A Neural-Symbolic Cognitive Agent with a Mind's Eye

H.L.H. de Penning[1], R.J.M. den Hollander[2], H. Bouma[2], G.J. Burghouts[2], A.S. d'Avila Garcez[3]

[1]TNO Behavioural and Societal Sciences, Soesterberg, NL, leo.depenning@tno.nl

[2]TNO Technical Sciences, Den Haag, NL, richard.denhollander@tno.nl, henri.bouma@tno.nl, gertjan.burghouts@tno.nl

[3]Department of Computing, City University, London, UK, aag@soi.city.ac.uk

Abstract

The DARPA Mind's Eye program seeks to develop in machines a capability that currently exists only in animals: visual intelligence. This paper describes a Neural-Symbolic Cognitive Agent that integrates neural learning, symbolic knowledge representation and temporal reasoning in a visual intelligent system that can reason about actions of entities observed in video. Results have shown that the system is able to learn and represent the underlying semantics of the actions from observation and use this for several visual intelligent tasks, like recognition, description, anomaly detection and gap-filling.

Introduction

The DARPA Mind's Eye program seeks to develop in machines a capability that currently exists only in animals: visual intelligence (Donlon, 2010). In particular, this program pursues the capability to learn generally applicable and generative representations of action between objects in a scene, directly from visual inputs, and then reason over those learned representations. A key distinction between this research and the state of the art in machine vision is that the latter has made continual progress in recognizing a wide range of objects and their properties, what might be thought of as the nouns in the description of a scene. The focus of Mind's Eye is to add the perceptual and cognitive underpinnings for recognizing and reasoning about the verbs in those scenes, enabling a more complete narrative of action in the visual experience.

The Neural-Symbolic Cognitive Agent (NSCA) is a cognitive model that is able to: perform learning of complex temporal relations from uncertain observations, reason probabilistically about the knowledge that has been learned, and represent the agent's knowledge in a logic-based format (de Penning, Garcez, Lamb, & Meyer, 2011).

This paper describes the application and results of the NSCA as part of a Visual Intelligence (VI) system, called CORTEX that is able to learn and reason about actions in order to; i) recognize actions based on properties of detected objects, ii) describe these actions in natural language, iii) detect anomalies and iv) fill-in gaps (e.g. video blackouts by missing frames, occlusion by moving objects, or entities receding behind objects).

A key requirement of visual intelligence in the context of Mind's Eye is the ability to describe the cognitive underpinnings of the reasoning process. Our objective is not only to provide accurate classifications but to provide descriptions of the process that can help explain the reasoning and shed new light into this important aspect of cognition. Hence, the use of NSCA with its knowledge extraction capacity will provide the descriptions that are required.

Learning and Reasoning

The NSCA (depicted in Figure 1) uses a Recurrent Temporal Restricted Boltzmann Machine (RTRBM) that can encode temporal relations as a joint probability distribution on beliefs B (represented by the visible units), hypotheses H (represented by the hidden units) and their probabilities in the previous time H^{t-1} (represented by recurrent connections in the hidden units).

