Learning from Incongruence

Tomáš Pajdla, Michal Havlena, and Jan Heller

Abstract. We present an approach to constructing a model of the universe for explaining observations and making decisions based on learning new concepts. We use a weak statistical model, e.g. a discriminative classifier, to distinguish errors in measurements from improper modeling. We use boolean logic to combine outcomes of direct detectors of relevant events, e.g. presence of sound and presence of human shape in the field of view, into more complex models explaining the states in which the universe may appear. The process of constructing a new concept is initiated when a significant disagreement – incongruence – has been observed between incoming data and the current model of the universe. Then, a new concept, i.e. a new direct detector, is trained on incongruent data and combined with existing models to remove the incongruence. We demonstrate the concept in an experiment with human audio-visual detection.

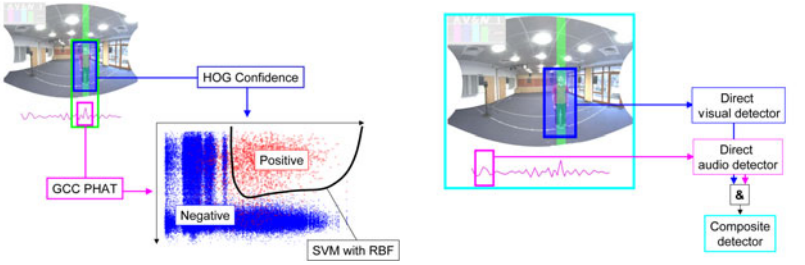
1 Introduction

Intelligent systems compare their model of the universe, the "theory of the universe", with observations and measurements they make. The comparison of conclusions made by reasoning about well established building blocks of the theory with direct measurements associated with the conclusions allow to falsify [1] current theory and to invoke a rectification of the theory by learning from observations or restructuring the derivation scheme of the theory. It is the disagreement – incongruence – between the theory, i.e. derived conclusions, and direct observations that allows for developing a richer and better model of the universe used by the system.

Works [2, 3] proposed an approach to modeling incongruences between classifiers (called detectors in this work) which decide about the occurrence of concepts

Tomáš Pajdla · Michal Havlena · Jan Heller
Center for Machine Perception, Department of Cybernetics, FEE, CTU in Prague,
Technická 2, 166 27 Prague 6, Czech Republic
e-mail: {pajdla,havlem1,hellej1}@cmp.felk.cvut.cz

D. Weinshall, J. Anemüller, and L. van Gool (Eds.): DIRAC, SCI 384, pp. 119–127.
springerlink.com



(a) The direct audio-visual (A/V) human speaker detector

(b) The composite audio-visual (A&V) human speaker detector

Fig. 1 (a) The direct audio-visual human speaker detector constructed by training an SVM classifier with a RBF kernel in the two-dimensional feature space of GCC-PHAT values (*x*-axis) and pedestrian detection scores (*y*-axis) for different positive (*red circles*) and negative (*blue crosses*) manually labeled examples [6]. (b) The composite audio-visual human speaker detector accepts if and only if the direct visual detector AND the direct audio detector both accept (but possibly at different places) in the field of view. See [6] or an accompanied paper for more details.

(events) via two different routes of reasoning. The first way uses a single *direct* detector trained on complete, usually complex and compound, data to decide about the presence of an event. The alternative way decides about the event by using a *composite* detector, which combines outputs of several (in [2, 3] direct but in general possibly also other composite) detectors in a probabilistic (logical) way, Figure 1.

Works [2, 3] assume direct detectors to be independent, and therefore combine probabilities by multiplication for the "part-membership hierarchy", resp. by addition for the "class-membership hierarchy". Assuming trivial probability space with values 0 and 1, this coincides with logical AND and logical OR. Such reasoning hence corresponds to the Boolean algebra [4]. In the next we will look at this simplified case. A more general case can be analyzed in a similar way.

The theory of incongruence [2, 3] can be used to improve low-level processing by detecting incorrect functionality and repairing it through re-defining the composite detector. In this work we look at an example of incongruence caused by the omission of an important concept in an example of audio-visual speaker detection and show how it can be improved. Figure 1 and Table 1 illustrate a prototypical system consisting of alternative detectors, which can lead to a disagreement between the alternative outcomes related to an event.

Three direct detectors and one composite detector are shown in Figure 2(a). The direct detector of "Sound in view", the direct detector of "Person in view", the direct detector of "Speaker", and the composite detector of "Speaker" are presented. The composite detector was constructed as a logical combination of direct detectors evaluated on the whole field of view, hence not capturing the spatial co-location of sound and look events defining a speaker in the scene. See [6] or an accompanied paper for more details.