Visual Intelligence using Neural-Symbolic Learning and Reasoning

H.L.H. (Leo) de Penning
TNO Behaviour and Societal Sciences
Kampweg 5, Soesterberg, The Netherlands.
leo.depenning@tno.nl

Abstract

The DARPA Mind's Eye program seeks to develop in machines a capability that currently exists only in animals: visual intelligence. This short paper describes the initial results of a Neural-Symbolic approach for action recognition and description to be demonstrated at the 7th international workshop on Neural-Symbolic Learning and Reasoning.

Introduction

Humans in particular perform a wide range of visual tasks with ease, which no current artificial intelligence can do in a robust way. Humans have inherently strong spatial judgment and are able to learn new spatiotemporal concepts directly from the visual experience. Humans can visualize scenes and objects, as well as the actions involving those objects. Humans possess a powerful ability to manipulate those imagined scenes mentally to solve problems. A machine-based implementation of such abilities would require major advances in each of the following technology focus areas: Robust recognition, Anomaly detection, Description, Gap-filling (i.e., interpolation, prediction, and post diction). These are human intelligence-inspired capabilities, which are envisaged in service of systems to directly support humans in complex perceptual and reasoning tasks (e.g. like Unmanned Ground Vehicles).

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 contribution of TNO, a Dutch research institute and one of the teams working on the Mind's Eye program, is called CORTEX and is presented in this paper. CORTEX is a Visual Intelligence (VI) system and consists of a visual processing pipeline and reasoning component that is able to reason about events detected in visual inputs (e.g. from a movie or live camera) in order to; i) recognize actions in terms of verbs, ii) describe these actions in natural language, iii) detect anomalies and iv) fill gaps (e.g. video blackouts by missing frames, occlusion by moving objects, or entities receding behind objects).

Neural-Symbolic Cognitive Agent

To learn spatiotemporal relations between detected events (e.g. size of bounding boxes, speed of moving entities, changes in relative distance between entities) and verbs describing actions (e.g. fall, bounce, dig) the reasoning component uses a Neural-Symbolic Cognitive Agent (NSCA) that is based on a Recurrent Temporal Restricted Boltzmann Machine (RTRBM) (described in [de Penning et al., 2011] and presented during the IJCAI 2011 poster session). This cognitive agent is able to learn hypotheses about temporal relations between observed events and related actions and can express those hypotheses in temporal logic or natural language. This enables the reasoning component, and thus CORTEX, to explain and describe the cognitive underpinnings of the recognition task as stated in the focus of the Mind's Eye program.

The hypotheses are modelled in a RTRBM, where each hidden unit H_j in the RTRBM represents a hypothesis about a specific relation between events e and verbs v being observed in the visible layer V and hypotheses h^{t-1} that have been true in the previous time frame. Based on a Bayesian inference mechanism, the NSCA can reason about observed actions by selecting the most likely hypotheses h using random Gaussian sampling of the posterior probability distribution (i.e. $h \sim P(H|V=e \wedge v, H^{t-1}=h^{t-1})$) and calculating the conditional probability or likelihood of all events and verbs assuming the selected hypotheses are true (i.e. $P(V|H=h)$). The difference between the detected events, available ground truth and the inferred events and verbs can be used by the NSCA to train the RTRBM (i.e. update its weights) in order to improve the hypotheses.