With the RTRBM, **deduction** is similar to Bayesian inference, where for all hypotheses H, the probability is calculated that the hypotheses are true given the observed beliefs b and the previously applied hypotheses H^{t-1} (i.e. $P(H|B=b,H^{t-1})$). From this posterior probability distribution, the RTRBM selects the most likely hypotheses h using random Gaussian sampling, i.e. $h \sim P(H|B=b,H^{t-1})$. Via **abduction** the RTRBM then infers the most likely beliefs based on h by calculating the conditional probability (i.e. $P(B|H=h)$). The differences between the observed and inferred beliefs are then used by the NSCA to determine the implications of the applied hypotheses, which are, in case of CORTEX, the recognized actions. **Induction** of new relations can be obtained by using these differences to improve the correlation between the selected hypotheses h and the observed beliefs b. It does so by updating the weights in the RTRBM using Contrastive Divergence and Backpropagation-Through-Time (Sutskever & Hinton, 2008).
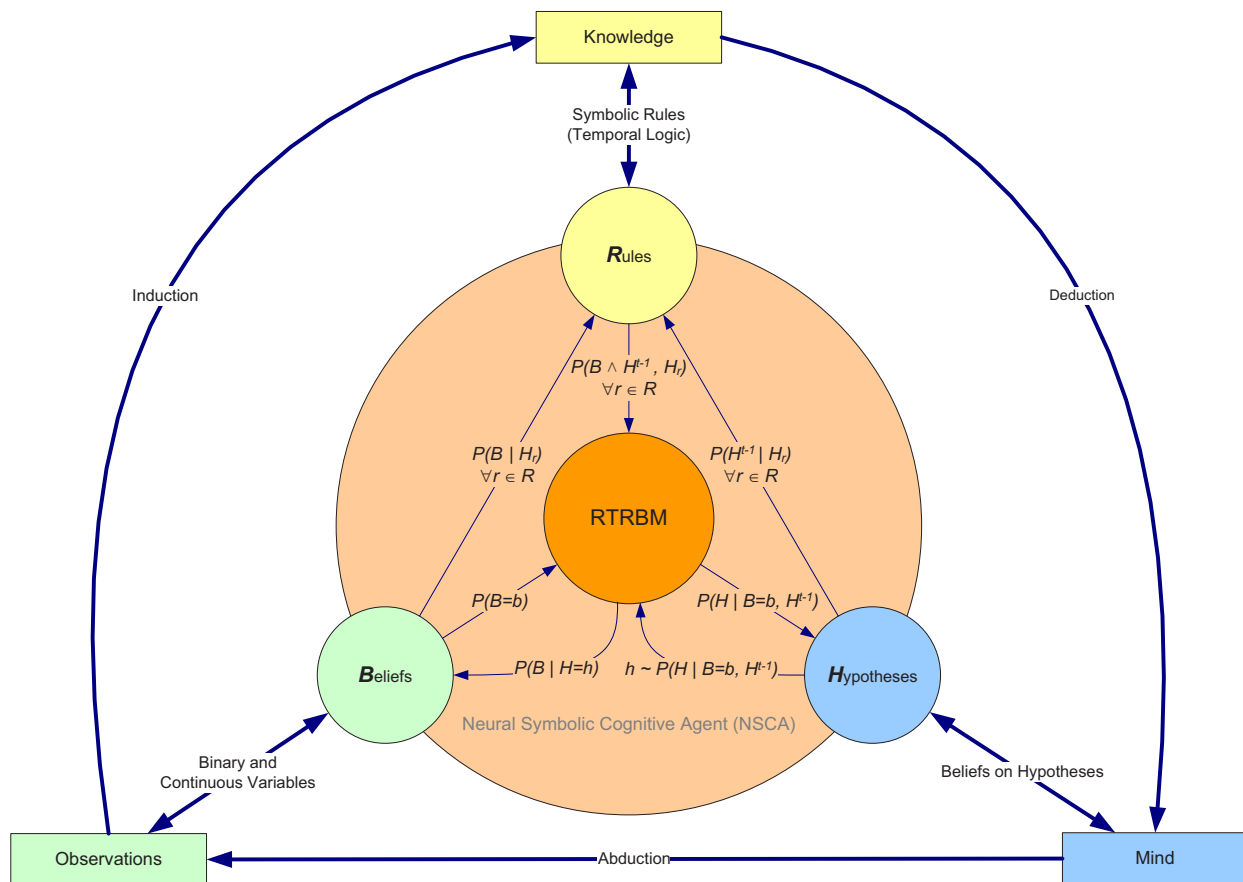
Knowledge

Symbolic Rules
(Temporal Logic)

