"*, 2011.

[16] J. Ball, S. Rodgers, and K. Gluck, "Integrating act-r and cyc in a large-scale model of language comprehension for use in intelligent systems," in *Papers from the AAAI workshop*. AAAI Press, 2004, pp. 19–25.

[17] S. Douglas, J. Ball, and S. Rodgers, "Large declarative memories in act-r," in *Proceedings of the 9th International Conference of Cognitive Modeling*, 2009.

[18] B. Best, N. Gerhart, and C. Lebiere, "Extracting the ontological structure of cyc in a large-scale model of language comprehension for use in intelligent agents," in *Proceedings of the 17th Conference on Behavioral Representation in Modeling and Simulation*, 2010.

[19] B. Edmond, "Wn-lexical: An act-r module built from the wordnet lexical database," in *Proceedings of the 7th International Conference of Cognitive Modeling*, 2006, pp. 359–360.

[20] C. Fellbaum, Ed., *WordNet, An Electronic Lexical Database*. MIT Press, Boston, 1998.

[21] J. Ruppenhofer, M. Ellsworth, M. Petruck, and C. Johnson, "Framenet: Theory and practice," June 2005.

[22] G. Vetere, A. Oltramari, I. Chiari, E. Jezek, L. Vieu, and F. M. Zanzotto, "Senso comune, an open knowledge base for italian," *TAL - Traitement Automatique des Langues*, vol. 39, no. Forthcoming, 2012.

[23] C. Masolo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider, "WonderWeb Deliverable D17: The WonderWeb Library of Foundational Ontologies," Tech. Rep., 2002.

[24] A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari, "Sweetening wordnet with dolce," *AI Magazine*, vol. 3, pp. 13–24, Fall 2003.

[25] P. Simons, Ed., *Parts: a Study in Ontology*. Clarendon Press, Oxford, 1987.

[26] C. J. Fillmore, "The case for case," in *Universals in Linguistic Theory*, E. Bach and T. Harms, Eds. New York: Rinehart and Wiston, 1968.

[27] A. Burchardt, K. Erk, and A. Frank, "A wordnet detour to framenet," in *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen.*, ser. Computer Studies in Language and Speech, B. S. Bernhard Fisseni, Hans-Christian Schmitz and P. Wagner, Eds. Frankfurt am Main: Peter Lang, 2005, vol. 8, pp. 408–421.

[28] A. Oltramari, "Lexipass methodology: a conceptual path from frames to senses and back," in *LREC 2006 (Fifth International Conference on Language Resources and Evaluation)*. Genoa (Italy): ELDA, 2006.

[29] J. F. Allen, "An interval based representation of temporal knowledge," in *7th International Joint Conference on Artificial Intelligence*. Vancouver: IJCAI, Morgan Kaufmann, 1983, pp. 221–226, vol.1.

[30] F. Baader, D. Calvanese, D. L. Mcguinness, D. Nardi, and P. F. Patel-Schneider, Eds., *The Description Logic Handbook : Theory, Implementation and Applications*. Cambridge University Press, 2003.

[31] S. Fahlman, "Using scones multiple-context mechanism to emulate human-like reasoning," in *First International Conference on Knowledge Science, Engineering and Management (KSEM'06)*. Guilin, China: Springer–Verlag (Lecture Notes in AI), 2006.

[32] D. Gildea and D. Jurafsky, "Automatic labelling of semantic roles," in *Proceedings of 38 th Annual Conference of the Association for Computational Linguistics (ACL-00)*, 2000, pp. 512–520.

[33] T. Pedersen, S. J. Patwardhan, and M. Michelizzi, "Wordnet :: Similarity: Measuring the relatedness of concepts," in *Demonstration Papers at HLT-NAACL*, 2004, pp. 38–41.

[34] L. Schooler and J. Anderson, "The disruptive potential of immediate feedback," in *Proceedings of the Twelfth Annual Conference of The Cognitive Science Society*, 1990, pp. 702–708.

[35] R. Casati and A. Varzi, Eds., *Events*. Aldershots, USA: Dartmouth, 1996.

[36] R. Casati and A. Varzi, *Parts and Places. The Structure of Spatial Representation*. Cambridge, MA: MIT Press, 1999.