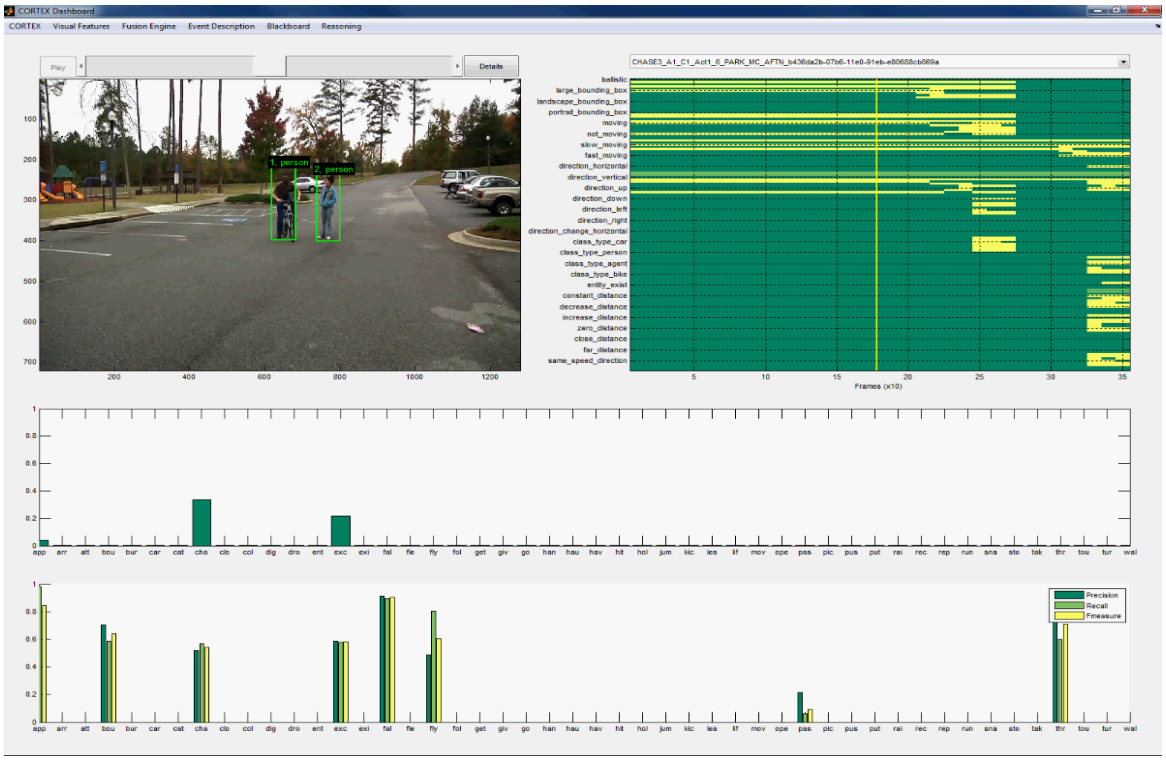

Figure 1. CORTEX Dashboard for testing and evaluation.

Experiments and Results

The CORTEX system and its reasoning component are currently being tested on a recognition task using several datasets of movies and related ground truth provided by DARPA. Figure 1 shows the CORTEX Dashboard, a user-interface for testing and evaluation of the CORTEX system. With the CORTEX Dashboard we are able to visualize the detected entities (depicted by bounding boxes in the upper-left image), probabilities on related events in each frame (depicted by intensities, green is 0 and yellow is 1, in the upper-right graph) and related verb probabilities calculated by the reasoning component (depicted by bars in the centre graph). The bottom graph shows the precision, recall and F-measure (i.e. harmonic mean of the precision and recall) for all verbs used to evaluate the output of the reasoning component. Also it can visualize the learned hypotheses and extract these in the form of temporal logic or natural language, which can be used to explain and describe the recognized actions.

Initial results show that the reasoning component is capable of learning hypotheses about events and related verbs and that it is able to reason with these hypotheses to correctly recognize actions based on detected events.
Furthermore the results show that the reasoning component is able to recognize actions that were not there in the ground truth for that specific input, but inferred from ground truth and related event patterns in other input. For example, reasoning about a movie that was trained to be recognized as a chase, resulted in some parts being recognized as fall, because one of the persons was tilting over when she started running, although fall was not part of the ground truth for this movie.

Conclusions and Future Work

With the NSCA architecture, the reasoning component is able to learn and reason about spatiotemporal events in visual inputs and recognize these in terms of actions denoted by verbs. It is also able to extract learned hypotheses on events and verbs that can be used to explain the perceptual and cognitive underpinnings of the recognition task and support other visual intelligence tasks, like description, anomaly detection and gap-filling, yet to be developed in the CORTEX system.

References

[Donlon, 2010] James Donlon. *DARPA Mind's Eye Program: Broad Agency Announcement*. Arlington, USA, 2010.

[de Penning et al., 2011] Leo de Penning, Artur S. d'Avila Garcez, Luís C. Lamb, and John-Jules C. Meyer. A Neural-Symbolic Cognitive Agent for Online Learning and Reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Barcelona, Spain, 2011.