Rules

Induction

Deduction

$P(B \wedge H^{t-1}, H_r)$
$\forall r \in R$

$P(B \mid H_r)$
$\forall r \in R$

$P(H^{t-1} \mid H_r)$
$\forall r \in R$

RTRBM

$P(B=b)$

$P(H \mid B=b, H^{t-1})$

Beliefs

$P(B \mid H=h)$ $h \sim P(H \mid B=b, H^{t-1})$

Hypotheses

Neural Symbolic Cognitive Agent (NSCA)

Binary and
Continuous Variables

Beliefs on Hypotheses

Observations

Abduction

Mind

Figure 1. The Neural-Symbolic Cognitive Agent Architecture

The NSCA architecture also enables the modelling of higher-order temporal relations using the probabilities on hypotheses (depicted as the current state of 'mind' in figure 1) of lower-level NSCAs as observations. Such a layered network of NSCAs is called a Deep Belief Network (or Deep Boltzmann Machine when RBMs are used) and is capable of meta-level learning and reasoning (Salakhutdinov, 2009).
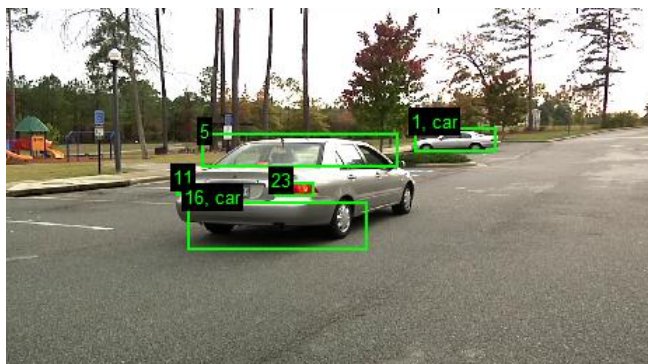

Figure 2. Detected entities (1, 5, 11, 16, 23) with object classification (car 1 and car 16).

Application to Visual Intelligence

For the Mind's Eye implementation of the NSCA in the CORTEX system the beliefs are based on observations of event properties (in total 134, e.g. movement up, size increasing, relative distance decreasing) of detected entities (see Figure 2). These properties are determined from raw video input (i.e. pixels) by the CORTEX pre-processor, which uses state-of-the-art visual processors (Felzenszwalb, Girshick, McAllester, & Ramanan, 2010; Radke, Andra, Al-Kofahi, & Roysam, 2005) to describe for each entity and video frame the probabilities of the event properties being active (see Figure 3).

To learn more generalized (i.e. first logic) relations that are independent from the number of entities, the NSCA will use beliefs that denote for each frame and event property the maximum probability over all entities, e.g. $MovingRight \equiv \exists e\ MovingRight(e)$. Later on we will describe how it is still possible to identify specific entities based on these generic beliefs (see Description).
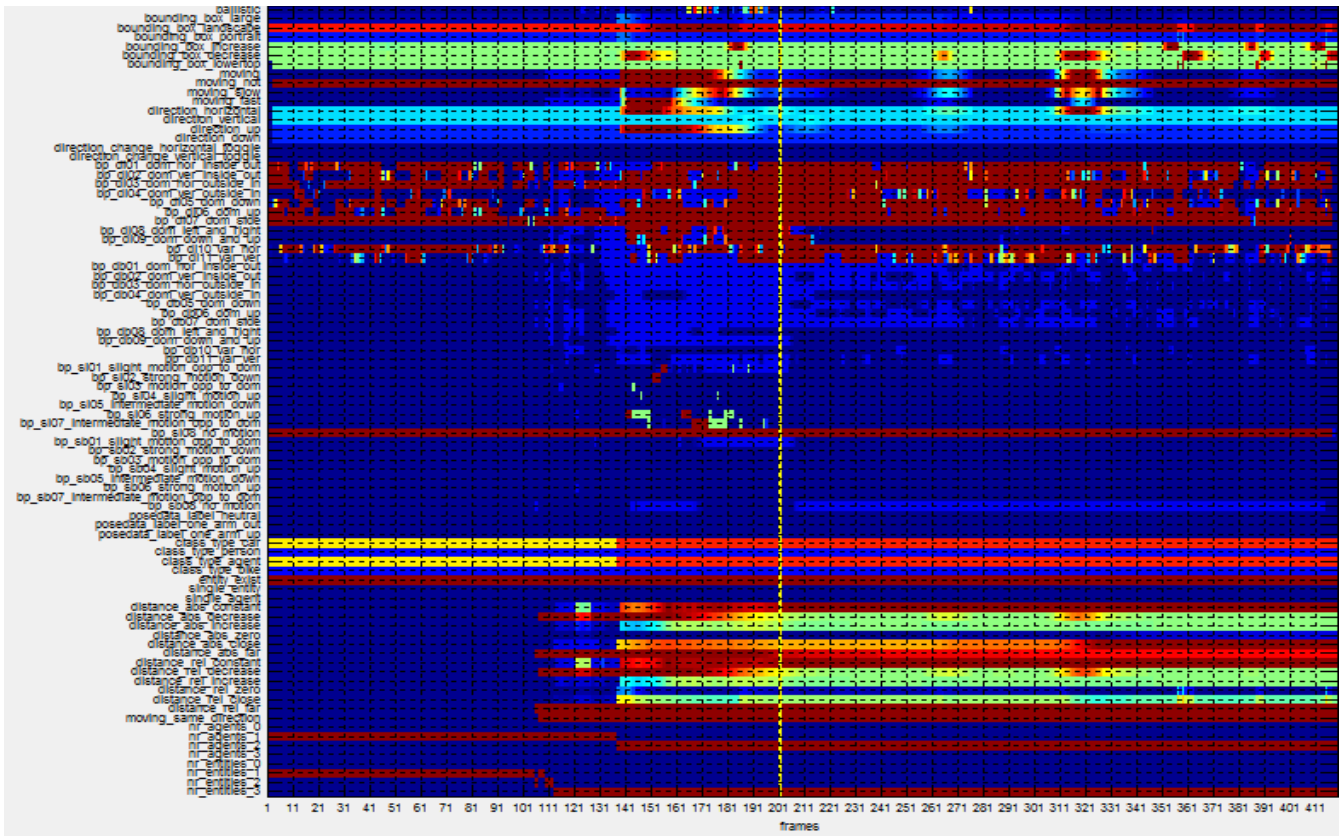
Figure 3. Beliefs on event properties (y-axis) for each video frame (x-axis), where blue is low probability and red is high probability. Yellow line depicts the current displayed frame 201 (see Figure 2).

During learning, the CORTEX system also includes human annotations on actions (e.g. *Chase*, *Fall*, *Dig*) in the observations, such that the NSCA can learn the temporal relations between the beliefs on event properties and actions (i.e. **induction**).

After learning, the NSCA can use the temporal knowledge encoded in the RTRBM to **deduce** hypotheses applicable to the current situation given the observed event properties p and the previous state of the hypotheses H^{t-1}, i.e. $P(H|B=p,H^{t-1})$. Then by selecting the most probable hypotheses h (using Gaussian sampling), the NSCA can **abduce** the related actions v from the RTRBM using its inference mechanism (i.e. $v=P(B|H=h)$).

The RTRBM's fast stochastic inference mechanism makes it possible for the NSCA, which is currently implemented in Java, to operate in real-time and deal with the uncertainties in real-world environments.

Temporal Knowledge Representation

As described in (de Penning et al., 2011) the NSCA can also **encode** and **extract** temporal knowledge about relations between beliefs on event properties and actions in the form of temporal logic clauses. This is a major advantage of the NSCA, compared to traditional machine learning, and allows us to reuse existing knowledge on actions, update the knowledge using the RTRBM, and extract the updated knowledge in symbolic form. For example, some temporal relations learned by the NSCA in CORTEX are:

0.737: H24 ↔ moving ∧ moving_fast ∧ nr_entities_3 ∧ chase ∧ flee ∧ follow ∧ go ∧ leave ∧ pass ∧ run ∧ •H0 ∧ •H1 ∧ •H2 ∧ •H3 ∧ •H4 ∧ •H5 ∧ •H6 ∧ •H8 ∧ •H11 ∧ •H12 ∧ •H13 ∧ •H14 ∧ •H15 ∧ •H16 ∧ •H17 ∧ •H19 ∧ •H20 ∧ •H21 ∧ •H22 ∧ •H23 ∧ •H24 ∧ •H25 ∧ •H27 ∧ •H29 ∧ •H31 ∧ •H32 ∧ •H33 ∧ •H34 ∧ •H36 ∧ •H37 ∧ •H38 ∧ •H39 ∧ •H40

0.561: H27 ↔ moving ∧ distance_rel_close ∧ moving_same_direction ∧ carry ∧ exit ∧ go ∧ have ∧ hold ∧ leave ∧ walk ∧ •H0 ∧ •H2 ∧ •H3 ∧ •H4 ∧ •H6 ∧ •H8 ∧ •H11 ∧ •H12 ∧ •H13 ∧ •H14 ∧ •H17 ∧ •H19 ∧ •H21 ∧ •H22 ∧ •H23 ∧ •H24 ∧ •H25 ∧ •H27 ∧ •H29 ∧ •H30 ∧ •H31 ∧ •H32 ∧ •H33 ∧ •H34 ∧ •H38 ∧ •H39 ∧ •H42 ∧ •H43 ∧ •H44 ∧ •H45 ∧ •H47 ∧ •H48
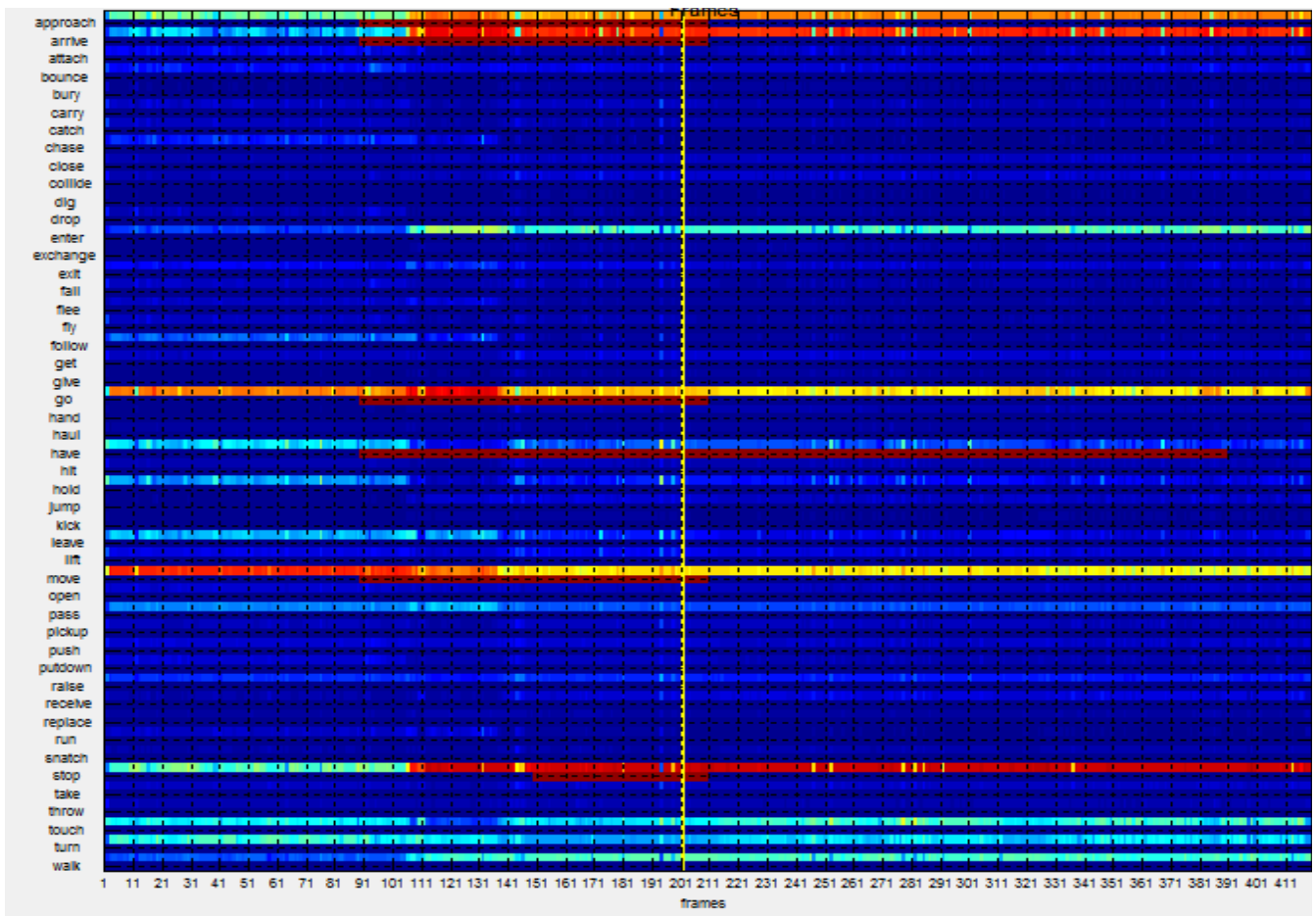
Figure 4. Depicts for each of the 48 actions (y-axis) and video frame (x-axis) the system response from the NSCA (upper line) and the temporal annotations (lower line), where blue is low probability and red is high probability. Yellow line depicts the current displayed frame 201 (see Figure 2).

Where a typical clause $H_1 \leftrightarrow B_1 \wedge B_3 \wedge \bullet H_1$ denotes that hypothesis H_1 holds at time t if and only if beliefs B_1 and B_3 hold at time t and hypothesis H_1 holds at time $t-1$, where we use the *previous time* temporal logic operator $\bullet$ to denote $t-1$ (Lamb, Borges, & d' Avila Garcez, 2007).

Evaluation and Results

During year one of the Mind's Eye program, the NSCA has been applied as part of the CORTEX system to all four visual intelligence tasks identified for the program; recognition, description, anomaly detection and gap-filling. The following subsections describe in detail how the temporal knowledge learned by the NSCA has been used to reason in each of these tasks about actions and their spatiotemporal relations with entities and their event properties detected by the CORTEX system. Also the most significant results for the NSCA of the evaluations done in year one are described. A full description of results and

comparison with other reasoning techniques can be found in (Bouma et al., 2012).

Recognition

In order to evaluate the NSCA on action recognition, it was trained with 78 beliefs on event properties (for each frame) and 48 beliefs on possible actions (for the whole clip) of 3481 video clips and tested on a set of 2588 video clips containing samples from the training set, but also previously unseen variants of these samples and completely new exemplars.

The results, show that the NSCA with its RTRBM was unable to discriminate between all 48 possible actions (*F1-measure* for the NSCA responses was 0.40, whereas the *F1-measure* of the average human response was 0.58, with *F1-measure = 2 × True Positives / (2 × True Positives + False Positives + False Negatives)*). Analysis showed that the RTRBM was able to learn good hypotheses on the training data and performed well for the more prevalent verbs, but for the less prevalent or highly semantic verbs

(e.g. bury, dig, kick, flee, haul, give) it did not. There are two possible reasons for this. First, the visual processing in CORTEX is still under development and often reports incorrect entities that introduce noise to the event properties. Secondly, the available verb annotations only described the actions present in the whole clip and not for each frame. This makes it harder for the RTRBM to learn temporal relations between event properties and verb annotations based on the state of the hypotheses in the previous frame. To investigate this another experiment has been done with temporal verb annotations for each frame that were available for a subset (~900 clips) of the training set.

The results of the quantitative analysis had improved (*F1-measure* was 0.465), and a qualitative analysis showed that the NSCA learned better temporal representations of the actions present in the videos.

As can be seen in Figure 4, the actions recognized by the NSCA not only correspond to the probability reported in the temporal annotations, but also correspond to right time frames (i.e. approach and arrive are almost in sync with the human annotation). Also we have seen that often the NSCA reports actions which have not been reported in the temporal annotations, but are clearly visible in the video (i.e. car 16 enters the scene, which was not annotated as such).

Description

Another very important visual intelligent task is to describe the observed actions in natural language. Mainly because the system must inform humans about the current situation as efficiently as possible. This requires that the system is not only able to recognize the observed actions, but also to relate these actions to the correct entities (e.g. persons, cars, object) in a scene to create a proper subject-verb-object sentence. In order to do so, the NSCA uses the temporal knowledge encoded in the RTRBM and the recognized actions to **abduce** beliefs on related event properties for each action. Then based on the Mahalanobis distance (Mahalanobis, 1936) between the abduced event properties and those of the detected entities, the NSCA determines the related subject, direct and indirect object for each recognized action. As can be seen in Figure 5, the generated descriptions, which apply to the scene depicted in Figure 2, this results in action specific sentences that only refer to the entities to which the action applies (i.e. 'arrives' applies to car 16 only and not to car 1, which is parked the whole scene).

To evaluate the descriptions produced by the NSCA, they were compared with human descriptions available from DARPA (~10 descriptions for 240 clips). This resulted in a union score of 39% (i.e. union score denotes the amount of combinations of reported verbs and related subject and object that have also been reported in at least one of the human descriptions for each clip).

In a later experiment this evaluation was repeated, but then with two modifications: To recognize the actions a model was used that was trained with temporal annotations (see Recognition) and instead of comparing event properties abduced for the whole clip, the event properties were abduced and matched with the detected entities for each frame in a clip. This resulted in a higher union score of 51%.

> *car 16 stops car 1 (0.958)*
> *car 16 goes to car 1 (0.933)*
> *car 16 arrives (0.916)*
> *car 16 moves to car 1 (0.854)*
> *car 16 approaches car 1 (0.778)*

Figure 5. Generated descriptions for recognized actions.

Anomaly Detection

The main goal of this task is to recognize anomalies based on the semantics of recognized actions. Using the temporal knowledge encoded in the NSCA, probabilities on recognized actions were abduced for each video frame and compared to the actions in a normative clip. Then based on the differences, the anomalous segments were determined. With the NSCA this resulted in an accuracy of 79.0% when compared to annotations from DARPA. This was similar to the accuracy of the other evaluated reasoners (~80%) described in (Bouma et al., 2012). Note that these results were obtained without the use of the improved RTRBM model (i.e. that was trained with temporal annotations) used in the re-evaluations for recognition and description. This is also the case for gap-filling. Re-evaluation of the NSCA in these tasks is considered as part of future work.

Gap-filling

This is the most difficult visual intelligent task and requires a combination of anomaly detection, recognition and description. Gap-filling is required when visual features are missing, because the camera blacks out, or entities are blocked by other objects, like buildings or fences. The CORTEX system must be able to fill in the gaps by detecting the missing features (i.e. anomaly detection), reason about possible actions and event properties during the gaps (i.e. recognition) and then give a proper description about these actions and related entities (i.e. description). This requires excellent performance on the other tasks, and therefore these tasks were the primary focus of attention in year one for the CORTEX team. Nonetheless, we still performed several experiments on the

gap-filling task during year one, since the main functionality required for gap-filling (e.g. robust tracking to determine the event properties and temporal reasoning) was already implemented in the system. We evaluated the generated descriptions with human descriptions provided by DARPA. The results were quite good, considering not much effort was put into optimizing the system for this task, and were similar to that of the descriptions generated for complete clips without gaps (i.e. union score of 39%).

Conclusions

The paper describes the application of a Neural Symbolic Cognitive Agent in the DARPA Mind's Eye program to perform several complex visual intelligent tasks. We have shown that the NSCA model is able to learn and reason about spatiotemporal actions and their underlying semantics from the properties of entities observed in visual input. Using a RTRBM, the NSCA is able to **induce** new knowledge from observations, **deduce** hypotheses applicable to the observed beliefs, and **abduce** new beliefs that describe the recognized actions in a video clip.

Evaluation results have shown that the NSCA has still some difficulty with discriminating between all 48 possible actions. We have also shown that the use of temporal annotations in the training set improves somewhat the quantitative results, but more importantly it improves the descriptive power of the model which is reflected in the natural language description provided in the description task. This descriptive capability also allows the NSCA to extract the temporal knowledge encoded in the RTRBM in a temporal logic form enabling it to explain humans the cognitive underpinnings of its reasoning process (the 'why'). And so, the NSCA, combining both sub-symbolic learning and symbolic reasoning, fulfills one of the key requirements in the Mind's Eye program: Visual Intelligence.

Future Work

As part of future work we are also considering the use of a NSCA to learn the event properties directly from the raw pixels in each video frame as RBMs have been reported to be good spatiotemporal feature detectors for visual input (Hinton, 2002, 2007; Mnih & Hinton, 2010). Then the event properties can be regarded as beliefs on hypotheses about temporal aspects in pixel patterns that can be observed as beliefs by higher-order NSCAs to reason about these hypotheses. This effectively creates a layered or deep belief network, which is more robust to changes in environmental conditions (e.g. lighting, camera position, type of objects).

In addition to its practical dimension to CORTEX and potential for comparative analysis, the use of such deep network architecture (Salakhutdinov, 2009) within NSCA opens up some interesting research questions. For example, the RTRBM would be allowed to influence and possibly improve the visual feature detection based on encoded knowledge of actions and related event properties.

References

Bouma, H., Hanckmann, P., Marck, J.-W., Penning, L. de, Hollander, R. den, Hove, J.-M. ten, Broek, S. van den, et al. (2012). Automatic human action recognition in a scene from visual inputs. *SPIE*, *8388*.

Donlon, J. (2010). *DARPA Mind's Eye Program: Broad Agency Announcement*. Arlington, USA.

Felzenszwalb, P., Girshick, R., McAllester, D., & Ramanan, D. (2010). Object Detection with Discriminatively Trained Part Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9).

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, *14*(8): 1771-1800.

Hinton, G. E. (2007). To recognize shapes, first learn to generate images. *Progress in brain research*, *165*: 535–547. Elsevier.

Lamb, L. C., Borges, R. V., & d' Avila Garcez, A. S. (2007). A connectionist cognitive model for temporal synchronisation and learning. *Proc. of the AAAI Conference on Artificial Intelligence*. AAAI Press.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science, Calcutta*.

Mnih, V., & Hinton, G. E. (2010). Learning to Detect Roads in High-Resolution Aerial Images. *Computer Vision–ECCV 2010*, 210–223. Springer.

Radke, R. J., Andra, S., Al-Kofahi, O., & Roysam, B. (2005). Image Change Detection Algorithms: A Systematic Survey. *IEEE Transactions on Image Processing*, *14*(3).

Salakhutdinov, R. (2009). Deep boltzmann machines. *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Sutskever, I., & Hinton, G. (2008). The recurrent temporal restricted boltzmann machine. *Advances in Neural Information*, *21*.

de Penning, L., Garcez, A. S. d'Avila, Lamb, L. C., & Meyer, J.-J. C. (2011). A Neural-Symbolic Cognitive Agent for Online Learning and Reasoning. *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*. Barcelona, Spain.